SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 1–23, January 1994

A GENERAL NECESSARY CONDITION FOR EXACT OBSERVABILITY*

DAVID L. RUSSELL† AND GEORGE WEISS‡

Abstract. Suppose A generates an exponentially stable strongly continuous semigroup on the Hilbert space X, Y is another Hilbert space, and $C : D(A) \to Y$ is an admissible observation operator for this semigroup. (This means that to any initial state in X we can associate an output function in $L^2([0,\infty), Y)$.) This paper gives a necessary condition for the exact observability of the system defined by A and C. This condition, called **(E)**, is related to the Hautus Lemma from finite dimensional systems theory. It is an estimate in terms of the operators A and C alone (in particular, it makes no reference to the semigroup). This paper shows that **(E)** implies approximate observability and, if A is bounded, it implies exact observability. This paper conjectures that **(E)** is in fact equivalent to exact observability. The paper also shows that for diagonal semigroups, **(E)** takes on a very simple form, and applies the results to sequences of complex exponential functions.

Key words. exact observability, admissible observation operators, diagonal semi-groups, Riesz bases of complex exponentials

AMS(MOS) subject classifications. 93B07, 93B28, 93C25

1. Introduction and statement of the main results.

Let X be a Hilbert space and suppose $\mathbb{T} = (\mathbb{T}_t)_{t\geq 0}$ is an exponentially stable, strongly continuous semigroup of operators on X, with generator $A : D(A) \to X$. Let Y be another Hilbert space and suppose $C : D(A) \to Y$ is a linear operator which is A-bounded, i.e.,

$$\|Cx\| \leq L \cdot \|Ax\| \tag{1.1}$$

holds for some $L \geq 0$ and any $x \in D(A)$.

This paper is concerned with the system described by the equations

$$\dot{z}(t) = Az(t), \quad z(0) = x, \tag{1.2a}$$

$$y(t) = Cz(t), \tag{1.2b}$$

where $t \geq 0$. The element $x \in X$ is called the *initial state*, $z(t)$ is called the *state* at time t, and y is the *output function*. By a solution of (1.2a) we mean of course $z(t) = \mathbb{T}_t x$ (this is a weak solution). Equation (1.2b) is more problematic: if $x \notin D(A)$ then it might happen that $z(t)$ is never in $D(A)$, so that $Cz(t)$ is not defined.

To ov rcome this difficulty, we assume that C is an *admissible observation operator* for $\mathbb{T}$, wh·ch means the following: there is a $K \geq 0$ such that

$$\int_0^\infty \|C\mathbb{T}_t x\|^2 dt \leq K \cdot \|x\|^2, \tag{1.3}$$

for any $x \in D(A)$. If C is admissible then the operator $\Psi_\infty : D(A) \to L^2([0,\infty), Y)$, defined by

$$(\Psi_\infty x)(t) = C\mathbb{T}_t x, \tag{1.4}$$

* Received by the editors April 15, 1991; accepted for publication April 8, 1992.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. This author's work was supported by Air Force Office of Scientific Research contract AFOSR-89-0031.

‡ Department of Electrical Engineering, Ben-Gurion University, Beer Sheva 84105, Israel. This author's work was supported by a Weizmann Fellowship and by Air Force Office of Scientific Research contract AFOSR-89-0031, and was carried out during a visit to the Interdisciplinary Center for Applied Mathematics (ICAM) at Virginia Tech.

has a continuous extension to X. This extension, still denoted Ψ_∞, is called the *extended output map* of A and C. For more details on admissibility, see §2. Now by the solution of (1.2b) we mean the function $y = \Psi_\infty x$.

DEFINITION 1.1. The system described by (1.2) is *exactly observable* on $[0, \infty)$, if there is a $\kappa > 0$ such that for any $x \in D(A)$,

$$\int_0^\infty \|C\mathbb{T}_t x\|^2 dt \geq \kappa \cdot \|x\|^2. \tag{1.5}$$

Clearly, (1.5) means that the extended output map Ψ_∞ is bounded from below, i.e., Ψ_∞ has a bounded left inverse, i.e., the problem of computing x, given y in (1.2), is well posed. Note the analogy between (1.3) and (1.5) (admissibility and exact observability).

As is well known, the concept of admissible observation operator is dual to that of admissible control operator (see, e.g., Salamon [16]), and the concept of exact observability is dual to that of exact controllability (see, e.g., Dolecki and Russell [3]). It will be understood that all the definitions and results in this paper have a dual counterpart.

The concept of exact observability (and its dual) has received considerable attention in recent years, see, e.g., the treatise of Lions [11], the survey papers of Lagnese [9], Lions [12], and Russell [18], or the general exposition of Bensoussan [1]. Usually, the emphasis is on exact observability on a finite time interval $[0, \tau)$ (or its dual property), which means that the integral in (1.5) is over $[0, \tau)$ only. However, exact observability on $[0, \infty)$ is equivalent to exact observability on some finite time interval (this follows from admissibility and exponential stability; see Proposition 2.8).

Throughout this paper, $\mathbb{C}_-$ denotes the left open half-plane in $\mathbb{C}$, and $\mathbb{C}_+$ denotes the right open half-plane in $\mathbb{C}$. Our main result follows.

THEOREM 1.2. *Let $X, \mathbb{T}, A, Y$ and C be as in* (1.1) *and* (1.3). *If the system described by* (1.2) *is exactly observable on $[0, \infty)$, then the following estimate is true:*

(E) *There is an $m > 0$ such that for any $s \in \mathbb{C}_-$ and any $x \in D(A)$,*

$$\frac{1}{|\text{Re } s|^2}\|(sI - A)x\|^2 + \frac{1}{|\text{Re } s|}\|Cx\|^2 \geq m \cdot \|x\|^2. \tag{1.6}$$

The proof of this theorem is given in §3. We conjecture that its converse is also true.

CONJECTURE 1.3. *Let $X, \mathbb{T}, A, Y$ and C be as in* (1.1) *and* (1.3). *If* **(E)** *holds, then the system described by* (1.2) *is exactly observable on $[0, \infty)$.*

Theorem 1.2 and Conjecture 1.3 are an attempt to generalize the Hautus Lemma, due to Popov [15] and Hautus [5], which concerns finite-dimensional linear systems. The Hautus Lemma states that if $A \in \mathcal{L}(\mathbb{C}^n)$ and $C \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^p)$, then the system defined by (1.2) is observable if and only if

$$\text{rank} \begin{bmatrix} sI - A \\ C \end{bmatrix} = n \quad \forall s \in \mathbb{C}.$$

Observing that it is sufficient to verify this condition for $s \in \sigma(A)$ (the spectrum of A), we can restate the Hautus Lemma for the case of stable A (i.e., $\sigma(A) \subset \mathbb{C}_-$) in the following form, visibly related to Theorem 1.2.

PROPOSITION 1.4. *Suppose $A \in \mathcal{L}(\mathbb{C}^n), C \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^p)$ and $\sigma(A) \subset \mathbb{C}_-$. Then the system described by* (1.2) *is observable if and only if, for any $s \in \mathbb{C}_-$ and any nonzero $x \in \mathbb{C}^n$,*

$$\|(sI - A)x\|^2 + \|Cx\|^2 > 0.$$

For a short proof of the Hautus Lemma and related material see, e.g., Sontag [21]. It is not difficult to prove that, with A and C matrices and A stable, (**E**) is equivalent to the condition in Proposition 1.4.

We now return to infinite-dimensional systems.

DEFINITION 1.5. The system described by (1.2) is *approximately observable* on $[0,\infty)$ if for any nonzero $x \in D(A)$,

$$\int_0^\infty \|C\mathbb{T}_t x\|^2 dt > 0.$$

The above condition is equivalent with $\operatorname{Ker} \Psi_\infty = \{0\}$, where Ψ_∞ is the extended output map; see Remark 3.2. Obviously, exact observability implies approximate observability. The following theorem may be regarded as a partial result for Conjecture 1.3.

THEOREM 1.6. *Let $X, \mathbb{T}, A, Y$ and C be as in* (1.1) *and* (1.3). *If the estimate* (**E**) *holds, then the system described by* (1.2) *is approximately observable on* $[0,\infty)$.

The proof is in §3. The following theorem shows that Conjecture 1.3 is true at least in the case when A (and hence also C) is bounded.

THEOREM 1.7. *Let X and Y be Hilbert spaces and suppose $A \in \mathcal{L}(X), C \in \mathcal{L}(X,Y)$ and $\sigma(A) \subset \mathbb{C}_-$. If for every $s \in \mathbb{C}_-$ there is an $m_s > 0$ such that for each $x \in X$,*

$$\|(sI - A)x\|^2 + \|Cx\|^2 \geq m_s \cdot \|x\|^2, \tag{1.7}$$

then the system described by (1.2) *is exactly observable on* $[0,\infty)$.

Note that (1.6) implies (1.7). This theorem follows from results of Rodman [17]. A different proof is given in §3.

In §4 we consider a special class of systems described by (1.2) namely, we assume that A is a diagonal operator of l^2 and that its eigenvalues λ_k are *properly spaced*. That means that the eigenvalues are not too close to each other: $|\lambda_j - \lambda_k| \geq \delta \cdot |\text{Re } \lambda_k|$ for all $j,k \in \mathbb{N}$ with $j \neq k$, where $\delta > 0$. We show that for such systems, the estimate (**E**) is equivalent to a simple and easily verifiable condition on the operator C: if (e_k) is the standard basis of l^2 then $\|Ce_k\|/|\text{Re } \lambda_k|$ should be bounded away from 0.

In §5 we consider sequences of functions of the form

$$p_k(t) = e^{\lambda_k t} c_k,$$

where $\lambda_k \in \mathbb{C}_-$ and $c_k \in Y, Y$ being a Hilbert space. The problem is to find necessary as well as sufficient conditions for the sequence (p_k) to be a Riesz basis in its closed span in $L^2([0,\infty), Y)$. A clear necessary condition is $\|\sum x_k p_k\|^2 \leq \sum |x_k|^2$, for any finite sequence (x_k) of complex numbers. Another simple necessary condition is that the norms $\|p_k\|$ should be bounded from below. Surprisingly, these two necessary conditions combined are also sufficient for certain sequences (λ_k). If Conjecture 1.3 is true for diagonal semigroups, then they are sufficient whenever the sequence (λ_k) is properly spaced and bounded away from the imaginary axis. We display an interesting example of an analytic diagonal semigroup for which Conjecture 1.3 is true.

Note. The authors propose Conjecture 1.3 as a challenge, to prove or disprove. Contact one of the authors for details.

2. Some background on admissibility and observability. First we return to the definition of admissible observation operators. We work in a slightly more general context than in §1, since we do not assume that the semigroup is exponentially stable. The concept of admissible observation operator (as defined here) has its origin in Dolecki and Russell

[3]. Since then, many authors have addressed the subject. Our notation and terminology follows Weiss [23].

Let X by a Hilbert space and suppose $\mathbb{T} = (\mathbb{T}_t)_{t\geq 0}$ is a strongly continuous semigroup of operators on X, with generator $A : D(A) \to X$. We define the Hilbert space X_1 to be $D(A)$ with the norm

$$\|x\|_1 = \|(\beta I - A)x\|, \tag{2.1}$$

where $\beta \in \rho(A)$, the resolvent set of A, and $\|\cdot\|$ denotes the norm on X. It is easy to check that $\|\cdot\|_1$ is equivalent with the graph norm of A, in particular, the topology of X_1 is independent of the choice of β. We have

$$X_1 \subset X,$$

densely and with continuous embedding.

DEFINITION 2.1. Let $X, \mathbb{T}, A$, and X_1 be as above. Let Y be a Hilbert space and suppose $C \in \mathcal{L}(X_1, Y)$. Then C is an *admissible observation operator* for $\mathbb{T}$ if for some (and hence any) $\tau > 0$, the operator $\Psi_\tau : X_1 \to L^2([0,\infty), Y)$ defined by

$$(\Psi_\tau x)(t) = C\mathbb{T}_t x \quad \text{for } t \in [0,\tau), \tag{2.2}$$

and $(\Psi_\tau x)(t) = 0$ for $t \geq \tau$, has a continuous extension to X.

In other words, admissibility of C means that for some $\tau > 0, K_\tau \geq 0$ and any $x \in D(A)$,

$$\int_0^\tau \|C\mathbb{T}_t x\|^2 dt \leq K_\tau \cdot \|x\|^2. \tag{2.3}$$

It is not difficult to verify that if (2.3) holds for some $\tau > 0$, then it holds for any other $\tau > 0$. We denote the extension of Ψ_τ to X by the same symbol. If $\mathbb{T}$ is exponentially stable, then we may take also $\tau = \infty$ in Definition 2.1, obtaining the equivalent definition of admissibility given in §1. In this case, the extended output map Ψ_∞ (defined in §1) is the strong limit of the operators Ψ_τ, as $\tau \to \infty$. For any $x \in X$ and any $\tau \geq 0, \Psi_\tau x$ can be obtained from $\Psi_\infty x$ by truncation to $[0,\tau)$.

The operators Ψ_τ satisfy an interesting functional equation, which is used to define linear observation systems in an abstract way, as follows.

DEFINITION 2.2. Let X and Y be Hilbert spaces. An *abstract linear observation system* with state space X and output space Y is a pair $(\mathbb{T}, \Psi)$, where $\mathbb{T} = (\mathbb{T}_t)_{t\geq 0}$ is a strongly continuous semigroup on X and $\Psi = (\Psi_t)_{t\geq 0}$ is a family of bounded operators from X to $L^2([0,\infty), Y)$, such that $\Psi_0 = 0$ and

$$(\Psi_{\tau+t}x)(\sigma) = \begin{cases} (\Psi_\tau x)(\sigma) & \text{for } \sigma \in [0,\tau), \\ (\Psi_t \mathbb{T}_\tau x)(\sigma - \tau) & \text{for } \sigma \geq \tau, \end{cases}$$

for any $x \in X$, any τ, $t \geq 0$ and almost every $\sigma \geq 0$.

With the notation of Definition 2.1, let Ψ be the family of operators defined by (2.2) and continuous extension to X. Then it is easy to verify that $(\mathbb{T}, \Psi)$ is an abstract linear observation system. Conversely, every abstract linear observation system is obtained in this way. This is the content of a *representation theorem* proved in Salamon [20] and Weiss [23] (we refrain from its formal statement). One consequence is that we can restrict our attention to operators $C \in \mathcal{L}(X_1, Y)$; there is no need to consider operators C defined on other dense $\mathbb{T}$-invariant subspaces of X.

When, in §1, we wrote "the system described by (1.2)" then, strictly speaking, we meant the system $(\mathbb{T}, \Psi)$ determined by A and C. We will continue to use this terminology.

PROPOSITION 2.3. *Let $\mathbb{T}$ be an exponentially stable, strongly continuous semigroup on the Hilbert space X, with generator A. Let Y be a Hilbert space and suppose $C \in \mathcal{L}(X_1, Y)$ is an admissible observation operator for $\mathbb{T}$. Then there is a $K \geq 0$ such that for any $s \in \mathbb{C}_+$,*

$$\|C(sI - A)^{-1}\|_{\mathcal{L}(X,Y)} \leq \frac{K}{\sqrt{\operatorname{Re} s}}. \tag{2.4}$$

Proof. We have for any $x \in D(A)$, using (1.3),

$$\begin{aligned} \|C(sI - A)^{-1}x\|^2 &= \left\| \int_0^\infty e^{-st} C\mathbb{T}_t x \, dt \right\|^2 \\ &\leq \left(\int_0^\infty |e^{-st}|^2 dt \right) \cdot \left(\int_0^\infty \|C\mathbb{T}_t x\|^2 dt \right) \\ &\leq \frac{1}{2 \operatorname{Re} s} \cdot K \cdot \|x\|^2, \end{aligned}$$

which implies (2.4). □

We mention that if $\mathbb{T}$ is normal and Y is finite-dimensional, then (2.4) is not only necessary but also sufficient for the admissibility of C. For this and other, related results see Weiss [24].

DEFINITION 2.4. Let $X, \mathbb{T}, A, Y$, and C be as in Proposition 2.3 and let Ψ_∞ be the extended output map of A and C. The *observability Gramian* of A and C is the operator $P \in \mathcal{L}(X)$ defined by

$$P = \Psi_\infty^* \Psi_\infty.$$

PROPOSITION 2.5. *With the notation of Definition* 2.4, *we have for any $x \in D(A)$,*

$$\|Cx\|^2 = -2 \operatorname{Re} \langle Px, Ax \rangle. \tag{2.5}$$

Proof. Take $x \in D(A^2)$ and define $f(t) = \|C\mathbb{T}_t x\|^2$, for any $t \geq 0$. Then f is continuously differentiable and

$$\frac{d}{dt} f(t) = \langle C\mathbb{T}_t Ax, C\mathbb{T}_t x \rangle + \langle C\mathbb{T}_t x, C\mathbb{T}_t Ax \rangle.$$

Due to the exponential stability of $\mathbb{T}$, we can integrate over $[0, \infty)$ and obtain

$$-f(0) = \langle \Psi_\infty Ax, \Psi_\infty x \rangle + \langle \Psi_\infty x, \Psi_\infty Ax \rangle,$$

i.e.,

$$-\|Cx\|^2 = 2 \operatorname{Re} \langle \Psi_\infty x, \Psi_\infty Ax \rangle,$$

which is the same as (2.5). Since both sides of (2.5) are continuous functions of x on X_1 and $D(A^2)$ is dense in X_1, (2.5) must hold for any $x \in X_1$. □

Remark 2.6. If X_1^* denotes the antilinear dual of X_1 and we identify X with its antilinear dual, then we have the dense inclusions $X_1 \subset X \subset X_1^*$. If we regard A as a

bounded operator from X_1 to X, then $A^* \in \mathcal{L}(X, X_1^*)$ is an extension of A^* computed as the adjoint of an unbounded operator in X. Formula (2.5) can now be rewritten as

$$A^*P + PA = -C^*C$$

where both sides are in $\mathcal{L}(X_1, X_1^*)$. This can be thought of as an equation in P, called a *Lyapunov equation.*

DEFINITION 2.7. Let $\mathbb{T}$ be a strongly continuous semigroup on the Hilbert space X, with generator A. Let Y be a Hilbert space and suppose $C \in \mathcal{L}(X_1, Y)$ is an admissible observation operator for $\mathbb{T}$. The system described by (1.2) is *exactly observable* on $[0, \tau)$ (where $\tau > 0$) if the operator Ψ_τ (defined by (2.2) and continuous extension to X) is bounded from below.

In other words, exact observability on $[0, \tau)$ means that

$$\int_0^\infty \|C\mathbb{T}_t x\|^2 dt \geq \kappa_\tau \cdot \|x\|^2 \tag{2.6}$$

holds for some $\kappa_\tau > 0$ and any $x \in D(A)$. Whereas in Definition 2.1, the choice of τ did not matter, in Definition 2.7, the choice of τ is important: if (2.6) holds for some $\tau > 0$, then obviously it holds for any bigger number, but not necessarily for a smaller one.

If $\mathbb{T}$ is exponentially stable, then we may take also $\tau = \infty$ in Definition 2.7, reobtaining the definition of exact observability on $[0, \infty)$. At first glance, this concept appears to be weaker than exact observability on some finite time interval, so that the following proposition is slightly surprising.

PROPOSITION 2.8. *Let $X, \mathbb{T}, A, Y$, and C be as in Proposition* 2.3. *If the system described by* (1.2) *is exactly observable on $[0, \infty)$, then there is a $\tau > 0$ such that this system is exactly observable on $[0, \tau)$.*

Proof. For any $x \in D(A)$ and any $\tau > 0$, we have

$$\int_0^\tau \|C\mathbb{T}_t x\|^2 dt = \int_0^\infty \|C\mathbb{T}_t x\|^2 dt - \int_0^\infty \|C\mathbb{T}_t \mathbb{T}_\tau x\|^2 dt.$$

Using (1.3) and (1.5), we obtain

$$\begin{aligned}\int_0^\tau \|C\mathbb{T}_t x\|^2 dt &\geq \kappa \cdot \|x\|^2 - K \cdot \|\mathbb{T}_\tau x\|^2 \\ &\geq (\kappa - K \cdot \|\mathbb{T}_\tau\|^2) \cdot \|x\|^2.\end{aligned}$$

Since $\mathbb{T}$ is exponentially stable, the parenthesis above becomes positive for τ sufficiently big, so that we get (2.6). □

Remark 2.9. The definition of exact observability on $[0, \infty)$ can be extended also to systems whose semigroup is not exponentially stable (we might have to allow the value ∞ on the left-hand side of (1.5)). However, Proposition 2.8 cannot be generalized to such semigroups. A simple example is as follows: Consider $X = L^2[0, \infty)$ and for any $t \geq 0$, let $\mathbb{T}_t$ be the left shift by t on X. Define C on $D(A) = H^1[0, \infty)$ by $Cx = x(0)$. Then the system described by (1.2) is exactly observable on $[0, \infty)$, but it is not exactly observable on any finite interval. A physically more relevant example involves the linear water wave equation, as described, e.g., in Reid and Russell [16].

Remark 2.10. Admissibility and exact observability are invariant under translations of the generator. More precisely, with the notation of Definition 2.7, C is admissible also for the semigroup generated by $A + \lambda I$, for any $\lambda \in \mathbb{C}$. Similarly, if the system described by

(1.2) is exactly observable on $[0,\tau)$, $\tau<\infty$, then, replacing A by $A+\lambda I$, the new system is exactly observable on $[0,\tau)$. These statements are easy to verify.

Example 2.11. Let Ω be a bounded domain in $\mathbb{R}^n$, with boundary Γ of class C^2. We put

$$X=\begin{matrix}H_0^1(\Omega)\\ \times\\ L^2(\Omega)\end{matrix},\qquad D(A)=\begin{matrix}H^2(\Omega)\cap H_0^1(\Omega)\\ \times\\ H_0^1(\Omega)\end{matrix}$$

and define $A: D(A)\to X$ by

$$A=\begin{pmatrix}0 & I\\ \Delta & 0\end{pmatrix},$$

where Δ is the Laplacian. Then A is the generator of a strongly continuous group of unitary operators on X. Fix $\xi_0\in\mathbb{R}^n$ and put

$$\Gamma_0=\{\xi\in\Gamma|(\xi-\xi_0)\cdot\nu(\xi)>0\},$$

where $\nu(\xi)$ is the outward normal to Γ at ξ. Let $Y=L^2(\Gamma_0)$ and define $C: D(A)\to Y$ by

$$C\begin{pmatrix}x_1\\ x_2\end{pmatrix}(\xi)=\frac{\partial x_1(\xi)}{\partial\nu},$$

where $\partial/\partial\nu$ denotes outward normal derivative on Γ and $\xi\in\Gamma_0$. Then C is an admissible observation operator for the group generated by A. (This statement remains true if Γ_0 is replaced by Γ). For details and further references on this see Lasiecka, Lions, and Triggiani [10]. Moreover, the system described by (1.2) is exactly observable on $[0,\tau)$, where τ depends on Ω. This was proved by Ho in [6] (see also Lions [11, p. 55]).

If we replace A by $\tilde{A}=A-I$, then the semigroup becomes exponentially stable. By Remark 2.10, the system determined by $\tilde{A}$ and C is exactly observable on $[0,\infty)$, so that Theorem 1.2 applies. The important values of s for (1.6) are those on the vertical line with $\mathrm{Re}\,s=-1$ (because this line contains the spectrum of $\tilde{A}$). By writing down (**E**) for this system and $s=-1+i\omega$, where $\omega\in\mathbb{R}$, we get the following estimate.

There is an $m>0$ such that for any $x_1\in H^2(\Omega)\cap H_0^1(\Omega)$, any $x_2\in H_0^1(\Omega)$ and any $\omega\in\mathbb{R}$,

$$\|i\omega x_1-x_2\|^2_{H_0^1(\Omega)}+\|\Delta x_1-i\omega x_2\|^2_{L^2(\Omega)}+\left\|\frac{\partial x_1}{\partial\nu}\right\|^2_{L^2(\Gamma_0)}\geq m(\|x_1\|^2_{H_0^1(\Omega)}+\|x_2\|^2_{L^2(\Omega)}).$$

In particular, taking $x_1=u, x_2=i\omega u$ and $\omega^2=\lambda$, we get that for any $u\in H^2(\Omega)\cap H_0^1(\Omega)$ and any $\lambda\geq 0$,

$$\|(\Delta+\lambda)u\|^2_{L^2(\Omega)}+\left\|\frac{\partial u}{\partial\nu}\right\|^2_{L^2(\Gamma_0)}\geq m(\|u\|^2_{H_0^1(\Omega)}+\lambda\cdot\|u\|^2_{L^2(\Omega)}). \tag{2.7}$$

Remark 2.12. The estimate (2.7) can be obtained also by direct computations, without using Theorem 1.2, as sketched below. Introducing the "multiplier" function $h(\xi)=\xi-\xi_0$ we can show, integrating by parts twice, that for any real-valued $u\in H^2(\Omega)\cap H_0^1(\Omega)$,

$$\int_\Omega(\Delta u)(h\cdot\nabla u)d\xi=\frac{1}{2}\int_\Gamma\left(\frac{\partial u}{\partial\nu}\right)^2(h\cdot\nu)d\gamma+\frac{n-2}{2}\int_\Omega\|\nabla u\|^2d\xi$$

(see, e.g., Komornik [8, Lemma 2.2]). Performing some more integrations by parts, we can get the identity

$$-2\int_\Omega (\Delta u + \lambda u)\left(h\cdot\nabla u + \frac{n-1}{2}u\right)d\xi + \int_\Gamma \left(\frac{\partial u}{\partial \nu}\right)^2 (h\cdot\nu)d\gamma = \lambda\int_\Omega u^2 d\xi + \int_\Omega \|\nabla u\|^2 d\xi, \tag{2.8}$$

valid for any real λ. Using that for any $r > 0$,

$$-2\int_\Omega (\Delta u + \lambda u)\left(h\cdot\nabla u + \frac{n-1}{2}u\right)d\xi \leq r\int_\Omega (\Delta u + \lambda u)^2 d\xi + \frac{1}{r}\int_\Omega \left(h\cdot\nabla u + \frac{n-1}{2}u\right)^2 d\xi$$

and choosing r sufficiently large to make the last term above smaller than the last term in (2.8), we get an inequality which is equivalent to (2.7).

Remark 2.13. Let us do a little speculation in connection with Conjecture 1.3. With the notation of this conjecture, suppose that (**E**) holds. Using (2.5) and elementary manipulations, we get that for any $s \in \mathbb{C}_-$ and any $x \in D(A)$,

$$\frac{1}{|\text{Re } s|^2}\|(sI - A)x\|^2 + \frac{2}{|\text{Re } s|}\,\text{Re}\,\langle Px, (sI - A)x\rangle + 2\langle Px, x\rangle \geq m\cdot\|x\|^2.$$

Denoting

$$v(s,x) = \frac{1}{|\text{Re } s|}(sI - A)x,$$

the above estimate can be written in the form

$$\|v(s,x)\|^2 + 2\,\text{Re}\,\langle Px, v(s,x)\rangle + 2\langle Px, x\rangle \geq m\cdot\|x\|^2.$$

This reveals the following meaning of (**E**): for $s \in \mathbb{C}_-$ and $x \in D(A)$ with $\|x\| = 1$, the quantities $\|v(s,x)\|$ and $\|Px\|$ cannot simultaneously be very small. Verly loosely speaking, $\|Px\|$ being small means that x is close to being an unobservable state, while $\|v(s,x)\|$ being small means that x is close to being an eigenvector of A. Perhaps this could be a starting point for the proof of the conjecture.

3. The proof of the main results. In this section we prove the theorems stated in §1.

Proof of Theorem 1.2. We will prove the following estimate: For any $s \in \mathbb{C}_-$ and any $x \in D(A)$,

$$\frac{1}{|\text{Re } s|^2}\|(sI - A)x\|^2 + \frac{1}{|\text{Re } s|}\|Cx\|^2 \geq \mu\cdot\|\Psi_\infty x\|_{L^2}^2, \tag{3.1}$$

where

$$\frac{1}{\mu} = \frac{1}{2} + \|\Psi_\infty\|^2.$$

Clearly, this implies Theorem 1.2. We denote

$$z = (A - SI)x,$$

and we define $\xi : [0, \infty) \to X$ by $\xi(t) = \mathbb{T}_t x$. Then

$$\begin{aligned}\dot{\xi}(t) &= \mathbb{T}_t A x \\ &= \mathbb{T}_t(sx + z) \\ &= s\xi(t) + \mathbb{T}_t z,\end{aligned}$$

whence

$$\xi(t) = e^{st}x + \int_0^t e^{s(t-\sigma)}\mathbb{T}_\sigma z \, d\sigma.$$

Without loss of generality we may assume that $x \in D(A^2)$ (by density in X_1) so that $z \in D(A)$. Then, using (1.4),

$$\begin{aligned}(\Psi_\infty x)(t) &= C\xi(t) \\ &= e^{st}Cx + \int_0^t e^{s(t-\sigma)}C\mathbb{T}_\sigma z \, d\sigma \\ &= e^{st}Cx + (e_x * \Psi_\infty z)(t),\end{aligned}$$

where $*$ denotes convolution and e_s denotes the function

$$e_s(t) = e^{st}$$

If we use the following well-known property of convolutions:

$$\|u * v\|_{L^2} \le \|u\|_{L^1} \cdot \|v\|_{L^2},$$

we obtain

$$\begin{aligned}\|\Psi_\infty x\|_{L^2} &\le \|e_s\|_{L^2} \cdot \|Cx\| + \|e_s\|_{L^1} \cdot \|\Psi_\infty z\|_{L^2} \\ &\le \frac{1}{\sqrt{2|\operatorname{Re} s|}}\|Cx\| + \frac{1}{|\operatorname{Re} s|}\|\Psi_\infty\| \cdot \|z\|.\end{aligned}$$

Using that $(\alpha a + \beta b)^2 \le (\alpha^2 + \beta^2)(a^2 + b^2)$, we get

$$\|\Psi_\infty x\|_{L^2}^2 \le \left(\frac{1}{2} + \|\Psi_\infty\|^2\right)\left[\frac{1}{|\operatorname{Re} s|^2}\|z\|^2 + \frac{1}{|\operatorname{Re} s|}\|Cx\|^2\right],$$

which is the same as (3.1). □

For the proof of Theorem 1.6 we need the following lemma.

LEMMA 3.1. *Let $\tilde{A}$ be the generator of a strongly continuous semigroup of operators on a Banach space Z. If $\|(sI - \tilde{A})^{-1}\|$ is bounded on $\rho(\tilde{A})$, then $Z = \{0\}$ (the trivial space).*

Proof. For any $s \in \rho(\tilde{A})$ we have

$$\|(sI - \tilde{A})^{-1}\| \ge \frac{1}{d(s, \sigma(\tilde{A}))},$$

where d denotes distance, see, e.g., Nagel [13, p. 67]. If $\|(sI - \tilde{A})^{-1}\|$ is bounded then it follows that $\sigma(\tilde{A}) = \emptyset$, so that $(sI - \tilde{A})^{-1}$ is a bounded entire function. By Liouville's theorem, $(SI - \tilde{A})^{-1}$ is constant. We know that $\|(\lambda I - \tilde{A})^{-1}\|$ decays like $1/\lambda$ for big

positive λ (see, e.g., Pazy [14, p. 20]), so that we must have $(sI - \tilde{A})^{-1} = 0$, for any $s \in \mathbb{C}$. Since the range of $(sI - \tilde{A})^{-1}$ is dense in Z, it follows that $Z = \{0\}$. □

Proof of Theorem 1.6. For any $x \in D(A)$ and any $t, \tau \geq 0$ we have $(\Psi_\infty \mathbb{T}_\tau x)(t) = (\Psi_\infty x)(t + \tau)$ (see (1.4)), whence (by integration)

$$\|\Psi_\infty \mathbb{T}_\tau x\|_{L^2} \leq \|\Psi_\infty x\|_{L^2}. \tag{3.2}$$

Since $D(A)$ is dense in X and both sides of (3.2) are continuous functions of x on X, it follows that (3.2) holds for any $x \in X$ and any $\tau \geq 0$.

If we denote $Z = \text{Ker } \Psi_\infty$ (so that Z is a closed subspace of X), then (3.2) implies that Z is invariant under $\mathbb{T}$. Let $\tilde{\mathbb{T}}$ be the restriction of $\mathbb{T}$ to Z, so $\tilde{\mathbb{T}}$ is a strongly continuous semigroup on Z, and let $\tilde{A}$ be the generator of $\tilde{\mathbb{T}}$. It is easy to see that

$$D(\tilde{A}) = D(A) \cap Z, \qquad D(\tilde{A}) \subset \text{Ker } C,$$

and $\tilde{A}$ is the restriction of A to $D(\tilde{A})$.

Now suppose that (**E**) holds. Then for any $s \in \mathbb{C}_-$ and any $x \in D(\tilde{A})$,

$$\frac{1}{|\text{Re } s|^2} \|(sI - \tilde{A})x\|^2 \geq m \cdot \|x\|^2,$$

or, equivalently, for any $s \in \rho(\tilde{A}) \cap \mathbb{C}_-$,

$$\|(sI - \tilde{A})^{-1}\| \leq \frac{1}{\sqrt{m}\,|\text{Re } s|}. \tag{3.3}$$

Since $\tilde{\mathbb{T}}$ is exponentially stable, $\|(sI - \tilde{A})^{-1}\|$ is defined and bounded on some half-plane $\{s \in \mathbb{C} | \text{ Re } s > \alpha\}$, where $\alpha < 0$ (see, e.g., Pazy [14, p. 20]). Together with (3.3) we obtain that $\|(sI - \tilde{A})^{-1}\|$ is bounded on all of $\rho(\tilde{A})$. By Lemma 3.1, $Z = \{0\}$, so that $\Psi_\infty x \neq 0$ for any nonzero $x \in X$. This implies the condition in Definition 1.5. □

Remark 3.2. As mentioned in §1, approximate observability on $[0, \infty)$ is equivalent with Ker $\Psi_\infty = \{0\}$. Indeed, if Ker $\Psi_\infty \neq \{0\}$ then introducing $Z, \tilde{\mathbb{T}}$, and $\tilde{A}$ as in the proof of Theorem 1.6, we have $D(\tilde{A}) \neq \{0\}$ and $\Psi_\infty x = 0$ for any $x \in D(\tilde{A}) \subset D(A)$, so that the system is not approximately observable on $[0, \infty)$. The converse is obvious.

Remark 3.3. In Theorem 1.6 (and in its proof) the estimate (**E**) can be replaced by the following slightly less restrictive condition.

There are functions $p_1, p_2 : \mathbb{C}_- \to [0, \infty)$ such that: (1) p_1 is bounded on any half-plane $\{s \in \mathbb{C} | \text{ Re } s \leq \alpha\}$ with $\alpha < 0$, (2) for any $s \in \mathbb{C}_-$ and any $x \in D(A)$,

$$p_1(s) \cdot \|(sI - A)x\|^2 + p_2(s) \cdot \|Cx\|^2 \geq \|x\|^2.$$

We now turn to the proof of Theorem 1.7. As already mentioned in §1, this theorem is essentially due to Rodman; more precisely, its dual is a direct consequence of [17, Thm. 7.1.2 and Ex. 7.1]. (We have been made aware of [17] after working out our own proof.) In the hope that our proof might contain useful ideas for dealing with the case of unbounded A and C, we reproduce it below. We need three lemmas.

LEMMA 3.4. *Let X be a Hilbert space. We denote by $l^\infty(X)$ the space of bounded X-valued sequences with the supremum norm, and by $c_0(X)$ the subspace of $l^\infty(X)$ consisting of sequences convergent to zero. We introduce the factor space*

$$\mathcal{F}(X) = l^\infty(X)/c_0(X)$$

and we denote by π the canonical surjection from $l^\infty(X)$ onto $\mathcal{F}(X)$. We endow $\mathcal{F}(X)$ with the factor norm

$$\|\pi(z)\|_{\mathcal{F}(X)} = \inf_{c\in c_0(X)} \|z + c\|_{l^\infty(X)}.$$

Then the factor norm can be computed by

$$\|\pi(z)\|_{\mathcal{F}(X)} = \limsup_{n\to\infty} \|z_n\| \tag{3.4}$$

(z_n is the nth term of z) and $\mathcal{F}(X)$ is a Banach space with this norm.

The proof of (3.4) is easy and we leave it to the reader. For the fact that $\mathcal{F}(X)$ is a Banach space see, e.g., Brown and Pearcy [2, p. 222].

LEMMA 3.5. *With the notation of Lemma* 3.4, *suppose $T \in \mathcal{L}(X)$. We define $\tilde{T} \in \mathcal{L}(l^\infty(X))$ by termwise application of T, i.e., for any $z \in l^\infty(X)$,*

$$(\tilde{T}z)_n = Tz_n \quad \forall n \in \mathbb{N}.$$

Then there is a unique $\tilde{\tilde{T}} \in \mathcal{L}(\mathcal{F}(X))$ such that the diagram

$$\begin{array}{ccc} l^\infty(X) & \xrightarrow{\tilde{T}} & l^\infty(X) \\ \downarrow \pi & & \downarrow \pi \\ \mathcal{F}(X) & \xrightarrow{\tilde{\tilde{T}}} & \mathcal{F}(X) \end{array}$$

commutes. Moreover,

$$\begin{aligned} \|\tilde{\tilde{T}}\| = \|\tilde{T}\| = \|T\|, \\ \sigma(\tilde{\tilde{T}}) = \sigma(\tilde{T}) = \sigma(T). \end{aligned} \tag{3.5}$$

It is routine to verify all the statements in this lemma. In view of the fact that a similar construction, with a strongly continuous semigroup in place of T, appears in Nagel [13, pp. 21 and 78], we omit the details.

LEMMA 3.6. *With the notation of Lemma* 3.5, *if N is a closed $\tilde{T}$-invariant subspace of $l^\infty(X)$ that contains $c_0(X)$, i.e.,*

$$c_0(X) \subset N \subset l^\infty(X), \qquad \tilde{T}N \subset N,$$

then $\mathcal{N} = \pi(N)$ is a closed $\tilde{\tilde{T}}$-invariant subspace of $\mathcal{F}(X)$.

Proof. Let $\xi \in \mathcal{N}$, so $\xi = \pi(z)$, where $z \in N$. We have $\tilde{T}z \in N$, so $\pi(\tilde{T}z) \in \mathcal{N}$. By the commutativity of the diagram in Lemma 3.5, $\tilde{\tilde{T}}\xi = \pi(\tilde{T}z)$, so that $\mathcal{N}$ is $\tilde{\tilde{T}}$-invariant.

Put $N^c = l^\infty(X)\backslash N$ and $\mathcal{N}^c = \mathcal{F}(X)\backslash\mathcal{N}$. It follows from $c_0(X) \subset N$ that $\pi(N^c) = \mathcal{N}^c$. Since N^c is open, but the open mapping theorem $\mathcal{N}^c$ is open too, so that $\mathcal{N}$ is closed. □

Proof of Theorem 1.7. Let X, Y, A, and C be as in the statement of the theorem and let $\mathbb{T}_t = e^{At}$. Let the Banach spaces $l^\infty(X), c_0(X), \mathcal{F}(X)$, and the surjection π be as in Lemma 3.4. For any operator $T \in \mathcal{L}(X)$, we define $\tilde{T}$ and $\tilde{\tilde{T}}$ as in Lemma 3.5. Then it is easy to see that

$$\tilde{\mathbb{T}}_t = e^{\tilde{A}t}. \tag{3.6}$$

Let P be the observability Gramian of A and C and define

$$N = \{z \in l^\infty(X) | Pz_n \to 0\}.$$

Then N is a closed subspace of $l^\infty(X)$, because it is the kernel of $\pi\tilde{P}$. It is clear that N contains $c_0(X)$. We show that N is $\tilde{A}$-invariant. Let Ψ_∞ be the extended output map of A and C and let $z \in N$. From $\langle Pz_n, z_n\rangle = \|\Psi_\infty z_n\|^2$ we see that $\Psi_\infty z_n \to 0$. By (3.2) we get that for any $\tau \geq 0$, $\Psi_\infty \mathbb{T}_\tau z_n \to 0$, whence $P\mathbb{T}_\tau z_n \to 0$, i.e., $\tilde{\mathbb{T}}_\tau z \in N$. Thus, N is $\tilde{\mathbb{T}}_\tau$-invariant, for any $\tau \geq 0$. By (3.6) this implies that N is $\tilde{A}$-invariant. Now by Lemma 3.6 we conclude that $\mathcal{N} = \pi(N)$ is $\tilde{A}$-invariant.

Let $\mathcal{A}$ be the restriction of $\tilde{\tilde{A}}$ to $\mathcal{N}$. Since we have assumed that $\sigma(A) \subset \mathbb{C}_-$, (3.5) shows that we have $\sigma(\tilde{\tilde{A}}) \subset \mathbb{C}_-$, i.e., the (uniformly continuous) group generated by $\tilde{\tilde{A}}$ is exponentially stable. It follows that the restriction of this semigroup to $\mathcal{N}$ is also exponentially stable. Since the generator of this restriction is $\mathcal{A}$, it follows that

$$\sigma(\mathcal{A}) \subset \mathbb{C}_-. \tag{3.7}$$

Our goal is to show that $\mathcal{N} = \{0\}$. To achieve this, we assume the contrary, i.e., $\mathcal{N}$ is nonzero, and we show that this leads to a contradiction. The approximate point spectrum of a bounded operator on a nonzero Banach space is nonvoid (because it contains the boundary of the spectrum, see, e.g., [13, p. 64]). Applying this to $\mathcal{A}$ and taking (3.7) into account, we get that there exists $\lambda \in \mathbb{C}_-$ and a sequence (η_k) with values in $\mathcal{N}$ such that $\|\eta_k\| = 1$ for any $k \in \mathbb{N}$ and

$$\lim_{k\to\infty} (\lambda I - \mathcal{A})\eta_k = 0. \tag{3.8}$$

We have assumed in the theorem that for every $s \in \mathbb{C}_-$ there is an $m_s > 0$ such that (1.7) holds. In particular, taking $s = \lambda$ and using (2.5), we get that for any $x \in X$,

$$\|(\lambda I - A)x\|^2 - 2\,\mathrm{Re}\,\langle Px, Ax\rangle \geq m_\lambda \cdot \|x\|^2. \tag{3.9}$$

Let $k \in \mathbb{N}$ be such that

$$\|(\lambda I - \mathcal{A})\eta_k\|^2 \leq \tfrac{1}{2} m_\lambda \tag{3.10}$$

(such a k exists by (3.8)). Let $z \in N$ be such that $\eta_k = \pi(z)$. By the commutativity of the diagram in Lemma 3.5 we have

$$(\lambda I - \mathcal{A})\eta_k = \pi((\lambda I - \tilde{A})z),$$

whence by (3.4) and (3.10) we obtain

$$\limsup_{n\to\infty} \|(\lambda I - A)z_n\|^2 \leq \tfrac{1}{2} m_\lambda. \tag{3.11}$$

The fact that $z \in N$ means that the sequence (z_n) is bounded and $Pz_n \to 0$. Hence,

$$\lim_{n\to\infty} \langle Pz_n, Az_n\rangle = 0. \tag{3.12}$$

Now (3.9), (3.11), and (3.12) imply that

$$\limsup_{n\to\infty} \|z_n\|^2 \leq \tfrac{1}{2},$$

whence by (3.4) $\|\eta_k\|^2 \leq \frac{1}{2}$. But the sequence (η_k) was such that $\|\eta_k\| = 1$, which is a contradiction.

Thus we have shown that $\mathcal{N} = \{0\}$. This means that P is bounded from below, i.e., Ψ_∞ is bounded from below, i.e., the system described by (1.2) is exactly observable on $[0, \infty)$. □

Remark 3.7. With the notation of Conjecture 1.3, it is possible to show that if A and C satisfy the estimate (**E**), then the bounded operators $A_b \in \mathcal{L}(X)$ and $C_b \in \mathcal{L}(X, Y)$ defined by

$$A_b = A^{-1}, \qquad C_b = CA^{-1},$$

satisfy a similar estimate, with a possibly different number m_b instead of m. A_b is not stable in general, but it can be verified that C_b is an *infinite time admissible observation operator* for the semigroup generated by A_b, i.e., (1.3) holds with $e^{A_b t}$ instead of $\mathbb{T}_t$ and C_b instead of C. The pair (A_b, C_b) is interesting because the observability Gramian of the system determined by A_b and C_b is the same as for A and C (as is not difficult to verify). Now it almost seems that we could apply Theorem 1.7 to A_b and C_b and thus prove Conjecture 1.3. But a more careful look reveals that this is not the case, because Theorem 1.7 requires A to be stable.

4. Systems with diagonal semigroup on l^2. First we introduce some terminology. A bounded operator D on l^2 is called *diagonal* if it is of the form

$$(\mathrm{D}x)_k = d_k x_k \quad \forall k \in \mathbb{N},$$

where x_k denotes the kth component of x and (d_k) is a bounded sequence of complex numbers. Obviously,

$$\|\mathrm{D}\| = \sup_{k \in \mathbb{N}} |d_k|.$$

A strongly continuous semigroup $\mathbb{T}$ on l^2 is called *diagonal* if $\mathbb{T}_t$ is diagonal for each $t \geq 0$. Then $\mathbb{T}_t$ is given by

$$(\mathbb{T}_t x)_k = e^{\lambda_k t} x_k \quad \forall k \in \mathbb{N}, \tag{4.1}$$

where (λ_k) is a sequence of complex numbers with real parts bounded above. We assume that $\mathbb{T}$ is exponentially stable, which is equivalent to

$$\sup_{k \in \mathbb{N}} \operatorname{Re} \lambda_k < 0. \tag{4.2}$$

The generator A of $\mathbb{T}$ is given by

$$D(A) = \{x \in l^2 | (\lambda_k x_k) \in l^2\}, \tag{4.3a}$$

$$(Ax)_k = \lambda_k x_k \quad \forall k \in \mathbb{N}. \tag{4.3b}$$

Clearly, $\{\lambda_k | k \in \mathbb{N}\}$ is the set of eigenvalues of A. The space denoted in general by X_1 will now be denoted l_1^2. That is, l_1^2 is $D(A)$ with the norm $\|x\|_1 = \|Ax\|_{l^2}$ (we have chosen $\beta = 0$ in (2.1)).

If Y is a Hilbert space and $C \in \mathcal{L}(l_1^2, Y)$, then C is uniquely determined by the sequence (c_k) of vectors in Y defined by

$$c_k = Ce_k, \tag{4.4}$$

where (e_k) is the standard basis sequence of l^2, i.e., $x = x_1e_1 + x_2e_2 + \dots$. It is not trivial to give necessary and sufficient conditions for the admissibility of C, in terms of the sequence (c_k). For finite-dimensional Y, this problem has been solved in Ho and Russell [7] and Weiss [22], [23]. For infinite-dimensional Y, the problem has been almost solved in Hansen and Weiss [4]. Here we mention only the following simple necessary condition for admissibility.

PROPOSITION 4.1. *Let $\mathbb{T}$ be an exponentially stable, diagonal semigroup on l^2, with generator A. Let Y be a Hilbert space and assume that $C \in \mathcal{L}(l_1^2, Y)$ is an admissible observation operator for $\mathbb{T}$, i.e.,* (1.3) *holds. Let the sequences $(\lambda_k), (e_k)$, and (c_k) be as in* (4.1) *and* (4.4). *Then*

$$\frac{1}{|\operatorname{Re} \lambda_k|}\|c_k\|^2 \le 2K \quad \forall k \in \mathbb{N}. \tag{4.5}$$

Proof. Taking in (1.3) $x = e_k$, we get by (4.4)

$$\|c_k\|^2 \int_0^\infty |e^{\lambda_k t}|^2 dt \le K \quad \forall k \in \mathbb{N},$$

which is equivalent to (4.5). □

PROPOSITION 4.2. *With the notation of Proposition* 4.1, *assume that the system determined by A and C is exactly observable on $[0, \infty)$, i.e.,* (1.5) *holds. Then*

$$\frac{1}{|\operatorname{Re} \lambda_k|}\|c_k\|^2 \ge 2\kappa \quad \forall k \in \mathbb{N}. \tag{4.6}$$

Proof. Taking in (1.5) $x = e_k$, we get by (4.4)

$$\|c_k\|^2 \int_0^\infty |e^{\lambda_k t}|^2 dt \ge \kappa \quad \forall k \in \mathbb{N},$$

which is equivalent to (4.6). □

Note the analogy between (4.5) and (4.6).

DEFINITION 4.3. Let (λ_k) be a sequence in $\mathbb{C}_-$. We say that (λ_k) is *properly spaced* if it satisfies

$$\inf_{j,k \in \mathbb{N}, j \ne k} \left| \frac{\lambda_j - \lambda_k}{\operatorname{Re} \lambda_k} \right| = \delta > 0. \tag{4.7}$$

In particular, (4.7) implies that (λ_k) has no multiple values, and no accumulation points in $\mathbb{C}_-$.

The goal of this section is to prove the following theorem.

THEOREM 4.4. *With the notation of Proposition* 4.1, *if (λ_k) is properly spaced then* (4.6) *is equivalent to the estimate* (**E**) *in Theorem* 1.2.

For the proof we need two lemmas.

LEMMA 4.5. *Let (λ_k) be a sequence in $\mathbb{C}_-$ satisfying* (4.7). *Define the function $N : \mathbb{C}_- \to \mathbb{N}$ such that for any $s \in \mathbb{C}_-$, $\lambda_{N(s)}$ is among the closest elements of $\{\lambda_k | k \in \mathbb{N}\}$ to s:*

$$|s - \lambda_{N(s)}| = \min_{k \in \mathbb{N}} |s - \lambda_k|. \tag{4.8}$$

Then for any $s \in \mathbb{C}_-$ and any $k \in \mathbb{N}$ with $k \ne N(s)$, we have

$$\left| \frac{\operatorname{Re} s}{s - \lambda_k} \right| \le 1 + \frac{2}{\delta}. \tag{4.9}$$

Proof. Take $s \in \mathbb{C}_-$ and $k \in \mathbb{N}$ with $k \neq N(s)$. Then

$$|\lambda_{N(s)} - \lambda_k| \leq |s - \lambda_{N(s)}| + |s - \lambda_k|,$$

whence by (4.8)

$$|s - \lambda_k| \geq \tfrac{1}{2}|\lambda_{N(s)} - \lambda_k|.$$

This shows that s is an element of the set S defined by

$$S = \{z \in \mathbb{C}_- m || z - \lambda_k| \geq \tfrac{1}{2}|\lambda_{N(s)} - \lambda_k|\}$$

(S is the complement in $\mathbb{C}_-$ of a disk). Therefore, we have

$$\left| \frac{\text{Re } s}{s - \lambda_k} \right| \leq \sup_{z \in S} \left| \frac{\text{Re } z}{z - \lambda_k} \right|.$$

It is an exercise in elementary calculus to check that this supremum is attained at $z = \lambda_k - \frac{1}{2}|\lambda_{N(s)} - \lambda_k|$, which yields

$$\left| \frac{\text{Re } s}{s - \lambda_k} \right| \leq 1 + 2 \left| \frac{\text{Re } \lambda_k}{\lambda_{N(s)} - \lambda_k} \right|$$

Since, by (4.7), we have

$$\left| \frac{\text{Re } \lambda_k}{\lambda_{N(s)} - \lambda_k} \right| \leq \frac{1}{\delta},$$

we get (4.9). □

LEMMA 4.6. *With the notation of Proposition* 4.1, *assume that* (λ_k) *is properly spaced and let* $N : \mathbb{C}_- \to \mathbb{N}$ *be a function satisfying* (4.8). *For each* $s \in \mathbb{C}_-$, *we define the subspace* $V(s) \subset l^2$ *by*

$$V(s) = \{e_{N(s)}\}^{\perp} \tag{4.10}$$

(*i.e.*, $V(s)$ *is the space spanned by all vectors* e_k *with* $k \neq N(s)$). *We denote by* A_s *the part of* A *in* $V(s)$, *i.e.*,

$$A_s : D(A) \cap V(s) \to V(s)$$

and $A_s x = Ax$ *for any* $x \in D(A) \cap V(s)$.

Then there is a $\mathcal{G} \geq 0$ *such that for any* $s \in \mathbb{C}_-$

$$\|C(sI - A_s)^{-1}\|_{\mathcal{L}(V(s),Y)} \leq \frac{\mathcal{G}}{\sqrt{|\text{Re } s|}}. \tag{4.11}$$

Proof. We have from the resolvent identity

$$(sI - A_s)^{-1} = (-\bar{s}I - A_s)^{-1}[I - (\bar{s} + s)(sI - A_s)^{-1}],$$

whence

$$\begin{aligned} &\|C(sI - A_s)^{-1}\|_{\mathcal{L}(V(s),Y)} \\ &\quad \leq \|C(-\bar{s}I - A_s)^{-1}\|_{\mathcal{L}(V(s),Y)} \cdot [1 + 2|\text{Re } s| \cdot \|(sI - A_s)^{-1}\|_{\mathcal{L}(V(s))}]. \end{aligned} \tag{4.12}$$

If P_s denotes the orthogonal projection from l^2 onto $V(s)$, then $(-\bar{s}I - A_s)^{-1} = (-\bar{s}I - A)^{-1}P_s$. Using (2.4) and the fact that $\|P_s\| = 1$, we have

$$\|C(-\bar{s}I - A_s)^{-1}\|_{\mathcal{L}(V(s),Y)} \leq \frac{\mathcal{K}}{\sqrt{|\text{Re } s|}}. \tag{4.13}$$

The operator $|\text{Re } s| \cdot (sI - A_s)^{-1}$ on $V(s)$ is diagonal and its diagonal elements d_k are given by

$$d_k = \frac{|\text{Re } s|}{s - \lambda_k},$$

where $k \neq N(s)$. By (4.9) we have $|d_k| \leq 1 + \frac{2}{\delta}$, whence

$$|\text{Re } s| \cdot \|(sI - A_s)^{-1}\|_{\mathcal{L}(V(s))} \leq 1 + \frac{2}{\delta}. \tag{4.14}$$

Combining (4.12)–(4.14) we get

$$\|C(sI - A_s)^{-1}\|_{\mathcal{L}(V(s),Y)} \leq \frac{\mathcal{K}}{\sqrt{|\text{Re } s|}}\left(3 + \frac{4}{\delta}\right),$$

which is the same as (4.11). □

Note the resemblance between (2.4) and (4.11).

Proof of Theorem 4.4. The implication **(E)** $\Rightarrow$ (4.6) is very easy: Taking in (1.6) $s = \lambda_k$ and $x = e_k$, we get (4.6) with $2\kappa = m$.

Conversely, suppose (4.6) holds but **(E)** is false. This means that there are sequences (s_n) and (z^n) such that $s_n \in \mathbb{C}_-$, $z^n \in D(A)$, $\|z^n\| = 1$ and

$$\frac{1}{|\text{Re } s_n|^2}\|(s_nI - A)z^n\|^2 + \frac{1}{|\text{Re } s_n|}\|Cz^n\|^2 = \epsilon_n^2, \tag{4.15}$$

where $\epsilon_n \geq 0$ and $\epsilon_n \to 0$.

The main idea of the proof is to show that for large n, s_n is almost equal to some eigenvalue of A (which may depend on n), and z^n is almost equal to a corresponding normalized eigenvector. We need a lot of notation: Let the function N and the spaces $V(s)$ be as in (4.8) and (4.10). For any $s \in \mathbb{C}_-$, let P_s denote the orthogonal projection of l^2 onto $V(s)$ and let A_s denote the part of A in $V(s)$ (as in Lemma 4.6). Further, we introduce

$$q^n = \frac{1}{|\text{Re } s_n|}(s_nI - A_{s_n})P_{s_n}z^n, \qquad \alpha_n = z^n_{N(s_n)},$$

so that

$$\frac{1}{|\text{Re } s_n|}(s_nI - A)z^n = e_{N(s_n)}\frac{s_n - \lambda_{N(s_n)}}{|\text{Re } s_n|}\alpha_n + q^n. \tag{4.16}$$

From (4.15) it follows, using that the two terms on the right-hand side of (4.16) are orthogonal, that

$$\|q^n\| \leq \frac{1}{|\text{Re } s_n|}\|(s_nI - A)z^n\| \leq \epsilon_n \tag{4.17}$$

and, by a similar argument,

$$\left| \frac{s_n - \lambda_{N(s_n)}}{\text{Re } s_n} \right| \cdot |\alpha_n| \leq \epsilon_n. \tag{4.18}$$

We have

$$P_{s_n} z^n = |\text{Re } s_n| \cdot (s_n I - A_{s_n})^{-1} q^n.$$

Using (4.17) and the inequality (4.14), obtained in the proof of Lemma 4.6, we get

$$\|P_{s_n} z^n\| \leq \left(1 + \frac{2}{\delta}\right) \epsilon_n,$$

whence $P_{s_n} z^n \to 0$. Since $\|z^n\| = 1$, it follows that $\|(I - P_{s_n}) z^n\| \to 1$, i.e.,

$$\lim_{n\to\infty} |\alpha_n| = 1. \tag{4.19}$$

Together with (4.18), this implies

$$\lim_{n\to\infty} \left| \frac{s_n - \lambda_{N(s_n)}}{\text{Re } s_n} \right| = 0.$$

It is now easy to see that

$$\lim_{n\to\infty} \frac{\text{Re } \lambda_{N(s_n)}}{\text{Re } s_n} = 1. \tag{4.20}$$

Now we turn our attention to the second term in (4.15). We have

$$\begin{aligned} \|Cz^n\| &= \|c_{N(s_n)} \alpha_n + C P_{s_n} z^n\| \\ &\geq \|c_{N(s_n)}\| \cdot |\alpha_n| - \|C(s_n I - A_{s_n})^{-1} (s_n I - A_{s_n}) P_{s_n} z^n\| \\ &= \|c_{N(s_n)}\| \cdot |\alpha_n| - |\text{Re } s_n| \cdot \|C(s_n I - A_{s_n})^{-1} q^n\|, \end{aligned}$$

whence by (4.11)

$$\frac{1}{\sqrt{|\text{Re } s_n|}} \|Cz^n\| \geq \frac{1}{\sqrt{|\text{Re } \lambda_{N(s_n)}|}} \|c_{N(s_n)}\| \cdot \left| \frac{\text{Re } \lambda_{N(s_n)}}{\text{Re } s_n} \right|^{1/2} |\alpha_n| - \mathcal{G} \cdot \|q^n\|. \tag{4.21}$$

By (4.19) and (4.20), there is an $n_0 \in \mathbb{N}$ such that for $n \geq n_0$ we have

$$\left| \frac{\text{Re } \lambda_{N(s_n)}}{\text{Re } s_n} \right|^{1/2} |\alpha_n| \geq \frac{1}{2}.$$

This, together with (4.6), (4.17), and (4.21), implies that for any $n \geq n_0$,

$$\frac{1}{\sqrt{|\text{Re } s_n|}} \|Cz^n\| \geq \sqrt{2\kappa} \cdot \frac{1}{2} - \mathcal{G} \cdot \epsilon_n.$$

Since $\epsilon_n \to 0$, it follows that for n sufficiently large we have

$$\frac{1}{|\text{Re } s_n|} \|Cz^n\|^2 \geq \frac{\kappa}{3}.$$

On the other hand, (4.15) implies that for each $n \in \mathbb{N}$,

$$\frac{1}{|\text{Re } s_n|} \|Cz^n\|^2 \leq \epsilon_n^2,$$

which is a contradiction. Therefore, (**E**) must be true. □

5. Riesz bases of complex exponentials.

DEFINITION 5.1. Let H be a separable Hilbert space and let (p_k) be a sequence in H. Then (p_k) is a *Riesz basis* in H if for some (and hence any) orthonormal basis (e_k) in H there is an invertible operator $T \in \mathcal{L}(H)$ such that

$$p_k = Te_k \quad \forall k \in \mathbb{N}.$$

The following two propositions are taken (with minor modifications) from Young [25, pp. 32 and 157].

PROPOSITION 5.2. *Let Z be a Hilbert space and let (p_k) be a sequence in Z. Then the following statements are equivalent:*

(S1) *There is a positive constant K such that for any finite sequence $(x_1, x_2, \dots x_N)$ in $\mathbb{C}$,*

$$\left\| \sum_{k=1}^{N} x_k p_k \right\|^2 \leq K \sum_{k=1}^{N} |x_k|^2. \tag{5.1}$$

(S2) *The Gramian matrix of (p_k), defined by*

$$P_{j,k} = \langle p_k, p_j \rangle \quad \forall j, k \in \mathbb{N}, \tag{5.2}$$

determines a bounded operator on l^2.

We mention that sequences satisfying (S1) are called *Bessel sequences* (see [25, p. 155]). In the second proposition, the statements (S1) and (S2) are replaced by stronger statements.

PROPOSITION 5.3. *Let Z be a Hilbert space and let (p_k) be a sequence in Z. Then the following statements are equivalent:*

(S3) *(p_k) is a Riesz basis in its closed span in Z.*

(S4) *The Gramian matrix of (p_k), defined by* (5.2), *determines an invertible bounded operator on l^2.*

The following theorem explains the connection between sequences of complex exponential functions and linear systems with diagonal semigroup.

THEOREM 5.4. *Let (λ_k) be a sequence in $\mathbb{C}$ satisfying* (4.2) *and let (c_k) be a sequence in a Hilbert space Y. Let (p_k) be the sequence in $L^2([0,\infty), Y)$ defined by*

$$p_k(t) = e^{\lambda_k t} c_k. \tag{5.3}$$

Then the statement (S1) *is equivalent to the following:*

(S5) *Let $\mathbb{T}$ be the diagonal semigroup on l^2 defined by* (4.1) *and let A be its generator. Then the matrix $[c_1, c_2, \dots]$ determines an operator $C \in \mathcal{L}(l^2_1, Y)$ that is admissible for $\mathbb{T}$.*

Moreover, if the above statement is true then the observability Gramian P of A and C (as defined in §2) is given by the Gramian matrix of the sequence (p_k) (as defined in (5.2)).

Proof. First we introduce some notation. Let F be the vector space of sequences in $\mathbb{C}$ with only finitely many nonzero terms. The matrix $[c_1, c_2, \dots]$ defines an operator $C' : F \to Y$. Since F is $\mathbb{T}$-invariant, we can defined $\Psi_\infty : F \to L^2([0,\infty), Y)$ by

$$(\Psi_\infty x)(t) = C' \mathbb{T}_t x \quad \forall x \in F. \tag{5.4}$$

Note that if (e_k) denotes the standard orthonormal basis of l^2, then $c_k = C' e_k$, whence

$$p_k = \Psi_\infty e_k. \tag{5.5}$$

Now suppose that (S1) holds. Using (5.5) and denoting $x = \sum_{k=1}^{n} x_k e_k$, (5.1) becomes

$$\|\Psi_\infty x\|_{L^2}^2 \le K \cdot \|x\|^2 \quad \forall x \in F.$$

Since F is dense in l^2, it follows that Ψ_∞ has a unique continuous extension to l^2, which we denote the same way.

For any $t \ge 0$, define $\Psi_t = P_t \Psi_\infty$, where P_t is the orthogonal projection from $L^2([0,\infty), Y)$ onto $L^2([0,t), Y)$ (considered as a subspace) and put $\Psi = (\Psi_t)_{t\ge 0}$. Then it is easy to check that $(\mathbb{T}, \Psi)$ is an abstract linear observation system, as defined in §2. By the representation theorem mentioned in §2, there is a $C \in \mathcal{L}(l_1^2, Y)$ such that

$$(\Psi_\infty x)(t) = C\mathbb{T}_t x \quad \forall x \in l_1^2. \tag{5.6}$$

Comparing this to (5.4) and taking $t = 0$ (this is possible since $\Psi_\infty x$ is continuous for $x \in l_1^2$) we conclude that C is an extension of C'. It follows that C is determined by the same matrix $[c_1, c_2, \ldots]$. Obviously, C is admissible, so that (S5) holds.

Conversely, suppose that (S5) holds. Then, by assumption, the operator C' defined earlier has a continuous extension $C \in \mathcal{L}(l_1^2, Y)$ and the operator Ψ_∞ defined by (5.6) has a continuous extension to l^2. Let P be the observability Gramian of A and C, i.e., $P = \Psi_\infty^* \Psi_\infty$. The matrix of P is

$$P_{j,k} = \langle Pe_k, e_j \rangle = \langle \Psi_\infty e_k, \Psi_\infty e_j \rangle.$$

Using (5.5), we obtain $P_{j,k} = \langle p_k, p_j \rangle$, i.e., the Gramian matrix of (p_k). Thus, (S2) holds and, by Proposition 5.2, (S1) holds as well. □

The last theorem reduces a (difficult) problem, checking (S1), to another (difficult) problem, checking (S5). However, for verifying (S5), powerful methods are available, especially the Carleson measure criterion and other criteria derived from it, see [4], [7], and [24].

Remark 5.5. An elementary computation shows that if (p_k) is defined by (5.3) then the Gramian matrix of (p_k) (see (5.2)) is given by

$$P_{j,k} = -\frac{\langle c_k, c_j \rangle}{\lambda_k + \overline{\lambda_j}}.$$

COROLLARY 5.6. *With the notation of Theorem 5.4, the sequence* (p_k) *is a Riesz basis in its closed span if and only if the following holds:*

(S6) *The statement* (S5) *is true (i.e.,* C *is admissible for* $\mathbb{T}$*) and the system determined by* A *and* C *is exactly observable on* $[0,\infty)$.

Proof. Suppose that (p_k) is a Riesz basis in its closed span. Then by Proposition 5.3, (S4) holds, which implies (S2). By Proposition 5.2, (S1) holds whence, by Theorem 5.4, (S5) is true. By the "moreover" part of Theorem 5.4, P is given by the Gramian matrix of (p_k). Since (S4) holds, P is invertible, so that (S6) holds.

Conversely, suppose (S6) is true. By Theorem 5.4, P is given by the Gramian matrix of (p_k). Since P is invertible, (S4) holds whence, by Proposition 5.3, (p_k) is a Riesz basis in its closed span. □

Remark 5.7. Theorem 5.4 and Corollary 5.6 can be slightly generalized, as follows: If (λ_k) is not required to satisfy (4.2), only $\lambda_k \in \mathbb{C}_-$, and "admissible" is replaced by "infinite time admissible", then both results are still true, with the same proof. C being infinite time admissible simply means that (1.3) holds. Since $\mathbb{T}$ is not necessarily stable, infinite time admissibility is a stronger condition than admissibility (see also Remark 3.7).

Let (p_k) be a sequence of complex exponentials (as in (5.3)) and suppose that we wish to check whether it is a Riesz basis in its closed span. We have seen that one necessary condition is (S1). Another simple necessary condition is $\inf \|p_k\| > 0$. If Conjecture 1.3 would be true (at least for diagonal semigroups) then for a large class of such sequences, the above two conditions together would be sufficient for the sequence to be a Riesz basis in its closed span. More precisely, we have the following result.

THEOREM 5.8. *Let (λ_k) be a sequence in $\mathbb{C}_-$ satisfying* (4.2) *and properly spaced* (*see* (4.7)). *Let (c_k) be a sequence in a Hilbert space Y and let the sequence (p_k) in $L^2([0,\infty),Y)$ be defined by* (5.3). *We assume that* (S1) *holds and*

$$\|p_k\|^2 \geq \kappa > 0 \quad \forall k \in \mathbb{N}. \tag{5.7}$$

If Conjecture 1.3 *is true for the diagonal semigroup $\mathbb{T}$ defined by* (4.1), *then (p_k) is a Riesz basis in its closed span.*

Proof. By Theorem 5.4, (S5) holds. Let A and C be as in (S5). By an elementary computation, the condition on $\|p_k\|$ in the theorem is the same as (4.6). Since (λ_k) is properly spaced, by Theorem 4.4 the estimate (**E**) holds for A and C. If Conjecture 1.3 is true for $\mathbb{T}$ then the system determined by A and C is exactly observable on $[0,\infty)$. It follows that P, the observability Gramian of A and C, is invertible. By Theorem 5.4, P is given by the Gramian matrix of (p_k). By Proposition 5.3, (S3) holds. □

For certain classes of properly spaced sequences (λ_k), we can show that Conjecture 1.3 is true for the semigroup defined by (4.1). Then, if (c_k) is such that (5.7) holds, Theorem 5.8 can be used to show that (p_k) (defined by (5.3)) is a Riesz basis in its closed span. We will briefly discuss two such classes of sequences: one for which the numbers λ_k lie on a vertical line in $\mathbb{C}_-$, and one for which they lie on the real line. The following theorem (the vertical line case) is a consequence of Ingham's theorem.

THEOREM 5.9. *Let (λ_k) be a properly spaced sequence in $\mathbb{C}_-$ such that for some $\rho < 0$,*

$$\operatorname{Re} \lambda_k = \rho \quad \forall k \in \mathbb{N}. \tag{5.8}$$

Then Conjecture 1.3 *is true for the diagonal semigroup $\mathbb{T}$ defined by* (4.1).

Proof. Let Y be a Hilbert space and let $C : l_1^2 \to Y$ be an admissible observation operator for $\mathbb{T}$. Let (e_k) be the standard basis of l^2 and put $c_k = Ce_k$. Then, by Proposition 4.1 and by (5.8), (c_k) is bounded. Let A be the generator of $\mathbb{T}$ and suppose that A and C satisfy the estimate (**E**). Then, by Theorem 4.4, $\|c_k\|$ is bounded from below by a positive number. Thus, we have the factorization $C = C_0 D$, where C_0 is defined by

$$C_0 e_k = \frac{c_k}{\|c_k\|} \quad \forall k \in \mathbb{N},$$

and D is an invertible bounded diagonal operator on l^2.

There are several ways to show that C_0 is admissible for $\mathbb{T}$. One of them is to use the fact that D commutes with the semigroup and C is admissible:

$$\begin{aligned}\int_0^\infty \|C_0 \mathbb{T}_t x\|^2 dt &= \int_0^\infty \|C \mathbb{T}_t D^{-1} x\|^2 dt \\ &\leq K \cdot \|D^{-1}\|^2 \cdot \|x\|^2.\end{aligned}$$

By Remark 2.10, C_0 is admissible for the (unitary) semigroup generated by $A - \rho I$. Moreover, the system determined by $A - \rho I$ and C_0 is exactly observable on a certain finite

interval $[0, \tau)$. This follows from the vector-valued version of Ingham's theorem (see [25, p. 162]) together with another (much simpler) result appearing in [25, p. 157]. By Remark 2.10, the system determined by A and C_0 is also exactly observable on $[0, \tau)$, and hence on $[0, \infty)$: for some $k > 0$ and any $x \in l_1^2$,

$$\int_0^\infty \|C_0 \mathbb{T}_t x\|^2 dt \geq k \cdot \|x\|^2. \tag{5.9}$$

To show that the system determined by A and C is exactly observable on $[0, \infty)$, we will again use the fact that D commutes with the semigroup. For any $x \in l_1^2$ we have, using (5.9),

$$\begin{aligned} \int_0^\infty \|C\mathbb{T}_t x\|^2 dt &= \int_0^\infty \|C_0 \mathbb{T}_t D x\|^2 dt \\ &\geq k \cdot \|Dx\|^2 \\ &\geq \frac{k}{\|D^{-1}\|^2} \|x\|^2. \end{aligned}$$ □

Our result for real λ_k depends on a lemma from complex function theory.

LEMMA 5.10. *Let $a > 1$ and let the function f on the unit circle be defined by*

$$f(z) = \sum_{k=-\infty}^{\infty} \frac{z^k}{a^k + a^{-k}}. \tag{5.10}$$

Then f is real and

$$\min_{|z|=1} f(z) > 0.$$

Sketch of the proof. The Laurent series in (5.10) permits us to extend f to a holomorphic function in the annulus $\mathcal{D}$ defined by

$$\mathcal{D} = \{z \in \mathbb{C} | a^{-1} < |z| < a\}.$$

Put $g(z) = \text{Re } f(z)$, so g is harmonic in $\mathcal{D}$. We have, for $a^{-1} < r < a$ and $\theta \in \mathbb{R}$,

$$g(re^{i\theta}) = \frac{1}{2} + \sum_{k=1}^{\infty} \frac{r^k + r^{-k}}{a^k + a^{-k}} \cos k\theta.$$

Since $\frac{1}{2} + \sum_{k=1}^{\infty} \cos k\theta$ is the Fourier expansion of π times the Dirac measure at $\theta = 0$, we get that the boundary distribution of g on the circle with radius a is $a\pi\delta_a$, where δ_a is the Dirac measure at $z = a$. Similarly, on the circle with radius a^{-1}, g becomes $a^{-1}\pi\delta_{a^{-1}}$. Thus, the boundary distribution of g is a positive measure (supported at two points). It follows that $g(z) \geq 0$ in $\mathcal{D}$. Since g is not constant, by the minimum principle it cannot attain its infimum in $\mathcal{D}$, so that in fact $g(z) > 0$ in $\mathcal{D}$. Since for $|z| = 1$ we have $f(z) = g(z)$, f satisfies the desired estimate. □

It seems to us that it is difficult to give a purely computational proof for the above lemma. The following theorem considers a very restricted class of systems, where the eigenvalues of the generator are negative and grow exponentially and the output space is one-dimensional.

THEOREM 5.11. *Let* $\alpha > 1$ *and let* (λ_k) *be defined by*

$$\lambda_k = -\alpha^k \quad \forall k \in \mathbb{N}. \tag{5.11}$$

Then Conjecture 1.3 *is true for the diagonal semigroup* $\mathbb{T}$ *defined by* (4.1) *and for* $Y = \mathbb{C}$.

Proof. Let $C : l^2_1 \to \mathbb{C}$ be an admissible observation operator for $\mathbb{T}$. Let (e_k) be the standard basis of l^2 and put $c_k = Ce_k$. Then, by Proposition 4.1 and by (5.11), there is a $K > 0$ such that $|c_k|^2 \leq 2K\alpha^k$ holds for any $k \in \mathbb{N}$. Let A be the generator of $\mathbb{T}$ and suppose that A and C satisfy the estimate (**E**). Since (λ_k) is properly spaced, by Theorem 4.4 (the easy direction) there is a $\kappa > 0$ such that $|c_k|^2 \geq 2\kappa\alpha^k$ holds for any $k \in \mathbb{N}$. Thus, we have the factorization $C = C_0 D$, where C_0 is defined by

$$C_0 e_k = \alpha^{k/2}, \quad \forall k \in \mathbb{N},$$

and D is an invertible bounded diagonal operator on l^2.

To show that C_0 is admissible for $\mathbb{T}$, we can use either the Carleson measure criterion (see [7]) or the fact that D commutes with the semigroup and C is admissible (as in the proof of Theorem 5.9).

Let us show that the system determined by A and C_0 is exactly observable on $[0, \infty)$. If P_0 denotes the observability Gramian of A and C_0, we have to show that this (nonnegative) operator is bounded from below. By Theorem 5.4 (the "moreover" part) and by Remark 5.5, the matrix of P_0 is

$$P_{j,k} = \frac{\alpha^{k/2}\alpha^{j/2}}{\alpha^k + \alpha^j} = \frac{1}{\alpha^{(k-j)/2} + \alpha^{(j-k)/2}}.$$

Thus, P_0 is a Toeplitz (or convolution) operator, with generating bilateral sequence (τ_k) given by

$$\tau_k = \frac{1}{\alpha^{k/2} + \alpha^{-k/2}}.$$

We define a continuous function φ on the unit circle by

$$\varphi(z) = \sum_{k=-\infty}^{\infty} \tau_k z^k.$$

It is easy to show that for any $x = (x_k) \in l^2$, denoting $\hat{x}(z) = \sum_{k=1}^{\infty} x_k z^k$, we have

$$\langle P_0 x, x \rangle = \frac{1}{2\pi} \int_0^{2\pi} \varphi(e^{i\theta}) \cdot |\hat{x}(e^{i\theta})|^2 d\theta,$$

so that if φ is bounded from below (by a positive number) then P_0 is bounded from below. Denoting $a = \alpha^{1/2}$, we see that φ becomes f defined in (5.10), so that by Lemma 5.10, φ is bounded from below.

Thus we have shown that the system determined by A and C_0 is exactly observable on $[0, \infty)$. To show that the system determined by A and C has the same property, we can now apply the exact same argument as in the last part of the proof of Theorem 5.9. □

It seems that Theorems 5.9 and 5.11 can be generalized in many directions, for example, the eigenvalues λ_k can be moved by amounts not exceeding (in absolute value) a small factor times their real part.

Acknowledgments. The idea of Remark 2.12 is due to Prof. J. E. Lagnese. A conversation with D. Shea was very helpful in regard to Lemma 5.10 and Theorem 5.11.

REFERENCES

[1] A. Bensoussan, *On the general theory of exact controllability for skew symmetric operators*, preprint, INRIA, 1991.
[2] A. Brown and C. Pearcy, *Introduction to Operator Theory* I, Graduate Texts in Mathematics Vol. 55, Springer-Verlag, New York, 1977.
[3] S. Dolecki and D. L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.
[4] S. Hansen and G. Weiss, *The operator Carleson measure criterion for admissibility of control operators for diagonal semigroups on l^2*, Systems Control Lett., 16 (1991), pp. 219–227.
[5] M. L. J. Hautus, *Controllability and observability conditions for linear autonomous systems*, Ned. Akad. Wetenschappen, Proç. Ser. A, 72 (1969), pp. 443–448.
[6] L. F. Ho, *Observabilité frontière de l'équation des ondes*, C.R. Acad. Sci. Paris, 302 (1986), pp. 443–446.
[7] L. F. Ho and D. L. Russell, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–640.
[8] V. Komornik, *Rapid boundary stabilization of the wave equation*, SIAM J. Control Optim., 29 (1991), pp. 197–221.
[9] J. E. Lagnese, *The Hilbert uniqueness method:* A retrospective, in Optimal Control of Partial Differential Equations, Proc. of the Conference in Irsee, Germany, April 1990, K. H. Hoffmann, W. Krabs, eds., Lecture Notes in Control and Inform. Sci., Vol. 149, Springer-Verlag, Berlin, 1991, pp. 158–181.
[10] I. Lasiecka, J. L. Lions, and R. Triggiani, *Non homogeneous boundary value problems for second order hyperbolic operators*, J. Math. pures et appl., 65 (1986), pp. 149–192.
[11] J. L. Lions, *Controlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués. Tome* 1, *Controlabilité Exacte*, Recherches en Mathématiques Appliquées Vol. 8, Masson, Paris, 1988.
[12] ———, *Exact controllability, stabilization and perturbations for distributed parameter systems*, SIAM Review, 30 (1988), pp. 1–68.
[13] R. Nagel, ed., *One-parameter Semigroups of Positive Operators*, Lecture Notes in Math., Vol. 1184, Springer-Verlag, Berlin, 1986.
[14] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci., Vol. 44, Springer-Verlag, New York, 1983.
[15] V. M. Popov, *Hyperstability of Control Systems*, Editura Academiei, Bucharest, 1966. (In Romanian.) Springer-Verlag, Berlin, 1973. (In English.)
[16] R. M. Reid and D. L. Russell, *Boundary control and stability of linear water waves*, SIAM J. Control Optim., 23 (1985), pp. 111–121.
[17] L. Rodman, *An Introduction to Operator Polynomials*, Operator Theory: Advances and Applications Vol. 38, Birkhäuser Verlag, Basel, 1989.
[18] D. L. Russell, *Review of "Controlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués," by J. L. Lions*, Bull. Amer. Math. Soc., 22 (1990), pp. 353–356.
[19] D. Salamon, *Infinite dimensional systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[20] ———, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.
[21] E. D. Sontag, *Mathematical Control Theory; Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.
[22] G. Weiss, *Admissibility of input elements for diagonal semigroups on l^2*, Systems Control. Lett., 10 (1988), pp. 79–82.
[23] ———, *Admissible observation operators for linear semigroups*, Israel J. Math. 65 (1989) pp. 17–43.
[24] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems, Proc. of the Conference in Vorau, Austria, July 1990, F. Kappel, K. Kunish, and W. Schappacher, eds., Birkhäuser Verlag, Basel, 1991, pp. 367–378.
[25] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 24–34, January 1994

002

EXACT CONTROLLABILITY FOR THE SCHRÖDINGER EQUATION*

ELAINE MACHTYNGIER†

Abstract. The exact controllability of Schrödinger equation in bounded domains with Dirichlet boundary condition is studied. Both the boundary controllability and the internal controllability problems are considered. Concerning the boundary controllability, the paper proves the exact controllability in $H^{-1}(\Omega)$ with L^2-boundary control. On the other hand, the exact controllability in $L^2(\Omega)$ is proved with L^2-controls supported in a neighborhood of the boundary. Both results hold for an arbitrarily small time. The method of proof combines both the HUM (Hilbert Uniqueness Method) and multiplier techniques.

Key words. Schrödinger equation, boundary controllability, interior controllability

AMS subject classifications. 35B45, 93B05

1. Introduction and main results. During the last few years, various authors have obtained exact controllability results for the wave and plate equations and for the elasticity system (see, e.g., Haraux [7]; Jaffard [9]; Lagnese [10]; Lagnese and Lions [11]; Lions [13], [14], [15]; Russell [20]; Zuazua [22], [23], and the bibliography therein).

As Rauch [19] pointed out, since every solution of Schrödinger equation

$$iy_t + \Delta y = 0$$

is also solution of the plate equation

$$y_{tt} + \Delta^2 y = 0,$$

we can get exact controllability results for the Schrödinger equation from those corresponding to plate equations.

The object of this work is to study directly the exact controllability problem for the Schrödinger equation by adapting the, by now well known, method (see for instance Lions [14]) which combines HUM (Hilbert Uniqueness Method) and multiplier techniques.

Let us formulate precisely the exact boundary controllability problem for the Schrödinger equation. Let Ω be an open bounded set of $\mathbf{R}^n$ with boundary $\Gamma = \partial\Omega$ of class C^3. We consider a partition (Γ_o, Γ_1) of Γ given by

$$\Gamma_o = \Gamma(x^o) = \{x \in \Gamma; m(x).\nu(x) > 0\}, \tag{1.1}$$

$$\Gamma_1 = \{x \in \Gamma; m(x).\nu(x) \leq 0\}, \tag{1.2}$$

where x^o is a fixed point of $\mathbf{R}^n$, $m(x) = x - x^o$, and $\nu(x)$ is the unit normal vector to Γ at $x \in \Gamma$ pointing towards the exterior of Ω, and "." denotes the scalar product in $\mathbf{R}^n$.

Let us consider the following Schrödinger equation with nonhomogeneous boundary conditions:

$$\begin{cases} iy_t + \Delta y = 0 & \text{in} \quad Q = \Omega \times (0,T) \\ y = \begin{cases} v \\ 0 \end{cases} & \begin{matrix} \text{on} \quad \Sigma_0 = \Gamma_o \times (0,T) \\ \text{on} \quad \Sigma_1 = \Gamma_1 \times (0,T) \end{matrix} \\ y(0) = y^o & \text{in} \quad \Omega. \end{cases} \tag{1.3}$$

* Received by the editors December 9, 1991; accepted for publication (in revised form) April 28, 1992.

† Instituto de Matemática, Universidade Federal do Rio de Janeiro, CP 68530, CEP 21945, Rio de Janeiro, Brasil.

For any initial data $y^o \in H^{-1}(\Omega)$ and $v \in L^2(\Sigma_o)$ there exists a unique weak solution of (1.3) in the class $y \in C([0,T];H^{-1}(\Omega))$. This solution is defined by transposition (see Lions and Magenes [16]).

Our main result is as follows.

THEOREM 1.1. *Let $T > 0$, Γ_o be defined by* (1.1) *and $\Sigma_o = \Gamma_o \times (0,T)$. Then, for any $y^o \in H^{-1}(\Omega)$, there exists $v \in L^2(\Sigma_o)$ such that the unique solution $y \in C([0,T];H^{-1}(\Omega))$ of* (1.3) *satisfies $y(T) = 0$.*

Let us now consider the exact controllability problem when the control acts in a subset of Ω.

We assume that the open subset $\omega \subset \Omega$ is a neighborhood of $\overline{\Gamma}_o$, that is, $\omega = \Omega \cap \mathcal{O}$ where $\mathcal{O}$ is an open set of $\mathbf{R}^n$ such that $\overline{\Gamma}_o \subset \mathcal{O}$, and let χ_ω be the characteristic function of ω.

Let us consider the following nonhomogeneous Schrödinger equation:

$$\text{(1.4)} \qquad \begin{cases} iy_t + \Delta y = h\chi_\omega & \text{in} \quad Q = \Omega \times (0,T) \\ y = 0 & \text{on} \quad \Sigma = \Gamma \times (0,T) \\ y(0) = y^o & \text{in} \quad \Omega. \end{cases}$$

It is well known (see Cazenave and Haraux [3], Cazenave [4]) that for any initial data $y^o \in L^2(\Omega)$ and $h \in L^2(\omega \times (0,T))$, there exists a unique weak solution of (1.4) in the class $y \in C([0,T];L^2(\Omega))$.

Concerning this problem our main exact controllability result is as follows.

THEOREM 1.2. *Let $T > 0$ and $\omega \subset \Omega$ be a neighborhood of $\overline{\Gamma}_o$. Then for any $y^o \in L^2(\Omega)$, there exists $h \in L^2(\omega \times (0,T))$ such that the solution of the problem* (1.4) *satisfies $y(T) = 0$.*

Recently some results connected to Theorems 1.1 and 1.2 have been obtained by Lebeau [12] and Fabre [5]. Lebeau, in [12], generalized the results about the exact controllability of hyperbolical problems of Bardos, Lebeau, and Rauch in [1] to the Schrödinger equation. He proved that when Ω is analytic and Γ_o (one open part of the boundary) controls geometrically Ω (cf. [1]); Theorem 1.1 holds. Our result is more restrictive in the sense that it only applies when $\Gamma_o = \Gamma(x^o)$, excluding many subsets Γ_o which geometrically control Ω and are not of the form (1.1). However, our technique presents several advantages:

(i) it applies provided Ω is of class C^3; and

(ii) it provides explicit estimates for the constants on the essential a priori estimations.

Fabre in [5] proved that the boundary control could be obtained as the limit of the internal controls with support in a neighborhood of the boundary of ε thickness by letting ε go to zero.

In the Hilbert spaces $L^2(\Omega)$ and $H_o^1(\Omega)$ we will consider the following inner products:

$$\langle u,v\rangle_{L^2(\Omega)} = \text{Re} \int_\Omega u(x)\overline{v(x)}dx, \quad \forall u,v \in L^2(\Omega)$$

and

$$\langle u,v\rangle_{H_o^1(\Omega)} = \text{Re} \int_\Omega \nabla u(x)\overline{\nabla v(x)}dx, \quad \forall u,v \in H_o^1(\Omega),$$

where $\bar{\bullet}$ denotes the conjugate of $\bullet$.

The rest of the paper is divided into two parts. In §2 we prove the Boundary Controllability Theorem 1.1. In §3 we prove the Internal Controllability Theorem 1.2.

Let us finally mention that the results of this paper were announced in Machtyngier [17].

2. Exact boundary controllability. To prove the Boundary Controllability Theorem 1.1 we need the following proposition.

PROPOSITION 2.1. *For every $T > 0$, there exist $c_i = c_i(T, \Omega) > 0$ $(i = 1, 2)$, such that*

$$\int_0^T \int_{\Gamma_o} \left| \frac{\partial \varphi}{\partial \nu} \right|^2 d\Sigma \le c_1 \|\varphi^o\|^2_{H_o^1(\Omega)} \tag{2.1}$$

and

$$\|\varphi^o\|^2_{H_o^1(\Omega)} \le c_2 \int_0^T \int_{\Gamma_o} \left\| \frac{\partial \varphi}{\partial \nu} \right|^2 d\Sigma \tag{2.2}$$

for every solution $\varphi = \varphi(x, t)$ of the problem

$$\begin{cases} i\varphi_t + \Delta\varphi = 0 & in \quad Q = \Omega \times (0, T) \\ \varphi = 0 & on \quad \Sigma = \Gamma \times (0, T) \\ \varphi(0) = \varphi^o & in \quad \Omega \end{cases} \tag{2.3}$$

with $\varphi^o \in H_o^1(\Omega)$.

(In (2.1) and (2.2) "$\partial\varphi/\partial\nu$" denotes the derivative on the direction ν, the normal unit vector to Γ oriented towards the exterior of Ω. We denote by $d\Sigma = d\Gamma\, dt$ the surface measure of Σ.)

Proof of Proposition 2.1. We proceed in several steps.

Step 1. First, we prove an identity for the solution of the problem

$$\begin{cases} i\varphi_t + \Delta\varphi = f & \text{in} \quad Q \\ \varphi = 0 & \text{on} \quad \Sigma \\ \varphi(0) = \varphi^o & \text{in} \quad \Omega. \end{cases} \tag{2.4}$$

LEMMA 2.2. *Let $q = q(x, t) \in C^2(\overline{Q}, \mathbf{R}^n)$. For every solution of* (2.4) *with $f \in \mathcal{D}(Q)$ and $\varphi^o \in \mathcal{D}(\Omega)$, the following identity holds:*

$$\begin{aligned} \frac{1}{2} \int_\Sigma (q.\nu) \left| \frac{\partial \varphi}{\partial \nu} \right|^2 d\Sigma &= \frac{1}{2} \operatorname{Im} \int_\Omega (\varphi q.\nabla\overline{\varphi}) dx \Big|_0^T \\ &+ \frac{1}{2} \operatorname{Im} \int_Q (q_t.\nabla\varphi\overline{\varphi}) dx dt \\ &+ \frac{1}{2} \operatorname{Re} \int_Q (\varphi \nabla(\operatorname{div}_x q).\nabla\overline{\varphi}) dx dt \\ &+ \operatorname{Re} \int_Q \sum_{j,k} \left(\frac{\partial q_k}{\partial x_j} \frac{\partial \overline{\varphi}}{\partial x_k} \frac{\partial \varphi}{\partial x_j} \right) dx dt + \operatorname{Re} \int_Q f\, q.\nabla\overline{\varphi}\, dx dt \\ &+ \frac{1}{2} \operatorname{Re} \int_Q f\overline{\varphi}(\operatorname{div}_x q) dx dt. \end{aligned} \tag{2.5}$$

In Lemma 2.2 we used the notation: $\operatorname{div}_x q = \sum_{j=1}^n (\partial q_j / \partial x_j)$ and

$$\begin{aligned} \frac{1}{2} \operatorname{Im} \int_\Omega (\varphi q.\nabla\overline{\varphi}) dx \Big|_0^T &= \frac{1}{2} \operatorname{Im} \int_\Omega (\varphi(x, T) q(x, T).\overline{\nabla\varphi(x, T)}) dx \\ &\quad - \frac{1}{2} \operatorname{Im} \int_\Omega (\varphi(x, 0) q(x, 0).\overline{\nabla\varphi(x, 0)}) dx. \end{aligned}$$

Proof of Lemma 2.2. We multiply (2.4) by $q.\nabla\overline{\varphi} + \frac{1}{2}\overline{\varphi}(\operatorname{div}_x q)$ and take the real part. Identity (2.5) holds by integration by parts. □

Step 2. For the proof of (2.1) we choose $q = q(x) \in C^2(\overline{Q}, \mathbf{R}^n)$ such that $q = \nu$ on Γ (see Lions [14] for the construction of this vector field) and $f \equiv 0$ in Q in the identity (2.5) and we obtain

$$\begin{aligned}\frac{1}{2}\int_\Sigma \left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\Sigma \le k_1 \|q\|_{L^\infty(\Omega)}(\|\varphi(T)\|^2_{L^2(\Omega)} + \|\nabla\varphi(T)\|^2_{L^2(\Omega)} \\ +\|\varphi(0)\|^2_{L^2(\Omega)} + \|\nabla\varphi(0)\|^2_{L^2(\Omega)}) \\ +k_2\|q\|_{W^{2,\infty}(\Omega)}\int_0^T \|\varphi(t)\|_{L^2(\Omega)}\|\nabla\varphi(t)\|_{L^2(\Omega)}dt \\ +k_3\|q\|_{W^{1,\infty}(\Omega)}\int_0^T \|\nabla\varphi(t)\|^2_{L^2(\Omega)}dt.\end{aligned} \tag{2.6}$$

We know, by classic results of the Schrödinger equation (see Cazenave [4]), that

$$\|\varphi(t)\|_{L^2(\Omega)} = \|\varphi^o\|_{L^2(\Omega)} \quad \forall t \in [0,T] \tag{2.7}$$

and

$$\|\nabla\varphi(t)\|_{L^2(\Omega)} = \|\nabla\varphi^o\|_{L^2(\Omega)} \quad \forall t \in [0,T]. \tag{2.8}$$

Hence, we obtain

$$\frac{1}{2}\int_\Sigma \left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\Sigma \le c_1\|\varphi^o\|^2_{H_o^1(\Omega)} \quad \forall\varphi^o \in \mathcal{D}(\Omega).$$

Since $\mathcal{D}(\Omega)$ is dense in $H_o^1(\Omega)$, the estimate (2.1) holds for every solution of the problem (2.3) with initial data $\varphi^o \in H_o^1(\Omega)$.

Remark 2.3. We remark that this estimate gives $(\partial\varphi/\partial\nu)|_\Sigma \in L^2(\Sigma)$. It is not a consequence of classic trace results.

Step 3. For the proof of (2.2) we choose $q(x,t) = m(x) = x - x^o$ and $f \equiv 0$ in Q in the identity (2.5) and, using (2.7), (2.8), we obtain

$$\frac{1}{2}\int_\Sigma (m.\nu)\left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\Sigma = \frac{1}{2}\,\mathrm{Im}\int_\Omega (\varphi m.\nabla\overline{\varphi})dx\Big|_0^T + T\|\varphi^o\|^2_{H_o^1(\Omega)} \tag{2.9}$$

Furthermore, let $\varepsilon > 0$ such that $(T-\varepsilon) > 0$ and

$$\left|\mathrm{Im}\int_\Omega (\varphi m.\nabla\overline{\varphi})dx\right| \le c_\varepsilon\|\varphi^o\|^2_{L^2(\Omega)} + \varepsilon\|\varphi^o\|^2_{H_o^1(\Omega)}. \tag{2.10}$$

Thus

$$(T-\varepsilon)\|\varphi^o\|^2_{H_o^1(\Omega)} \le \frac{1}{2}\int_{\Sigma_o}(m.\nu)\left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\Sigma + c_\varepsilon\|\varphi^o\|^2_{L^2(\Omega)}. \tag{2.11}$$

Step 4. To conclude the proof it is enough to prove the following estimate:

$$\|\varphi^o\|^2_{L^2(\Omega)} \le K\int_{\Sigma_o}\left|\frac{\partial\varphi}{\partial\nu}\right|^2 d\Sigma. \tag{2.12}$$

We argue by contradiction. If (2.12) is not satisfied for any $K > 0$, there exists a sequence $\{\varphi_n\}$ of solutions of (2.3) such that

$$\|\varphi_n(0)\|_{L^2(\Omega)} = 1 \quad \forall n \in \mathbf{N} \tag{2.13}$$

and

$$\int_{\Sigma_o} \left| \frac{\partial \varphi_n}{\partial \nu} \right|^2 d\Sigma \to 0 \quad \text{as} \quad n \to +\infty. \tag{2.14}$$

From (2.11) we deduce that $\{\varphi_n(0)\}$ is bounded in $H_o^1(\Omega)$ and then

$$\{\varphi_n\} \text{ is bounded in } L^\infty(0,T;H_o^1(\Omega)) \cap W^{1,\infty}(0,T;H^{-1}(\Omega)).$$

Thus, by extracting a subsequence (that we will still note by $\{\varphi_n\}$) we will have

$$\begin{cases} \varphi_n \to \varphi & \text{in } L^\infty(0,T;H_o^1(\Omega)) \quad \text{weak}^* \\ (\varphi_n)_t \to \varphi_t & \text{in } L^\infty(0,T;H^{-1}(\Omega)) \quad \text{weak}^*. \end{cases}$$

The function $\varphi \in L^\infty(0,T;H_0^1(\Omega)) \cap W^{1,\infty}(0,T;H^{-1}(\Omega))$ is clearly a solution of (2.3) and, from the compactness of the embedding (see Simon [21])

$$L^\infty(0,T;H_o^1(\Omega)) \cap W^{1,\infty}(0,T;H^{-1}(\Omega)) \to C([0,T];L^2(\Omega))$$

and (2.13), we deduce

$$\|\varphi(0)\|_{L^2(\Omega)} = 1. \tag{2.15}$$

On the other hand, (2.14) implies

$$\frac{\partial \varphi}{\partial \nu} = 0 \quad \text{on } \Sigma_o,$$

which, combined with (2.3), implies $\varphi \equiv 0$, from Holmgren's Uniqueness Theorem (see Hörmander [8, Chap. V, Thm. 5.3.3] and Lions [14, Chap. I, Thm. 8.2]). This is in contradiction with (2.15). This ends the proof of inequality (2.2). □

Proof of Theorem 1.1. Let us now apply HUM to deduce the exact boundary controllability result.

Let φ be the solution of the problem (2.3) with $\varphi^o \in H_o^1(\Omega)$. From step 2 of Proposition 2.1 we have $(\partial\varphi/\partial\nu)|_\Sigma \in L^2(\Sigma)$.

We consider the problem

$$\begin{cases} iy_t + \Delta y = 0 & \text{in} \quad Q \\ y = \begin{cases} \dfrac{\partial \varphi}{\partial \nu} & \text{on} \quad \Sigma_o \\ 0 & \text{on} \quad \Sigma_1 \end{cases} \\ y(T) = 0 & \text{in} \quad \Omega \end{cases} \tag{2.16}$$

from the following proposition.

PROPOSITION 2.4. *Let $v \in L^2(\Sigma)$. Then, there exists an unique solution*

$$y \in C([0,T];H^{-1}(\Omega)),$$

in the transposition sense, of the problem

$$\begin{cases} iy_t + \Delta y = 0 & in \quad Q \\ y = v & on \quad \Sigma \\ y(0) = 0 & in \quad \Omega. \end{cases} \tag{2.17}$$

Furthermore, the map $v \mapsto y$ is linear and continuous from $L^2(\Sigma)$ into $C([0,T];H^{-1}(\Omega))$.

Proof. We say that $y \in L^\infty(0,T;H^{-1}(\Omega))$ is a solution of (2.17) in the transposition sense if and only if

$$\text{Re} \int_0^T \langle y(t), \overline{f}(t)\rangle_{(H^{-1}(\Omega),H_o^1(\Omega))} dt = \text{Re} \int_\Sigma v \frac{\partial \overline{\theta}}{\partial \nu} d\Sigma \tag{2.18}$$

for every $f \in L^1(0,T;H_0^1(\Omega))$, where $\theta = \theta(x,t)$ is the solution of the problem

$$\begin{cases} i\theta_t + \Delta\theta = f & \text{in} \quad Q \\ \theta = 0 & \text{on} \quad \Sigma \\ \theta(T) = 0 & \text{in} \quad \Omega. \end{cases} \tag{2.19}$$

Applying the identity (2.5) with a vector field $q = \nu$ on Γ and using the following classic result (see Cazenave and Haraux [3, Chap. IV]):

$$\|\theta(t)\|_{H_o^1(\Omega)} \le \|f\|_{L^1(0,T;H_o^1(\Omega))}, \quad \forall t \in [0,T]$$

we obtain

$$\left\| \frac{\partial \theta}{\partial \nu} \right\|_{L^2(\Sigma)} \le c \|f\|_{L^1(0,T;H_o^1(\Omega))}.$$

Hence, we have

$$\begin{aligned} \left| \text{Re} \int_\Sigma v \frac{\partial \overline{\theta}}{\partial \nu} d\Sigma \right| &\le \|v\|_{L^2(\Sigma)} \left\| \frac{\partial \theta}{\partial \nu} \right\|_{L^2(\Sigma)} \\ &\le c \|v\|_{L^2(\Sigma)} \|f\|_{L^1(0,T;H_o^1(\Omega))}. \end{aligned} \tag{2.20}$$

From (2.20), we obtain that the map

$$f \mapsto \text{Re} \int_\Sigma v \frac{\partial \overline{\theta}}{\partial \nu} d\Sigma$$

is linear and continuous from $L^1(0,T;H_o^1(\Omega))$ into $\mathbf{R}$.

Hence, there exists a unique $y \in L^\infty(0,T;H^{-1}(\Omega))$ that satisfies (2.18) for every $f \in L^1(0,T;H_o^1(\Omega))$.

From (2.18) and (2.20) we have

$$\|y\|_{L^\infty(0,T;H^{-1}(\Omega))} \le c \|v\|_{L^2(\Sigma)}. \tag{2.21}$$

Thus, the map $v \mapsto y$ is continuous from $L^2(\Sigma)$ into $L^\infty(0,T;H^{-1}(\Omega))$.

Moreover, $y \in C([0,T];H^{-1}(\Omega))$. Indeed, we consider $\{v_n\}_{n \in N} \subset \mathcal{D}(0,T;C^2(\Gamma))$ such that

$$v_n \to v \text{ strong in } L^2(\Sigma). \tag{2.22}$$

Let y_n be the solution of (2.17) with boundary condition v_n. Since v_n is regular, in particular, we have $y_n \in C([0,T];H^{-1}(\Omega))$.

From (2.21) and (2.22), we have

$$y_n \to y \quad \text{in } L^\infty(0,T;H^{-1}(\Omega)).$$

Since $C([0,T];H^{-1}(\Omega))$ is a closed subspace of $L^\infty(0,T;H^{-1}(\Omega))$, we have

$$y \in C([0,T];H^{-1}(\Omega)).$$

We obtain that the solution of (2.16) is in the class $y \in C([0,T];H^{-1}(\Omega))$.

It is easy to see that, by multiplying (2.16) by $\overline{\varphi}$, taking the real part, and integrating it by parts, the following identity is satisfied:

$$\langle -iy(0), \varphi^o \rangle = \int_{\Sigma_o} \left| \frac{\partial \varphi}{\partial \nu} \right|^2 d\Sigma \quad \forall \varphi^o \in \mathcal{D}(\Omega).$$

Let Λ be a linear continuous operator from $H_o^1(\Omega)$ into $H^{-1}(\Omega)$ defined by

$$\Lambda \varphi^o = -iy(0),$$

where $y = y(x,t)$ is the solution of the problem (2.16).

From Proposition 2.1 we have

$$\langle \Lambda \varphi^o, \varphi^o \rangle \geq c \|\varphi^o\|^2_{H_o^1(\Omega)}.$$

Hence Λ is an isomorphism from $H_o^1(\Omega)$ to $H^{-1}(\Omega)$ and the theorem is proved. Indeed, given $y^o \in H^{-1}(\Omega)$, we choose the control $v = \partial\varphi/\partial\nu$ on Σ_o where φ is the solution of problem (2.3) with initial data $\varphi^o = \Lambda^{-1}(-iy^o)$. □

3. Exact interior controllability. In order to prove Theorem 1.2 we need to establish the following proposition.

PROPOSITION 3.1. *Let $\omega \subset \Omega$ be a neighborhood of $\overline{\Gamma}_o$. Then for every $T > 0$ there exists $c = c(T) > 0$ such that*

$$\|\varphi^o\|^2_{L^2(\Omega)} \leq c \int_0^T \int_\omega |\varphi|^2 dx\, dt \tag{3.1}$$

for every $\varphi = \varphi(x,t)$ solution of (2.3) *with initial data $\varphi^o \in L^2(\Omega)$.*

Proof. We proceed in several steps.

Step 1. First, we remark that the inequality

$$\|\varphi^o\|^2_{H_o^1(\Omega)} \leq c \int_0^T \int_{\hat{\omega}} |\nabla \varphi|^2 dx\, dt \tag{3.2}$$

holds for every solution $\varphi = \varphi(x,t)$ of the problem (2.3) with $\varphi^o \in H_o^1(\Omega)$, where $\hat{\omega} \subset \Omega$ is a neighborhood of $\overline{\Gamma}_o$.

Indeed, applying identity (2.5) with $f \equiv 0$ in Q and a vector field q satisfying

$$\begin{cases} q(x,t) = \nu(x) & \forall (x,t) \in \Gamma_o \times (\varepsilon, T-\varepsilon) \\ q(x,t).\nu(x) \geq 0 & \forall (x,t) \in \Gamma \times (0,T) \\ q(x,0) = q(x,T) = 0 & \forall x \in \Omega \\ q(x,t) = 0 & \forall (x,t) \in (\Omega \backslash \hat{\omega}) \times (0,T) \end{cases}$$

and using Proposition 2.1, we obtain (3.2).

Step 2. The inequality

$$\|\varphi^o\|^2_{H_o^1(\Omega)} \leq c \int_0^T (\|\varphi_t(t)\|^2_{H^{-1}(\omega)} + \|\varphi^o\|^2_{L^2(\Omega)})\, dt \tag{3.3}$$

is obtained in the following way.

From Step 1 we have

$$\|\varphi^o\|^2_{H^1_o(\Omega)} \leq c\int_0^T \|\varphi(t)\|^2_{H^1(\hat{\omega})}dt. \tag{3.4}$$

Moreover, we have the following result.

LEMMA 3.2. *Let $\Omega \subset \boldsymbol{R}^n$ be a regular domain, $f \in H^{-1}(\Omega)$ and let $u \in H^1_o(\Omega)$ be the solution of*

$$\begin{cases} -\Delta u = f & in \quad \Omega \\ u = 0 & on \quad \Gamma = \partial\Omega \end{cases} \tag{3.5}$$

Then, there exists $c > 0$ (which does not depend on f) such that

$$\|u\|^2_{H^1(\hat{\omega})} \leq c[\|f\|^2_{H^{-1}(\omega)} + \|u\|^2_{L^2(\Omega)}] \tag{3.6}$$

where ω and $\hat{\omega}$ are neighborhood of Γ such that $(\Omega \cap \overline{\hat{\omega}}) \subset \omega$.

Proof. Let us consider the function $\eta \in C^\infty(\mathbf{R}^n)$ such that

$$\eta(x) = \begin{cases} 1 & \text{in} \quad \hat{\omega} \\ 0 & \text{in} \quad \Omega/\omega. \end{cases}$$

Thus, the function $v = \eta u$, where u is the solution of (3.5), satisfies

$$\begin{cases} -\Delta v = -(\Delta\eta)u - 2\nabla\eta.\nabla u + \eta f \quad \text{in} \quad \omega \\ v \in H^1_o(\omega). \end{cases} \tag{3.7}$$

Since the operator $-\Delta$ is an isomorphism from $H^1_0(\omega)$ to $H^{-1}(\omega)$, we have

$$\|v\|_{H^1(\omega)} \leq c_1\|f\|_{H^{-1}(\omega)} + c_2\|u\|_{L^2(\Omega)} \tag{3.8}$$

Hence, we obtain (3.6) from (3.8) because

$$\|u\|_{H^1(\hat{\omega})} = \|v\|_{H^1(\hat{\omega})} \leq \|v\|_{H^1(\omega)}.$$

Remark 3.3. We also can use this lemma when ω and $\hat{\omega}$ are neighborhoods of $\overline{\Gamma}_o$ such that $(\Omega \cap \overline{\hat{\omega}}) \subset \omega$.

Thus, by combining (3.4), (3.6), and (2.3) we obtain

$$\|\varphi^o\|^2_{H^1_o(\Omega)} \leq c\int_0^T (\|\varphi_t(t)\|^2_{H^{-1}(\omega)} + \|\varphi(t)\|^2_{L^2(\Omega)})dt. \tag{3.9}$$

From (2.7), (3.9) we conclude the proof of (3.3).

Step 3. From (3.9), proceeding by contradiction and using compactness and Holmgren's Uniqueness Theorem, we obtain the following estimate:

$$\|\varphi^o\|^2_{H^1_0(\Omega)} \leq c\int_0^T \|\varphi_t(t)\|^2_{H^{-1}(\omega)}dt. \tag{3.10}$$

On the other hand, as the operator $-\Delta$ is an isomorphism from $H^1_o(\Omega)$ to $H^{-1}(\Omega)$, from (3.10) we have

$$\|\varphi_t(0)\|^2_{H^{-1}(\Omega)} \leq c\int_0^T \|\varphi_t(t)\|^2_{H^{-1}(\omega)}dt. \tag{3.11}$$

Step 4. Let $\varphi \in C([0,T];H^{-1}(\Omega))$ be the solution of (2.3) with $\varphi(0)=\varphi^o \in H^{-1}(\Omega)$ and define

$$\psi(t)=\int_0^t \varphi(s)ds + X,$$

where

$$\begin{cases} \Delta X = -i\varphi(0) \\ X \in H_o^1(\Omega). \end{cases}$$

Thus, ψ is a solution of problem (2.3) with $\psi(0)=X\in H_o^1(\Omega)$ and $\psi_t=\varphi$. Then, applying (3.11) to ψ we have

$$\|\varphi^o\|^2_{H^{-1}(\Omega)} \le c\int_0^T \|\varphi(t)\|^2_{H^{-1}(\omega)}dt. \tag{3.12}$$

Step 5. From (3.4) and (3.12) we have

$$\|\varphi^o\|^2_{H_o^1(\Omega)} \le c\int_0^T \|\varphi(t)\|^2_{H^1(\omega)}dt = c\|\varphi\|_{L^2(0,T;H^1(\omega))}, \tag{3.13}$$

$$\|\varphi^o\|^2_{H^{-1}(\Omega)} \le c\int_0^T \|\varphi(t)\|^2_{H^{-1}(\omega)}dt = c\|\varphi\|_{L^2(0,T;H^{-1}(\omega))}. \tag{3.14}$$

We are going to prove (3.1) by interpolation. Let us consider the linear operator

$$L: H^{-1}(\Omega) \to L^2(0,T;H^{-1}(\omega))$$

defined by

$$L\varphi(t)=(e^{it\Delta}\varphi)|_\omega.$$

It is clear that

$$\|L\varphi\|_{L^2(0,T;H^{-1}(\omega))} \le C\|\varphi\|_{H^{-1}(\Omega)}.$$

Furthermore, it follows from (3.12) that

$$\|L\varphi\|_{L^2(0,T;H^{-1}(\omega))} \ge c\|\varphi\|_{H^{-1}(\Omega)}.$$

Therefore, we can consider the closed subspace $X_o = L(H^{-1}(\Omega))$ of $L^2(0,T;H^{-1}(\omega))$, and the linear operator $\Pi = L^{-1}$ (since L is an isomorphism $H^{-1}(\Omega)\to X_o$).

Thus

$$\Pi \in \mathcal{L}(X_o, Y_o) \tag{3.15}$$

with $Y_o = H^{-1}(\Omega)$.

Next, we can set $X_1 = X_o \cap L^2(0,T;H^1(\omega))$, and it follows from (3.13) that

$$\Pi \in \mathcal{L}(X_1, Y_1) \tag{3.16}$$

with $Y_1 = H_o^1(\Omega)$.

From (3.15) and (3.16), cf. [16, Thm. 51, p. 27], we have

$$\Pi \in \mathcal{L}([X_o, X_1]_{1/2}, [Y_o, Y_1]_{1/2}).$$

From [16, Thm. 12.3, p. 27], we have

$$[Y_o, Y_1]_{1/2} = L^2(\Omega).$$

Furthermore, by [2, Thm. 5.1.2, p. 107], we have

$$[L^2(0,T;H^1(\omega)), L^2(0,T;H^{-1}(\omega))]_{1/2} = L^2(0,T;[H^1(\omega);H^{-1}(\omega)]_{1/2})$$

and from [16, Thm. 12.4], we obtain

$$[H^1(\omega);H^{-1}(\omega)]_{1/2} = L^2(\omega).$$

Hence, since X_o is a closed subspace of $L^2(0,T;H^{-1}(\omega))$ and X_1 a closed subspace of $L^2(0,T;H^1(\omega))$, using [16, Thm. 15.1, p. 107], we verify that the norm of the space $[X_o, X_1]_{1/2}$ is equivalent to the norm of $L^2(0,T;L^2(\omega))$, and since $\Pi \in \mathcal{L}([X_o, X_1]_{1/2}; L^2(\Omega))$, we have

$$\|\varphi^o\|^2_{L^2(\Omega)} \leq c \int_0^T \int_\omega |\varphi|^2 \, dxdt.$$

This ends the proof of the proposition. □

Proof of Theorem 1.2. We apply HUM to deduce the exact controllability as follows. We define the linear continuous operator from $L^2(\Omega)$ into $L^2(\Omega)$ by

$$\Lambda\varphi^o = -iy(0),$$

where $y = y(x,t)$ is the solution of the problem

$$\text{(3.17)} \qquad \begin{cases} iy_t + \Delta y = \varphi\chi_\omega & \text{in} \quad Q \\ y = 0 & \text{on} \quad \Sigma \\ y(T) = 0 & \text{in} \quad \Omega \end{cases}$$

and φ is the solution of (2.3) with initial data $\varphi^o \in L^2(\Omega)$.

It is easy to see that, by multiplying (3.17) by $\overline{\varphi}$, taking the real part and integrating it by parts, one following identity is satisfied:

$$\text{(3.18)} \qquad \langle \Lambda\varphi^o, \varphi^o \rangle = \int_o^T \int_\omega |\varphi|^2 \, dx \, dt, \quad \forall \varphi^o \in L^2(\Omega).$$

By combining estimate (3.1) and identity (3.18) we deduce that Λ is an isomorphism from $L^2(\Omega)$ to $L^2(\Omega)$. Hence, given $y^o \in L^2(\Omega)$, we choose the control $h = \varphi|_\omega$ where φ is the solution of (2.3) with initial data $\varphi^o = \Lambda^{-1}(-iy^o)$. Then the proof is finished. □

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Contrôle et stabilisation dans les problèmes hyperboliques*, Appendix 2, in J. L. Lions, Lecture Notes of the Collège de France, Vol. 1, Masson, Paris, 1988.

[2] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Springer-Verlag, New York, 1976.

[3] T. Cazenave and A. Haraux, *Introduction aux problèmes d'évolution semilinéaires* in, Mathématiques et Applications, No. 1, Ellipses, 1990.

[4] T. Cazenave, *An introduction to nonlinear Schrödinger equations*, in Textos de Métodos Matemáticos da Universidade Federal do Rio Janeiro, No. 22, 1989.
[5] C. Fabre, *Comportement au voisinage du bord des solutions de quelques équations d'évolution linéaires. Application à certains problèmes de contrôlabilité exacte et perturbations singulières*, Thèse de l'Université Paris VI, November, 1990.
[6] ———, *Quelques résultats de contrôlabilité exacte de l'équation de Schrödinger. Application à l'équation des plaques vibrantes*, C.R. Acad. Sci. Paris, 321(1991), pp. 61–66.
[7] A. Haraux, *Series lacunaires et contrôle semi interne es vibrations d'une plaque rectangulaire*, J. Math. Pures et Appl., 68(1989), pp. 457–465.
[8] L. Hörmander, *Linear Partial Differential Operators*; Springer-Verlag, Berlin, New York, 1969.
[9] S. Jaffard, *Contrôle interne exact des vibrations d'une plaque carrée*, C.R. Acad. Sci. Paris, 307(1988), pp. 759–762.
[10] J. Lagnese, *Uniform asymptotic energy estimates for solutions of the equation of dynamic plane elasticity with nonlinear dissipation at the boundary*, Nonlinear Anal. 16(1991), pp. 35–54.
[11] J. Lagnese and J.L. Lions, *Modelling, Analysis and Control of Thin Plates*, Masson, Collection de Recherches en Mathematiques Appliquées 6, Paris, 1988.
[12] G. Lebeau - *Contrôle de l'équation de Schrödinger*, preprint.
[13] J.L. Lions, Contrôlabilité exacte de systèmes distribués; C.R. Acad. Sci. Paris, 302(1986), pp. 471–475.
[14] ———, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués; Tome* 1, *Contrôlabilité Exacte*, Collection de Recherches en Mathematiques Appliquées 8, Masson, Paris, 1988.
[15] ———, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.
[16] J.L. Lions and E. Magenes, *Problèmes aux Limites Non Homogènes et Applications*, Dunod, Paris, 1968.
[17] E. Machtyngier, *Contrôlabilité exacte et stabilisation frontière de l'équation de Schrödinger*, C.R. Acad. Sci. Paris, 310(1990), pp. 806–811.
[18] ———, *Controlabilidade exata e estabilização da equação de Schrödinger*, in Teses de Doutorado No. 20 do Instituto de Matemática da Universidade Federal do Rio de Janeiro, 1991.
[19] J. Rauch, private communication, 1990.
[20] D.L. Russel, *Controllability and stabilization theory for linear partial differential equations. Recent progress and open questions*, SIAM Rev., 20(1978), pp. 639–739.
[21] J. Simon, *Compact sets in the space* $L^p(0,T;B)$; Ann. Mat. Pura Appl. (IV), CXLVI (1987), pp. 65–96.
[22] E. Zuazua, *An Introduction to the exact controllability for distributed systems*, in Textos e Notas No. 44, CMAF da Universidade de Lisboa, 1990.
[23] E. Zuazua, *Controlabilidad Exacta y Estabilización de la ecuación de ondas*, in Textos de Métodos Matemáticos da Universidade Federal do Rio de Janeiro, No. 23, 1992.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 35–50, January 1994

003

APPROXIMATE BOUNDARY CONTROLLABILITY FOR THE WAVE EQUATION IN PERFORATED DOMAINS*

DOINA CIORANESCU†, PATRIZIA DONATO‡, AND ENRIQUE ZUAZUA§

Abstract. This paper considers the wave equation in a perforated domain with holes of size $r(\epsilon)$ distributed with ϵ-periodicity, with the assumption that there exists a neighborhood of the exterior boundary without holes. The following question is asked: Is it possible to approximately control the wave equation in the perforated domain in such a way that when ϵ goes to zero the exact controllability of the limit system is obtained? Two main theorems give a positive answer to this question when $r(\epsilon)$ is the critical size that transforms at the limit the operator $(\partial^2/\partial t^2) - \Delta$ into $(\partial^2/\partial t^2) - \Delta + \mu$, where μ is a positive measure. In the first theorem, in a suitable sense, $L^2(\Omega) \times H^{-1}(\Omega)$ is approximated by the space of the admissible data for the exact controllability in the perforated domain. In the second, it is shown that the limit control, supported only by the exterior boundary of the perforated domain, is such that the related state in the perforated domain goes at the limit to the equilbrium state at the time T.

Key words. approximate controllability, wave equation, perforated domains, homogenization

AMS subject classifications. 93B05, 93C20, 35L05, 35B27

1. Introduction and main results. We consider here some questions related to the approximate and exact Dirichlet boundary controllability of the wave equation in perforated domains.

Let Ω be a bounded domain in $\mathbb{R}^n, n \geq 2$ with boundary $\partial\Omega$ of class C^2. Let Ω_ϵ be the domain obtained by removing from Ω a set S_ϵ of closed smooth subsets (the "holes"), i.e., $\Omega_\epsilon = \Omega \backslash S_\epsilon$.

We assume that the measure of the set of holes goes to zero as the parameter ϵ tends to zero, and we are interested in describing the asymptotic behavior of the system under some geometrical assumptions on the domains Ω_ϵ.

Let us consider the wave equation in the perforated domain $\Omega_\epsilon \times (0,T)$ for a given $T > 0$

$$\begin{cases} y_\epsilon'' - \Delta y_\epsilon = 0 & \text{in } \Omega_\epsilon \times (0,T) \\ y_\epsilon(0) = y_\epsilon^0 & \text{in } \Omega_\epsilon \\ y_\epsilon'(0) = y_\epsilon^1 & \text{in } \Omega_\epsilon \end{cases} \tag{1.1}$$

with Dirichlet boundary conditions

$$y_\epsilon = v_\epsilon \quad \text{on } \Sigma_\epsilon = \partial\Omega_\epsilon \times (0,T), \tag{1.2}$$

where $\partial\Omega_\epsilon = \partial\Omega \cup \partial S_\epsilon$.

The exact controllability problem for system (1.1)–(1.2) consists of finding $T > 0$ large enough such that for every initial data $\{y_\epsilon^0, y_\epsilon^1\}$ in a given space there exists a control v_ϵ such that the solution of (1.2) satisfies

$$y_\epsilon(T) = y_\epsilon'(T) = 0.$$

* Received by the editors November 18, 1991; accepted for publication (in revised form) May 29, 1992. This work is part of the Project Eurhomogenization SC1*-CT91-0732 of the Programme SCIENCE of the Commission of the European Community.

† Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Tour 55-65, 5ème étage, 4 place Jussieu, 75252 Paris Cedex 05, France.

‡ Istituto di Matematica, Facoltà di Scienze M.F.N., Università di Salerno, 84081 Baronissi (Salerno), Italy.

§ Departamento de Matemática Aplicada, Universidad Complutense, 28040 Madrid, Spain. The work of this author was partially supported by project PB90-0245 of the "Dirección General de Investigación Científica y Técnica (MEC-España)."

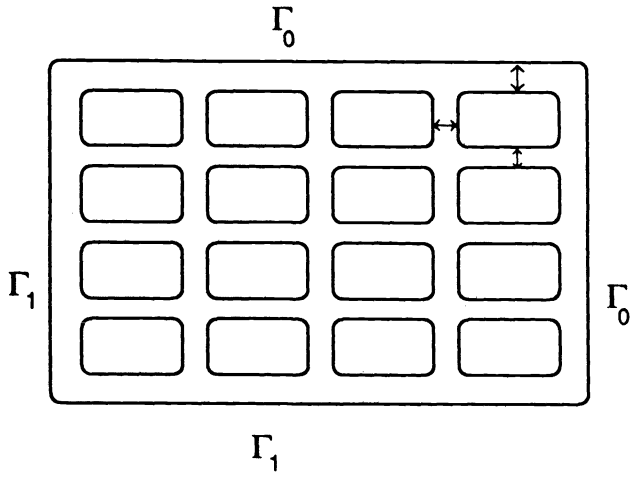


FIG. 1

The Hilbert Uniqueness Method (HUM) (see Lions [7]) furnishes such an exact control $v_\epsilon \in L^2(\Sigma_\epsilon)$ for T large enough ($T >$ diam Ω =diameter of Ω) and for every pair $\{y_\epsilon^0, y_\epsilon^1\} \in L^2(\Omega_\epsilon) \times H^{-1}(\Omega_\epsilon)$. This control minimizes the $L^2(\Sigma_\epsilon)$-norm among all the exact controls for (1.1)–(1.2). Moreover, in [3] we constructed special exact controls for which, under suitable geometric assumptions on S_ϵ (see (1.4)), the asymptotic behavior of the system as $\epsilon \to 0$ can be described.

Let us point out that these controls constructed in [3] are supported by the external boundary $\partial\Omega$ and also by the boundary of *all* the holes. Hence, they are not realistic from a practical point of view.

The support of the control can be restricted to a "large enough" subset Γ_o^ϵ of $\partial\Omega_\epsilon$. More precisely, as a consequence of the results proved by Bardos, Lebeau, and Rauch in [1], in order to have such an exact controllability property in $L^2(\Omega_\epsilon) \times H^{-1}(\Omega_\epsilon)$ with L^2-controls supported on Γ_0^ϵ and in time T_0, Γ_0^ϵ and T_0 must satisfy the following geometric property: Any generalized ray of geometric optics meets Γ_0^ϵ in a time $T \leq T_0$.

Therefore, if S_ϵ contains a sole "nontrapping" hole (for instance, a star-shaped hole), then exact controllability in $L^2(\Omega_\epsilon) \times H^{-1}(\Omega_\epsilon)$ can be achieved with L^2-controls by acting only on the exterior boundary. However, if S_ϵ contains more than one hole, in order to control the trapped generalized rays the exact control must also have a part supported on the boundary of the holes (see Fig. 1).

In practice, what is natural (and realistic) is to try to control system (1.1) by acting only on the external boundary $\Sigma = \partial\Omega \times (0,T)$. Thus, instead of (1.2) we must add to equations (1.1) the following boundary conditions:

$$(1.2)' \qquad \begin{cases} y_\epsilon = v_\epsilon & \text{on } \partial\Omega \times (0,T) \\ y_\epsilon = 0 & \text{on } \partial S_\epsilon \times (0,T). \end{cases}$$

System (1.1)–(1.2)′ is approximately controllable. More precisely, applying HUM again (see [7, Chap. 1]), it can be shown that if $T >$ diam Ω, there exists a Hilbert space F_ϵ such that, for every $\{y_\epsilon^0, y_\epsilon^1\}$ satisfying

$$\{y_\epsilon^1, -y_\epsilon^0\} \in F_\epsilon',$$

there exists an exact control $v_\epsilon \in L^2(\Sigma)$ so that the solution of (1.1)–(1.2)′ satisfies

$$y_\epsilon(T) = y'_\epsilon(T) = 0. \tag{1.3}$$

Moreover F'_ϵ is dense in $H^{-1}(\Omega_\epsilon) \times L^2(\Omega_\epsilon)$ and, as we saw above, F'_ϵ is strictly contained in $H^{-1}(\Omega_\epsilon) \times L^2(\Omega_\epsilon)$ as soon as S_ϵ contains more than one hole. On the other hand, F'_ϵ cannot be characterized in a usable way.

However, when letting ϵ go to zero, under suitable assumptions on the holes S_ϵ, the limit system of the wave equation is a second-order (in time) hyperbolic equation in the whole domain Ω. The exact controllability of this limit system can be achieved in $L^2(\Omega) \times H^{-1}(\Omega)$ with L^2-controls supported on the boundary of Ω. Therefore, the following general question arises: Is it possible to approximately control system (1.1) in such a way that when ϵ goes to zero we obtain the exact controllability of the limit system? This general question can be formulated more precisely in the following two ways.

The first question is: Given T large enough ($T >$ diam Ω) and $\{y^0, y^1\} \in L^2(\Omega) \times H^{-1}(\Omega)$, find approximate initial data $\{y^0_\epsilon, y^1_\epsilon\}$ such that $\{y^1_\epsilon, -y^0_\epsilon\} \in F'_\epsilon$ and

$$\{y^0_\epsilon, y^1_\epsilon\} \to \{y^0, y^1\}$$
$$y_\epsilon \to y$$
$$v_\epsilon \to v,$$

as $\epsilon \to 0$ (in a sense to be made precise), where v_ϵ is the exact control given by HUM for (1.1)–(1.2)′, y is the solution of the limit wave equation in the whole of Ω with initial data $\{y^0, y^1\}$, and v is the exact Dirichlet control for the limit state y so that $y(T) = y'(T) = 0$. Roughly, this question asks whether F'_ϵ converges to $L^2(\Omega) \times H^{-1}(\Omega)$ as ϵ goes to zero.

A second natural question, complementary to the first one, can also be posed. Suppose that we are given $\{y^0_\epsilon, y^1_\epsilon\} \in L^2(\Omega_\epsilon) \times H^{-1}(\Omega_\epsilon)$ so that $\{y^0_\epsilon, y^1_\epsilon\}$ converges in some sense to $\{y^0, y^1\} \in L^2(\Omega) \times H^{-1}(\Omega)$ (in a sense that will be made precise). As shown before, we cannot construct an exact L^2-control with support in $\partial\Omega \times (0,T)$ only. The question is as follows: Can we construct a Dirichlet boundary control v_ϵ supported only by $\partial\Omega \times (0,T)$ such that, if y_ϵ satisfies (1.1), (1.2)′, then

$$y_\epsilon(T) \to 0$$
$$y'_\epsilon(T) \to 0?$$

We give here a positive answer to these two questions under some assumptions on the geometry of the holes.

Our first assumption is as follows:

$$\left\{\begin{array}{l} \text{There exists a sequence of test functions } w_\epsilon \text{ such that} \\ \quad \text{(i) } w_\epsilon \in H^1(\Omega) \cap L^\infty(\Omega), \qquad \|w_\epsilon\|_{L^\infty(\Omega)} \le M_0 \\ \quad \text{(ii) } w_\epsilon = 0 \quad \text{on } S_\epsilon \\ \quad \text{(iii) } w_\epsilon \rightharpoonup 1 \quad \text{weakly in } H^1(\Omega) \text{ and almost everywhere in } \Omega \\ \quad \text{(iv) } -\Delta w_\epsilon = \mu_\epsilon - \gamma_\epsilon \quad \text{where} \quad \mu_\epsilon, \gamma_\epsilon \in H^{-1}(\Omega) \text{ with} \\ \qquad \mu_\epsilon \to \mu \quad \text{strongly in } H^{-1}(\Omega), \mu \in L^\infty(\Omega) \text{ and} \\ \qquad < \gamma_\epsilon, u_\epsilon >_\Omega = 0 \quad \text{for any } u_\epsilon \in H^1_0(\Omega) \text{ such that } u_\epsilon = 0 \text{ on } S_\epsilon. \end{array}\right. \tag{1.4}$$

This assumption implies that, at the limit, $-\Delta : H^1_0(\Omega_\epsilon) \to H^{-1}(\Omega_\epsilon)$ is transformed into $-\Delta + \mu I : H^1_0(\Omega) \to H^{-1}(\Omega)$ (for details see Cioranescu and Murat [6]).

The second assumption is the following:

$$\text{There exists } \alpha > 0 \text{ such that } d(S_\epsilon, \partial\Omega) \ge \alpha \quad \forall \epsilon > 0, \tag{1.5}$$

where $d(S_\epsilon, \partial\Omega)$ is the distance between S_ϵ and $\partial\Omega$.

Remark 1.1. This hypothesis signifies that we have a "safety zone"

$$D_\alpha = \{x \in \Omega : d(x, \partial\Omega) \leq \alpha\}$$

around $\partial\Omega$ where there are no holes. Moreover, it is easy to check that if Ω satisfies hypotheses (1.4) and (1.5) then the support of μ is contained in $\Omega \backslash D_\alpha$, the holes being located in $\Omega \backslash D_\alpha$ only.

We observe also that it is not restrictive to assume that $w_\epsilon \equiv 1$ in $D_{\alpha-\delta}$ for a fixed $\delta \in]0, \alpha[$. In fact, if not, we can define $\overline{w}_\epsilon = \psi + (1 - \psi) w_\epsilon$ where $\psi \in C^\infty(\Omega)$ satisfies

$$\psi \equiv 0 \quad \text{in } \Omega \backslash D_\alpha, \quad \psi \equiv 1 \quad \text{in } \Omega \backslash D_{\alpha-\delta}, \quad 0 \leq \psi \leq 1$$

and it is obvious that hypothesis (1.4) is still satisifed with $\overline{w}_\epsilon$ instead of w_ϵ. Consequently, in the following we will assume that $w_\epsilon \equiv 1$ in $D_{\alpha-\delta}$ for a fixed $\delta \in]0, \alpha[$.

Remark 1.2. Condition (1.4), without the assumption $\mu \in L^\infty(\Omega)$, was introduced by Cioranescu and Murat in [6] in the context of elliptic homogenization and ensures that the measure of the holes is asymptotically small enough. Note that if μ does not belong to $L^\infty(\Omega)$, at the limit we get the elliptic operator $-\Delta + \mu I : V \to V'$ with $V = H_0^1(\Omega) \cap L^2(\Omega; d\mu)$ (see [6] for the elliptic case and Cioranescu et al. [2] for the wave equation). The assumption $\mu \in L^\infty(\Omega)$ is necessary in the present paper to provide the exact controllability for the homogenized wave equation.

Let us now consider the wave equation related to the limit elliptic operator $-\Delta + \mu I$ in the domain $\Omega \times (0, T)$, with $T > 0$ and Dirichlet boundary condition

$$\begin{cases} y'' - \Delta y + \mu y = 0 \quad \text{in } \Omega \times (0, T) \\ y = v \quad \text{on } \partial\Omega \times (0, T) \\ y(0) = y^0 \quad \text{in } \Omega \\ y'(0) = y^1 \quad \text{in } \Omega \end{cases} \tag{1.6}$$

with $\{y^0, y^1\} \in L^2(\Omega) \times H^{-1}(\Omega)$.

If $T > \operatorname{diam} \Omega$, applying HUM to (1.6) we get the existence of a control $v \in L^2(\Sigma)$ such that the solution of (1.6) satisfies

$$y(T) = y'(T) = 0.$$

At this point, we use the fact that $\mu \in L^\infty(\Omega)$. Indeed, when μ is only a measure in $H^{-1}(\Omega)$, no exact boundary controllability result is known.

In order to state the main results of this paper we need to introduce the quasi-extension operator P_ϵ defined in [2]:

$$P_\epsilon \psi = w_\epsilon \tilde{\psi} \quad \forall \psi \in L^2(\Omega_\epsilon) \tag{1.7}$$

where $\tilde{\psi}$ is the extension of ψ by zero in the holes S_ϵ and w_ϵ are the test functions from hypothesis (1.4).

This operator extends as follows to an operator P_ϵ defined on $H^{-1}(\Omega_\epsilon)$:

$$\langle P_\epsilon \psi, \varphi \rangle_{W^{-1,q}(\Omega), W_0^{1,q/(q-1)}(\Omega)} = \langle \psi, w_\epsilon \varphi \rangle_{H^{-1}(\Omega_\epsilon), H_0^1(\Omega_\epsilon)}.$$

This operator satisfies

$$\begin{cases} P_\epsilon \in \mathcal{L}(H^{-1}(\Omega_\epsilon); W^{-1,q}(\Omega)) \\ \|P_\epsilon\|_{\mathcal{L}(H^{-1}(\Omega_\epsilon); W^{-1,q}(\Omega))} \leq C_q \end{cases}$$

for any $q \in (1, n/(n-1))$ (for details we refer the reader to [2]).

We are now able to formulate the main results. The following theorem answers the first question.

THEOREM 1.3. *Let y be the solution of system* (1.6) *with* $\{y^0, y^1\} \in L^2(\Omega) \times H^{-1}(\Omega)$ *and* $v \in L^2(\Sigma)$ *the exact control of* (1.6) *given by* HUM. *Let* y_ϵ *be the solution of*

$$\text{(1.8)} \qquad \begin{cases} y_\epsilon'' - \Delta y_\epsilon = 0 \quad \text{in } \Omega_\epsilon \times (0,T) \\ y_\epsilon = v \quad \text{on } \partial\Omega \times (0,T) \\ y_\epsilon = 0 \quad \text{on } \partial S_\epsilon \times (0,T) \\ y_\epsilon(T) = y_\epsilon'(T) = 0 \quad \text{in } \Omega_\epsilon. \end{cases}$$

Then, under hypotheses (1.4) *and* (1.5)

$$\{y_\epsilon'(0), -y_\epsilon(0)\} \in F_\epsilon'$$

and, moreover,

$$\text{(1.9)} \qquad \begin{cases} \text{(i)} & \widetilde{y_\epsilon} \to y \quad \text{strongly in } L^2(\Omega \times (0,T)) \\ \text{(ii)} & \widetilde{y_\epsilon(0)} \to y^0 \quad \text{strongly in } L^2(\Omega) \\ \text{(iii)} & P_\epsilon y_\epsilon'(0) \to y^1 \quad \text{strongly in } W^{-1,q}(\Omega) \end{cases}$$

for any $q \in (1, n/(n-1))$.

In the following, if $g \in H^{-1}(\Omega)$, we denote $g|_{\Omega_\epsilon}$ by $\in H^{-1}(\Omega_\epsilon)$ the functional defined by

$$\langle g|\Omega_\epsilon, \varphi\rangle_{H^{-1}(\Omega_\epsilon), H_0^1(\Omega_\epsilon)} = \langle g, \tilde{\varphi}\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \quad \forall \varphi \in H_0^1(\Omega_\epsilon).$$

Let us now answer the second question.

THEOREM 1.4. *Let* $\{y^0, y^1\} \in L^2(\Omega) \times H^{-1}(\Omega)$ *and* y *be the solution of* (1.6) *where* v *is an exact Dirichlet control. Let* y_ϵ *be the solution of system* (1.1) *with initial data* $\{y_\epsilon^0, y_\epsilon^1\} = \{y^0|_{\Omega_\epsilon}, y^1|_{\Omega_\epsilon}\}$ *and boundary data*

$$\begin{cases} y_\epsilon = v \quad \text{on } \partial\Omega \times (0,T) \\ y_\epsilon = 0 \quad \text{on } \partial S_\epsilon \times (0,T). \end{cases}$$

Then

$$\text{(1.10)} \qquad \begin{cases} \text{(i)} & \widetilde{y_\epsilon} \to y \quad \text{strongly in } L^2(\Omega \times (0,T)) \\ \text{(ii)} & P_\epsilon y_\epsilon' \to y' \quad \text{strongly in } L^2(0,T; W^{-1,q}(\Omega)) \end{cases}$$

and, moreover,

$$\text{(1.11)} \qquad \begin{cases} \text{(i)} & \widetilde{y_\epsilon}(T) \to 0 \quad \text{strongly in } L^2(\Omega) \\ \text{(ii)} & P_\epsilon y_\epsilon'(T) \to 0 \quad \text{strongly in } W^{-1,q}(\Omega) \end{cases}$$

for all $q \in (1, n/(n-1))$.

Remark 1.5. Regarding the sequence $\{\tilde{y}_\epsilon\}$, it would be interesting to have strong convergence in $C^0([0,T]; L^2(\Omega))$; this is still an open problem.

On the other hand, Theorems 1.3 and 1.4 are essentially homogenization results for the wave equation with nonhomogeneous Dirichlet boundary conditions in a perforated domain provided with a safety zone. We do not know if the theorems are true in domains without this property. A possible approach to this question is the following.

Suppose that (1.5) is not satisfied. Let Ω^1 be an open bounded set such that $\overline{\Omega} \subset \Omega^1$, and consider the new perforated domain $\Omega_\epsilon^1 = \Omega^1 \backslash S_\epsilon$. Clearly (1.4) and (1.5) hold.

Given $(y^0, y^1) \in L^2(\Omega) \times H^{-1}(\Omega)$ we may extend them to $(z^0, z^1) \in L^2(\Omega^1) \times H^{-1}(\Omega^1)$. Applying Theorem 1.4 to (z^0, z^1) in the cylinder $\Omega^1 \times (0,T)$, we find solutions z_ϵ of the wave equation in $\Omega_\epsilon^1 \times (0,T)$ with initial data (z^0, z^1) such that

$$(\tilde{z}_\epsilon(T), P_\epsilon z_\epsilon'(T)) \to (0,0) \quad \text{in } L^2(\Omega^1) \times W^{-1,q}(\Omega^1)$$

for every $1 < q < n/(n-1)$.

Now, clearly,

$$v_\epsilon = z_{\epsilon|_{\partial\Omega \times (0,T)}}$$

are approximate controls for the wave equation in $\Omega_\epsilon \times (0,T)$ such that $y_\epsilon = z_{\epsilon|_{\Omega_\epsilon \times (0,T)}}$ is a solution of (1.1)–(1.2)′ that satisfies

$$(\tilde{y}_\epsilon(T), P_\epsilon y_\epsilon'(T)) \to (0,0) \quad \text{in } L^2(\Omega) \times W^{-1,q}(\Omega).$$

However, we do not know the optimal regularity of v_ϵ and the smallest space in which they are uniformly bounded. This seems to be an interesting open problem.

Remark 1.6. As a consequence of the results by Bardos, Lebeau, and Rauch [1], system (1.6) is exactly controllable in $L^2(\Omega) \times H^{-1}(\Omega)$ with L^2-controls supported in a subset Γ_0 of $\partial\Omega$ in a time T if the geometric control property is satisfied. Theorems 1.3 and 1.4 can be easily adapted to this situation. The same convergence results hold for y_ϵ satisfying, instead of the boundary conditions of Theorems 1.3 and 1.4,

$$\begin{cases} y_\epsilon = v & \text{on } \Gamma_0 \times (0,T) \\ y_\epsilon = 0 & \text{on } (\partial\Omega \backslash \Gamma_0) \times (0,T) \\ y_\epsilon = 0 & \text{on } \partial S_\epsilon \times (0,T), \end{cases}$$

where $v \in L^2(\Gamma_0 \times (0,T))$ is the control of the limit system (1.6). In this case condition (1.5) may be relaxed to the following:

$$\text{There exists } \alpha > 0 \text{ such that } d(S_\epsilon, \Gamma_0) \geq \alpha \quad \forall \epsilon > 0.$$

Remark 1.7. In the case of the wave equation with oscillating coefficients, similar approximate controllability results are given in Cioranescu, Donato, and Zuazua [5] in the framework of the classical homogenization (the coefficients are ϵ-periodic). They are proved when assuming that the coefficients are constant in a neighborhood of $\partial\Omega$. This assumption is of the same type as (1.5): Around $\partial\Omega$ there is a safety zone where no oscillation occurs.

Remark 1.8. We mention also that a strong convergence result of solutions has been proved in [4] in the case of special controls introduced in [3].

The rest of the paper is organized as follows. The proofs of Theorems 1.3 and 1.4 lay on homogenization results for the wave equation with homogeneous Dirichlet boundary conditions in perforated domains given in Cioranescu, Donato, Murat, and Zuazua [2] and on some convergence results from Cioranescu, Donato, and Zuazua [3]. We recall them in §2 where we also give some complementary results. Section 3 is devoted to the proof of the main results. In §4 we give some examples.

2. Homogenization results for the wave equation. Consider the wave equation

$$\begin{cases} u_\epsilon'' - \Delta u_\epsilon = f_\epsilon & \text{in } \Omega_\epsilon \times (0,T) \\ u_\epsilon = 0 & \text{in } \partial\Omega_\epsilon \times (0,T) \\ u_\epsilon(0) = u_\epsilon^0 & \text{in } \Omega_\epsilon \\ u_\epsilon'(0) = u_\epsilon^1 & \text{in } \Omega_\epsilon \end{cases} \tag{2.1}$$

with data $\{u_\epsilon^0, u_\epsilon^1\} \in H_0^1(\Omega_\epsilon) \times L^2(\Omega_\epsilon)$, $f_\epsilon \in L^1(0,T; L^2(\Omega_\epsilon))$. We recall here some results concerning the asymptotic behaviour of u_ϵ as $\epsilon \to 0$. We refer to [2] for the proofs and for further details.

2.1. Convergence results in Ω. We have the following convergence results.

PROPOSITION 2.1 (see [2]). *Assume hypothesis* (1.4) *and suppose that the data* $\{u_\epsilon^0, u_\epsilon^1\}$ *and* f_ϵ *are such that*

$$\begin{cases} \{\tilde{u}_\epsilon^0, \tilde{u}_\epsilon^1\} \rightharpoonup \{u^0, u^1\} & \text{weakly in } H_0^1(\Omega) \times L^2(\Omega) \\ \tilde{f}_\epsilon \rightharpoonup f & \text{weakly in } L^1(0,T;L^2(\Omega)). \end{cases}$$

Then

$$\begin{cases} \tilde{u}_\epsilon \rightharpoonup u & \text{weakly } * \text{ in } L^\infty(0,T;H_0^1(\Omega)) \\ \tilde{u}_\epsilon' \rightharpoonup u' & \text{weakly } * \text{ in } L^\infty(0,T;L^2(\Omega)) \end{cases}$$

and

$$\begin{cases} \tilde{u}_\epsilon(t) \rightharpoonup u(t) & \text{weakly in } H_0^1(\Omega) \times L^2(\Omega) \\ \tilde{u}_\epsilon'(t) \rightharpoonup u'(t) & \text{weakly in } L^2(\Omega) \end{cases}$$

for all $t \in [0,T]$, *where* u *satisfies*

$$\begin{cases} u'' - \Delta u + \mu u = f & \text{in } \Omega \times (0,T) \\ u = 0 & \text{on } \partial\Omega \times (0,T) \\ u(0) = u^0 & \text{in } \Omega \\ u'(0) = u^1 & \text{in } \Omega. \end{cases} \tag{2.2}$$

Remark 2.2. Obviously, Proposition 2.1 holds for the backward wave equation.

Under stronger asssumptions on the convergence of the data, we can improve the convergence result for u_ϵ. Namely, we have the following proposition.

PROPOSITION 2.3 (see [2]). *Suppose that*

$$\begin{cases} \text{(i)} & \tilde{f}_\epsilon \to f \quad \text{strongly in } L^2(\Omega \times (0,T)) \\ \text{(ii)} & \tilde{u}_\epsilon^1 \to u^1 \quad \text{strongly in } L^2(\Omega) \\ \text{(iii)} & u_\epsilon^0 \in H_0^1(\Omega_\epsilon) \quad \text{and there exists } g_\epsilon \in H^{-1}(\Omega) \text{ such that} \\ & \qquad -\Delta u_\epsilon^0 = g_\epsilon \quad \text{in } \mathcal{D}'(\Omega_\epsilon) \text{ with } g_\epsilon \to g \text{ strongly in } H^{-1}(\Omega). \end{cases} \tag{2.3}$$

Then, under hypothesis (1.4),

$$\begin{cases} \tilde{u}_\epsilon' \to u' & \text{strongly in } C^0([0,T];L^2(\Omega)), \\ \int_\Omega |\nabla \tilde{u}_\epsilon|^2 dx \to \int_\Omega |\nabla u|^2 dx + < \mu, u^2 > & \text{in } C^0([0,T]), \end{cases} \tag{2.4}$$

where u *is a solution of* (2.2) *with* $u^0 \in H_0^1(\Omega)$ *and such that*

$$-\Delta u^0 + \mu u^0 = g \quad \text{in } H^{-1}(\Omega).$$

Moreover,

$$\tilde{\mu}_\epsilon(x,t) = w_\epsilon(x) u(x,t) + \rho_\epsilon(x,t) \tag{2.5}$$

with

$$\begin{cases} \text{(i)} & \rho_\epsilon' \to 0 \quad \text{strongly in } C^0([0,T];L^2(\Omega)) \\ \text{(ii)} & \nabla \rho_\epsilon \to 0 \quad \text{strongly in } C^0([0,T];L^1(\Omega)). \end{cases} \tag{2.6}$$

Remark 2.4. Let us point out that hypothesis (1.4) implies that

$$\chi_{\Omega_\epsilon} \to 1 \text{ strongly in } L^p(\Omega) \quad \text{for all } p \text{ with } 1 \leq p, < +\infty.$$

Indeed, since $w_\epsilon \chi_{S_\epsilon} \equiv 0$ in Ω, we have

$$\int_\Omega |1 - \chi_{\Omega_\epsilon}|^p dx = m(S_\epsilon) = \int_{S_\epsilon} (1 - w_\epsilon)^2 dx \leq \int_\Omega (1 - w_\epsilon)^2 dx,$$

and the last integral tends to zero since $w_\epsilon \to 1$ strongly in $L^2(\Omega)$. Consequently, if $\{h_\epsilon\} \subset L^2(\Omega)$ with

$$h_\epsilon \rightharpoonup h \quad \text{weakly in } L^2(\Omega),$$

and if $H_\epsilon = h_\epsilon|_{\Omega_\epsilon}$, then

$$\tilde{H}_\epsilon \rightharpoonup h \quad \text{weakly in } L^2(\Omega).$$

Remark 2.5. From (2.5) and Remark 1.1 it follows that

$$\widetilde{u_\epsilon}(x,t) = u(x,t) + \rho_\epsilon(x,t) \quad \text{in } D_{\alpha-\delta}$$

with δ sufficiently small. Then, from Proposition 2.3 we deduce that

$$\widetilde{u_\epsilon} \to u \quad \text{strongly in } C^0([0,T]; W^{1,1}(D_{\alpha-\delta})).$$

This convergence holds even in $C^0([0,T]; H^1(D_{\alpha-\delta}))$. This is a simple consequence of the fact that $w_\epsilon \equiv 1$ in $D_{\alpha-\delta}$ and follows easily from the proof of Theorem 4.1 in [2].

Similarly, also using hypothesis (2.3)(iii) we have

$$u_\epsilon^0|_{D_{\alpha-\delta}} \to u^0 \quad \text{strongly in } H^1(D_{\alpha-\delta}).$$

2.2. Convergence of the normal derivative on $\partial\Omega$. In [3] it is shown that if the convergence in hypothesis (1.4)(iii) is strong in $H^1(\Omega)$, and if

$$\{\widetilde{u_\epsilon^0}, \widetilde{u_\epsilon^1}, \tilde{f}_\epsilon\} \rightharpoonup \{u^0, u^1, f\} \quad \text{weakly in } H_0^1(\Omega) \times L^2(\Omega) \times L^2(\Omega \times (0,T)), \tag{2.7}$$

then

$$\left.\frac{\partial u_\epsilon}{\partial \nu}\right|_{\partial\Omega\times(0,T)} \rightharpoonup \left.\frac{\partial u}{\partial \nu}\right|_{\partial\Omega\times(0,T)} \quad \text{weakly in } L^2(\partial\Omega \times (0,T)). \tag{2.8}$$

Note that the strong convergence in (1.4)(iii) implies that $\mu \equiv 0$. Now, as pointed out in Remark 1.1, if Ω_ϵ satisfies (1.4) and (1.5), supp $\mu \subset \Omega \backslash D_\alpha$. This allows us to state the following result.

PROPOSITION 2.6. *Let u_ϵ be a solution of* (2.1) *and suppose that $\{u_\epsilon^0, u_\epsilon^1, f_\epsilon\}$ satisifes* (2.7). *Then if Ω_ϵ satisfies hypotheses* (1.4) *and* (1.5), *we have* (2.8). *Assume further that the data satisfy* (2.3). *Then the convergence* (2.8) *is strong.*

Sketch of proof. We follow along the lines of the proof of Lemma 4.6 in [3]. Essentially, the result is a consequence of the following inequality (see Lemma 4.5 in [3]):

$$\begin{cases} \left\| \frac{\partial u_\epsilon}{\partial \nu} \right\|_{L^2(\partial\Omega\times(0,T))} \leq C\{\|u_\epsilon\|_{H^1(D_{\alpha-\delta}\times(0,T))} + \|u_\epsilon^0\|_{H^1(D_{\alpha-\delta})} + \|u_\epsilon^1\|_{L^2(D_{\alpha-\delta})} \\ \qquad\qquad + \|u_\epsilon(T)\|_{H^1(D_{\alpha-\delta})} + \|u_\epsilon'(T)\|_{L^2(D_{\alpha-\delta})} \\ \qquad\qquad + \|f_\epsilon\|_{L^1(0,T;L^2(D_{\alpha-\delta}))}\}. \end{cases} \tag{2.9}$$

This inequality can easily be obtained by multiplying by $h \cdot \nabla u_\epsilon$ the equation defining u_ϵ, where $h \in (W^{1,\infty}(\Omega))^n$, $h = \nu$ on $\partial\Omega$, and supp $h \subset D_{\alpha-\delta}$. This implies, due to the hypotheses, that for a subsequence

$$\frac{\partial u_\epsilon}{\partial \nu} \rightharpoonup \xi \quad \text{weakly in } L^2(\partial\Omega \times (0,T)).$$

The same argument as in [3] shows that, in fact, $\xi = (\partial u/\partial \nu)$ and therefore the whole sequence converges.

To prove the last statement of the proposition, we again apply Lemma 4.5 from [3] to obtain the equivalent of (2.9) written for $u_\epsilon - u$:

$$\begin{cases} \left\| \frac{\partial(u_\epsilon - u)}{\partial \nu} \right\|_{L^2(\partial\Omega\times(0,T))} \le C\{\|u_\epsilon - u\|_{H^1(D_{\alpha-\delta}\times(0,T))} + \|u_\epsilon^0 - u^0\|_{H^1(D_{\alpha-\delta})} \\ \qquad + \|u_\epsilon^1 - u^1\|_{L^2(D_{\alpha-\delta})} \\ \qquad + \|u_\epsilon(T) - u(T)\|_{H^1(D_{\alpha-\delta})} \\ \qquad + \|u_\epsilon'(T) - u'(T)\|_{L^2(D_{\alpha-\delta})} \\ \qquad + \|f_\epsilon - f\|_{L^1(0,T;L^2(D_{\alpha-\delta}))}\}. \end{cases} \tag{2.10}$$

Let $\epsilon \to 0$. Then, as a consequence of hypothesis (2.3), Proposition 2.3, and Remark 2.5, each term in the right-hand side of the inequality converges to zero.

3. Proof of the main results.

3.1. Definition of F_ϵ. Before proving Theorems 1.3 and 1.4, let us briefly describe the construction of the space F_ϵ introduced in HUM.

Consider the wave equation

$$\begin{cases} \varphi_\epsilon'' - \Delta\varphi_\epsilon = 0 \quad \text{in } \Omega_\epsilon \times (0,T) \\ \varphi_\epsilon = 0 \quad \text{on } \partial\Omega_\epsilon \times (0,T) \\ \varphi_\epsilon(0) = \varphi_\epsilon^0 \quad \text{in } \Omega_\epsilon \\ \varphi_\epsilon'(0) = \varphi_\epsilon^1 \quad \text{in } \Omega_\epsilon \end{cases} \tag{3.1}$$

and define

$$\|\{\varphi_\epsilon^0, \varphi_\epsilon^1\}\|_{F_\epsilon} = \left(\int_0^T \int_{\Gamma_0} \left(\frac{\partial \varphi_\epsilon}{\partial \nu} \right)^2 d\sigma\, dt \right)^{1/2}, \tag{3.2}$$

where Γ_0 is a nonempty open subset of $\partial\Omega_\epsilon$ ($d\sigma$ denotes the surface measure on $\partial\Omega_\epsilon$). If T is large enough, by Holmgren's uniqueness theorem it follows that $\|\{\cdot,\cdot\}\|_{F_\epsilon}$ is a norm in $\mathcal{D}(\Omega_\epsilon) \times \mathcal{D}(\Omega_\epsilon)$. Then F_ϵ is defined as the Hilbert space obtained by completion of $\mathcal{D}(\Omega_\epsilon) \times \mathcal{D}(\Omega_\epsilon)$ for this norm. It is proved that there is exact controllability with controls in $L^2(\partial\Omega \times (0,T))$ if and only if the initial data $\{y_\epsilon^1, y_\epsilon^0\} \in F_\epsilon'$ (for details we refer the reader to Lions [7], [8]).

As mentioned above, when Ω_ϵ is smooth, $F_\epsilon = H_0^1(\Omega_\epsilon) \times L^2(\Omega_\epsilon)$ if and only if any generalized ray of geometric optics meets Γ_0 in a time less or equal to T (cf. [1]). On the other hand, if Ω_ϵ is C^2, by multiplier techniques, we can show that $F_\epsilon = H_0^1(\Omega_\epsilon) \times L^2(\Omega_\epsilon)$ if $T > 2\|x - x_0\|_{L^\infty(\Omega_\epsilon)}$ and $\Gamma_0^\epsilon = \{x \in \partial\Omega_\epsilon : (x - x_0) \cdot \nu_\epsilon(x) > 0\}$ for some $x_0 \in \mathbb{R}^n$, $\nu_\epsilon(x)$ being the outward unit vector to Ω_ϵ at point $x \in \partial\Omega_\epsilon$.

If we take $\Gamma_0^\epsilon = \Gamma_0 \subset \partial\Omega$, this condition is not satisified. Moreover, as $\epsilon \to 0$, there is an increasing number of generalized rays that never reach $\partial\Omega$ (see Fig. 1, where $\Gamma_1 = \partial\Omega \backslash \Gamma_0$), the space F_ϵ becomes larger and larger, and consequently F_ϵ' is smaller and smaller. It is quite impossible to characterize these spaces. Note, for example, that

F_ϵ may contain elements which are not distributions on Ω_ϵ. We know that F'_ϵ is strictly contained in $H^{-1}(\Omega_\epsilon)\times L^2(\Omega_\epsilon)$. What Theorem 1.3 shows is that, actually, each element of $H^{-1}(\Omega)\times L^2(\Omega)$ is a (strong) limit of appropriate elements of F'_ϵ, namely, $\{y'_\epsilon(0), -y_\epsilon(0)\}$, where y_ϵ is solution of (1.8).

3.2. Proof of Theorem 1.3. The function y_ϵ is a solution of (1.8) in the transposition sense, i.e., y_ϵ satisfies

$$\langle\{y'_\epsilon(0), -y_\epsilon(0)\}, \{\theta^0_\epsilon, \theta^1_\epsilon\}\rangle - \int_0^T \int_{\Omega_\epsilon} y_\epsilon f_\epsilon \, dx\, dt = \int_0^T \int_{\partial\Omega} v \frac{\partial\theta_\epsilon}{\partial\nu} d\sigma\, dt \tag{3.3}$$

for every $f_\epsilon \in L^1(0,T;L^2(\Omega_\epsilon)), \theta^0_\epsilon \in H^1_0(\Omega_\epsilon), \theta^1_\epsilon \in L^2(\Omega_\epsilon)$, where θ_ϵ is the unique solution of

$$\begin{cases} \theta''_\epsilon - \Delta\theta_\epsilon = f_\epsilon & \text{in } \Omega_\epsilon \times (0,T) \\ \theta_\epsilon = 0 & \text{on } \partial\Omega_\epsilon \times (0,T) \\ \theta_\epsilon(0) = \theta^0_\epsilon \\ \theta'_\epsilon(0) = \theta^1_\epsilon & \text{in } \Omega_\epsilon. \end{cases} \tag{3.4}$$

From [7] it is known that there exists a unique $y_\epsilon \in C([0,T];L^2(\Omega_\epsilon))\cap C^1([0,T];H^{-1}(\Omega_\epsilon))$ satisfying (3.3).

Let us first show that $\{y'_\epsilon(0), -y_\epsilon(0)\} \in F'_\epsilon$.

By the definition of F'_ϵ, we know that $\{y'_\epsilon(0), y_\epsilon(0)\} \in F'_\epsilon$ if and only if

$$|\langle\{y'_\epsilon(0), -y_\epsilon(0)\}, \{\varphi^0_\epsilon, \varphi^1_\epsilon\}\rangle| \le C\left(\int_0^T \int_{\Gamma_0} \left|\frac{\partial\varphi_\epsilon}{\partial\nu}\right|^2 d\sigma\, dt\right)^{1/2} \tag{3.5}$$

for all $\{\varphi^0_\epsilon, \varphi^1_\epsilon\} \in H^1_0(\Omega_\epsilon)\times L^2(\Omega_\epsilon)$ with φ_ϵ solution of (3.1).

By transposition we have the identity

$$\langle\{y'_\epsilon(0), -y_\epsilon(0)\}, \{\varphi^0_\epsilon, \varphi^1_\epsilon\}\rangle = \int_0^T \int_{\Gamma_0} v \frac{\partial\varphi_\epsilon}{\partial\nu} d\sigma\, dt,$$

from which (3.5) follows obviously since $v \in L^2(\Gamma_0 \times (0,T))$.

In order to prove convergence (1.9)(i) let us choose in (3.3) $\theta^0_\epsilon = \theta^1_\epsilon = 0$ and f_ϵ such that

$$\begin{cases} f_\epsilon \in L^2(\Omega\times(0,T)) \\ f_\epsilon \rightharpoonup f \quad \text{weakly in } L^2(\Omega\times(0,T)). \end{cases} \tag{3.6}$$

We claim that

$$\int_0^T \int_\Omega \widetilde{y}_\epsilon f_\epsilon \, dx\, dt \to \int_0^T \int_\Omega y\, f \, dx\, dt, \tag{3.7}$$

which is precisely convergence (1.9)(i).

To begin with the proof of this claim, we remark that Proposition 2.1 can be applied to θ_ϵ, a solution of system (3.4), all its assumptions being fulfilled. It follows that

$$\tilde{\theta}_\epsilon \rightharpoonup \theta \quad \text{weakly * in } L^\infty(0,T;H^1_0(\Omega)),$$

where θ satisfies the limit system

$$\begin{cases} \theta'' - \Delta\theta + \mu\theta = f & \text{in } \Omega\times(0,T) \\ \theta = 0 & \text{on } \partial\Omega\times(0,T) \\ \theta(0) = \theta'(0) = 0 & \text{in } \Omega. \end{cases}$$

Moreover, Proposition 2.6 asserts that

$$\left.\frac{\partial\theta_\epsilon}{\partial\nu}\right|_{\partial\Omega\times(0,T)} \rightharpoonup \left.\frac{\partial\theta}{\partial\nu}\right|_{\partial\Omega\times(0,T)} \quad \text{weakly in } L^2(\partial\Omega\times(0,T)),$$

the domain Ω_ϵ satisfying hypothesis (1.5). This convergence shows that

$$\lim_{\epsilon\to 0}\int_0^T\int_{\Omega_\epsilon} y_\epsilon f_\epsilon dx\,dt = \lim_{\epsilon\to 0}\int_0^T\int_\Omega \widetilde{y}_\epsilon f_\epsilon\,dx\,dt = -\int_0^T\int_{\partial\Omega} v\frac{\partial\theta}{\partial\nu}d\sigma\,dt.$$

Let us now consider the solution y of system (1.6) introduced in Theorem 1.3. Since v is the exact control, we have $y(T) = y'(T) = 0$. Thus y, which is also defined by transposition, satisfies

$$\langle\{y'(0), -y(0)\},\{\eta^0,\eta^1\}\rangle - \int_0^T\int_\Omega y\,g\,dx\,dt = \int_0^T\int_{\partial\Omega} v\frac{\partial\eta}{\partial\nu}d\sigma\,dt \tag{3.8}$$

for all $g \in L^1(0,T;L^2(\Omega)), \eta^0 \in H_0^1(\Omega), \eta^1 \in L^2(\Omega)$ with η the unique solution of

$$\begin{cases} \eta'' - \Delta\eta + \mu\eta = g & \text{in } \Omega\times(0,T)\\ \eta = 0 & \text{on } \partial\Omega\times(0,T)\\ \eta(0) = \eta^0 & \\ \eta'(0) = \eta^1 & \text{in } \Omega. \end{cases}$$

In particular, (3.8) holds for $g = f$ and $\eta^0 = \eta^1 = 0$, in which case, by uniqueness, $\eta = \theta$, and thus (3.7) holds.

It remains to prove the convergences (1.9)(ii) and (1.9)(iii). To do that, we will make use several times of the formulation by transposition of the system defining y_ϵ, with particular choices of the test function. First let ρ_ϵ satisfy the following system:

$$\begin{cases} \rho_\epsilon'' - \Delta\rho_\epsilon = 0 & \text{in } \Omega_\epsilon\times(0,T)\\ \rho_\epsilon = 0 & \text{on } \partial\Omega_\epsilon\times(0,T)\\ \rho_\epsilon(0) = 0 & \text{in } \Omega_\epsilon\\ \rho_\epsilon'(0) = \rho_\epsilon^1 & \text{in } \Omega_\epsilon. \end{cases}$$

With this ρ_ϵ as test function in the definition by transposition of y_ϵ, we have

$$0 = \int_{\Omega_\epsilon} y_\epsilon(0)\rho_\epsilon^1 dx + \int_0^T\int_{\partial\Omega} v\frac{\partial\rho_\epsilon}{\partial\nu}d\sigma\,dt.$$

Suppose that $\rho_\epsilon^1 = \overline{\rho_\epsilon^1}|_{\Omega_\epsilon}$ with

$$\overline{\rho_\epsilon^1} \rightharpoonup \rho^1 \quad \text{weakly in} L^2(\Omega).$$

Propositions 2.3 and 2.6 show that

$$\begin{cases} \left.\frac{\partial\rho_\epsilon}{\partial\nu}\right|_{\partial\Omega\times(0,T)} \rightharpoonup \left.\frac{\partial\rho}{\partial\nu}\right|_{\partial\Omega\times(0,T)} & \text{weakly in } L^2(\partial\Omega\times(0,T))\\ \widetilde{\rho}_\epsilon \rightharpoonup \rho \quad \text{weakly * in } L^\infty(0,T;H_0^1(\Omega)) \end{cases}$$

with the ρ solution of

$$\begin{cases} \rho'' - \Delta\rho + \mu\rho = 0 & \text{in } \Omega\times(0,T)\\ \rho = 0 & \text{on } \partial\Omega\times(0,T)\\ \rho(0) = 0 & \text{in } \Omega\\ \rho'(0) = \rho^1 & \text{in } \Omega \end{cases} \tag{3.9}$$

and, arguing as before, we have

$$\lim_{\epsilon\to 0}\int_{\Omega_\epsilon} y_\epsilon(0)\rho_\epsilon^1\,dx = \lim_{\epsilon\to 0}\int_\Omega \widetilde{y_\epsilon(0)}\overline{\rho_\epsilon^1}dx = -\int_0^T\int_{\partial\Omega} v\frac{\partial\rho}{\partial\nu}\,d\sigma\,dt. \tag{3.10}$$

From the definition, by transposition of y written with the ρ solution of (3.9) as a test function, we have immediately that

$$0 = \int_\Omega y^0\rho^1\,dx + \int_{\partial\Omega} v\frac{\partial\rho}{\partial\nu}d\sigma\,dt,$$

which, combined with (3.10), gives

$$\lim_{\epsilon\to 0}\int_\Omega \widetilde{y_\epsilon(0)}\overline{\rho_\epsilon^1}dx = \int_\Omega y^0\rho^1 dx;$$

thus convergence (1.9)(ii) is established.

To prove (1.9)(iii), let $\psi_\epsilon^0 \in W_0^{1,q'}(\Omega)$ be a sequence such that

$$\psi_\epsilon^0 \rightharpoonup \psi^0 \quad \text{weakly in} W_0^{1,q'}(\Omega), \tag{3.11}$$

where $1/q' + 1/q = 1$, with $q \in (1, n/(n-1))$.

Now let ξ_ϵ be a solution of

$$\begin{cases} \xi_\epsilon'' - \Delta\xi_\epsilon = 0 \quad \text{in } \Omega_\epsilon \times (0,T) \\ \xi_\epsilon = 0 \quad \text{on } \partial\Omega_\epsilon \times (0,T) \\ \xi_\epsilon(0) = \psi_\epsilon^0 w_\epsilon \quad \text{in } \Omega_\epsilon \\ \xi_\epsilon'(0) = 0 \quad \text{in } \Omega, \end{cases}$$

where w_ϵ are the functions introduced in hypothesis (1.4).

With this ξ_ϵ as a test function, y_ϵ satisfies

$$0 = \langle y_\epsilon'(0), \psi_\epsilon^0 w_\epsilon\rangle_{H^{-1}(\Omega_\epsilon),H_0^1(\Omega_\epsilon)} - \int_0^T\int_{\partial\Omega} v\frac{\partial\xi_\epsilon}{\partial\nu}d\sigma\,dt. \tag{3.12}$$

By the definition (1.7) of the extension operator P_ϵ given in §1, we have

$$\begin{aligned} \langle P_\epsilon y_\epsilon'(0), \psi_\epsilon^0\rangle_{W^{-1,q}(\Omega),W_0^{1,q'}(\Omega)} &= \langle w_\epsilon\widetilde{y_\epsilon'(0)}, \psi_\epsilon^0\rangle_{W^{-1,q}(\Omega),W_0^{1,q'}(\Omega)} \\ &= \langle y_\epsilon'(0), w_\epsilon\psi_\epsilon^0|_{\Omega_\epsilon}\rangle_{H^{-1}(\Omega_\epsilon),H_0^1(\Omega_\epsilon)}. \end{aligned} \tag{3.13}$$

The second duality makes sense because the definitions of w_ϵ and ψ_ϵ^0 imply that $w_\epsilon\psi_\epsilon^0 \in H_0^1(\Omega_\epsilon)$. Moreover, by hypothesis (1.4), $w_\epsilon\psi_\epsilon^0 \rightharpoonup \psi^0$ weakly in $H_0^1(\Omega)$.

Once again, by Propositions 2.1 and 2.6, we have, from (3.12) and (3.13), that

$$\lim_{\epsilon\to 0}\langle P_\epsilon y_\epsilon'(0), \psi_\epsilon^0\rangle_{W^{-1,q}(\Omega),W_0^{1,q'}(\Omega)} = \int_0^T\int_{\partial\Omega} v\frac{\partial\xi}{\partial\nu}d\sigma\,dt, \tag{3.14}$$

where ξ is a solution of

$$\begin{cases} \xi'' - \Delta\xi + \mu\xi = 0 \quad \text{in } \Omega \times (0,T) \\ \xi = 0 \quad \text{on } \partial\Omega \times (0,T) \\ \xi(0) = \psi^0 \quad \text{in } \Omega \\ \xi'(0) = 0 \quad \text{in } \Omega. \end{cases}$$

Now, as before, take ξ as a test function for the definition by transposition of y. It comes about that

$$\langle y^1, \psi^0\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} = \int_0^T\int_{\partial\Omega} v\frac{\partial\xi}{\partial\nu}d\sigma\,dt \tag{3.15}$$

which, with (3.14), gives the claimed result and completes the proof of Theorem 1.3. □

3.3. Proof of Theorem 1.4. The tool for proving Theorem 1.4 is exactly the same as in the previous proof, i.e., the choice of particular test functions in the formulation by transposition of the solution y_ϵ. We give here only the main steps. We have, by definition,

$$\left\{\begin{array}{l} \displaystyle\int_0^T\int_{\Omega_\epsilon} y_\epsilon f_\epsilon\,dx\,dt = -\int_{\Omega_\epsilon} y_\epsilon^0\theta_\epsilon'(0)dx \\ \qquad\qquad + \langle y_\epsilon^1, \theta_\epsilon(0)\rangle_{H^{-1}(\Omega_\epsilon),H_0^1(\Omega_\epsilon)} - \displaystyle\int_0^T\int_{\partial\Omega} v\frac{\partial\theta_\epsilon}{\partial\nu}d\sigma\,dt, \\ \qquad\qquad \forall f_\epsilon \in L^1(0,T;L^2(\Omega_\epsilon)) \end{array}\right. \tag{3.16}$$

with θ_ϵ solution of

$$\left\{\begin{array}{l} \theta_\epsilon'' - \Delta\theta_\epsilon = f_\epsilon \quad \text{in } \Omega_\epsilon\times(0,T) \\ \theta_\epsilon = 0 \quad \text{on } \partial\Omega_\epsilon\times(0,T) \\ \theta_\epsilon(T) = \theta_\epsilon'(T) = 0 \quad \text{in } \Omega_\epsilon. \end{array}\right. \tag{3.17}$$

Take in (3.17) f_ϵ satisfying (3.6). We will pass to the limit in identity (3.16). To do that, apply Proposition 2.1 (see also Remark 2.2) to θ_ϵ. We have

$$\widetilde{\theta_\epsilon} \rightharpoonup \theta \quad \text{weakly * in } L^\infty(0,T;H_0^1(\Omega))$$

with θ the unique solution of

$$\left\{\begin{array}{l} \theta'' - \Delta\theta + \mu\theta = f \quad \text{in } \Omega\times(0,T) \\ \theta = 0 \quad \text{on } \partial\Omega\times(0,T) \\ \theta(T) = \theta'(T) = 0 \quad \text{in } \Omega \end{array}\right.$$

and, by Proposition 2.6,

$$\left.\frac{\partial\theta_\epsilon}{\partial\nu}\right|_{\partial\Omega\times(0,T)} \rightharpoonup \left.\frac{\partial\theta}{\partial\nu}\right|_{\partial\Omega\times(0,T)} \quad \text{weakly in } L^2(\partial\Omega\times(0,T)).$$

Moreover, from Proposition 2.1 we also have the following pointwise convergences:

$$\begin{array}{l} \widetilde{\theta_\epsilon}(t) \rightharpoonup \theta(t) \quad \text{weakly in } H_0^1(\Omega) \\ \widetilde{\theta_\epsilon'}(t) \rightharpoonup \theta'(t) \quad \text{weakly in } L^2(\Omega) \end{array}$$

for all $t\in[0,T]$. These convergences enable us to pass to the limit in the right-hand side term of identity (3.16). Recalling the definitions of y_ϵ^0 and y_ϵ^1, we have, successively,

$$\begin{array}{l} \langle y_\epsilon^0, \theta_\epsilon'(0)\rangle_{L^2(\Omega_\epsilon),L^2(\Omega_\epsilon)} = \displaystyle\int_\Omega y^0\widetilde{\theta_\epsilon'(0)}dx \to \int_\Omega y^0\theta'(0)dx \\ \langle y_\epsilon^1, \theta_\epsilon(0)\rangle_{H^{-1}(\Omega_\epsilon),H_0^1(\Omega_\epsilon)} = \langle y^1, \widetilde{\theta_\epsilon(0)}\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} \to \langle y^1,\theta(0)\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}, \end{array}$$

hence

$$\left\{\begin{array}{l} \displaystyle\lim_{\epsilon\to0}\int_0^T\int_{\Omega_\epsilon}\widetilde{y_\epsilon}f_\epsilon dx\,dt = -\int_\Omega y^0\theta'(0)dx \\ \qquad\qquad +\langle y^1,\theta(0)\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} - \displaystyle\int v\frac{\partial\theta}{\partial\nu}d\sigma\,dt. \end{array}\right. \tag{3.18}$$

Consider the formulation by transposition of system (1.6) with θ as a test function:

$$\int_0^T \int_\Omega yf\,dx\,dt = -\int_\Omega y^0\theta'(0)dx + \langle y^1, \theta(0)\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} - \int_0^T \int_{\partial\Omega} v\frac{\partial\theta}{\partial\nu}d\sigma\,dt.$$

This identity combined with (3.18) gives the strong convergence in $L^2(\Omega)$ of $\widetilde{y_\epsilon}$ and thus (1.10) is proved.

To prove convergence (1.11), first let ρ_ϵ be the solution of

$$\text{(3.19)}\qquad \begin{cases} \rho_\epsilon'' - \Delta\rho_\epsilon = 0 \quad \text{in } \Omega_\epsilon \times (0,T) \\ \rho_\epsilon = 0 \quad \text{on } \partial\Omega_\epsilon \times (0,T) \\ \rho_\epsilon(T) = 0 \quad \text{in } \Omega_\epsilon \\ \rho_\epsilon'(T) = \rho_\epsilon^1|_{\Omega_\epsilon} \quad \text{in } \Omega_\epsilon \end{cases}$$

where

$$\rho_\epsilon^1 \rightharpoonup \rho^1 \quad \text{weakly in } L^2(\Omega).$$

By transposition, y_ϵ satisfies, in particular,

$$0 = \int_{\Omega_\epsilon} y_\epsilon(T)\rho_\epsilon^1\,dx - \int_{\Omega_\epsilon} y_\epsilon^0\rho_\epsilon'(0)\,dx + \langle y_\epsilon^1, \rho_\epsilon(0)\rangle_{H^{-1}(\Omega_\epsilon),H_0^1(\Omega_\epsilon)} - \int_0^T \int_{\partial\Omega} v\frac{\partial\rho_\epsilon}{\partial\nu}d\sigma\,dt.$$

Passing to the limit as before, we have that

$$\text{(3.20)}\qquad \begin{cases} \lim_{\epsilon\to 0}\int_{\Omega_\epsilon} y_\epsilon(T)\rho_\epsilon^1\,dx = \lim_{\epsilon\to 0}\int_\Omega \widetilde{y_\epsilon(T)}\rho_\epsilon^1\,dx \\ \qquad = \int_\Omega y^0\rho'(0)dx - \langle y^1, \rho(0)\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} + \int_0^T \int_{\partial\Omega} v\frac{\partial\rho}{\partial\nu}d\sigma\,dt, \end{cases}$$

where ρ is the solution of the limit system for (3.19) given by Proposition 2.1, i.e., ρ satisfies

$$\begin{cases} \rho'' - \Delta\rho + \mu\rho = 0 \quad \text{in } \Omega \times (0,T) \\ \rho = 0 \quad \text{on } \partial\Omega \times (0,T) \\ \rho(T) = 0 \quad \text{in } \Omega \\ \rho'(T) = \rho^1 \quad \text{in } \Omega. \end{cases}$$

On the other hand, with this ρ as a test function in the definition of y we obtain

$$-\int_\Omega y^0\rho'(0)dx + \langle y^1, \rho(0)\rangle_{H^{-1}(\Omega),H_0^1(\Omega)} - \int v\frac{\partial\rho}{\partial\nu}d\sigma\,dt = 0,$$

which, combined with (3.20), leads to

$$\lim_{\epsilon\to 0}\int_\Omega \widetilde{y_\epsilon(T)}\rho_\epsilon^1\,dx = 0,$$

hence (1.11)(i) is proved.

Finally, to prove (1.11)(ii), take, as in the last step of the proof of Theorem 1.3, $\psi_\epsilon^0 \in W_0^{1,q'}(\Omega)$ satisfying (3.11) and choose as a test function in (3.16) the solution ξ_ϵ of the system

$$\begin{cases} \xi_\epsilon'' - \Delta\xi_\epsilon = 0 \quad \text{in } \Omega_\epsilon \times (0,T) \\ \xi_\epsilon = 0 \quad \text{on } \partial\Omega_\epsilon \times (0,T) \\ \xi_\epsilon(T) = \psi_\epsilon^0 w_\epsilon \quad \text{in } \Omega_\epsilon \\ \xi_\epsilon'(T) = 0 \quad \text{in } \Omega_\epsilon. \end{cases}$$

Arguing for the proof of (3.15), we obtain

$$\lim_{\epsilon\to 0} \langle P_\epsilon y_\epsilon'(T), \psi_\epsilon^0 \rangle_{W^{-1,q}(\Omega), W_0^{1,q'}(\Omega)} = 0,$$

which ends the proof of Theorem 1.5. □

4. Examples and Related Problems. Here we give some examples where hypothesis (1.4) is satisfied. For instance, that is the case when S_ϵ is the union of periodically distributed holes (the period is 2ϵ) of critical size

$$S_\epsilon = \cup_{i=1}^{N_\epsilon} T_i^\epsilon,$$

where T_i^ϵ are balls of radius $\gamma^\epsilon = a^\epsilon$ with

$$a_\epsilon = \begin{cases} \delta_\epsilon \exp\left(-\frac{c_0}{\epsilon^2}\right) & \text{if } n = 2 \\ c_0 \epsilon^{n/n-2} & \text{if } n > 2, \end{cases}$$

where c_0 is a positive constant and δ_ϵ is such that $\epsilon^2 \log \delta_\epsilon \to 0$. Hypothesis (1.4) holds with

$$\begin{cases} \mu = \frac{\pi}{2}\frac{1}{c_0} & \text{if } n = 2 \\ \mu = \frac{s_n(n-2)}{2n} c_0^{n-2} & \text{if } n \geq 3, \end{cases}$$

where s_n is the surface of the unit sphere in $\mathbb{R}^n$. It is this size a_ϵ which, in the elliptic case, transforms $-\Delta$ in $-\Delta + \mu$, as it was proved in [6] where more examples can be found as well as explicit computations of w_ϵ. Of course, in this example when ϵ goes to zero the number of holes tends to ∞.

When $a_\epsilon \ll \gamma_\epsilon$, hypothesis (1.4) holds with $\mu = 0$. For this situation, with Ω satisfying hypothesis (1.5), the equivalent of Theorems 1.3 and 1.4 were partially proved in [4].

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Contrôle et stabilisation dans les problèmes hyperboliques,* Appendix II in [7], pp. 497–537.

[2] D. Cioranescu, P. Donato, F. Murat, and E. Zuazua, *Homogenization and correctors for the wave equation in domains with small holes*, Ann. Scuola Norm. Sup. Pisa, 18 (1991), pp. 251–293.

[3] D. Cioranescu, P. Donato, and E. Zuazua, *Exact boundary controllability for the wave equation in domains with small holes,* J. Math. Pures Appl., 71 (1992), pp. 343–377.

[4] ———, *Strong convergence on the exact controllability of the wave equation in perforated domains,* in Proceedings of the International Conference on Control and Estimation of Distributed Parameter Systems, Vorau (Styria), 1990, W. Desch, F. Kappel, and K. Kunisch, eds., ISNM Series, Vol. 100, Birkhaüser Verlag, Basel, 1991, pp. 99–113.

[5] ———, *Approximate controllability for the wave equation with oscillating coefficients,* in Boundary Control and Boundary Variations, J. P. Zolesio, ed., Proceedings of the IFIP Conference, Sophia-Antipolis, France, 1990, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, to appear.

[6] D. Cioranescu and F. Murat, *Un terme étrange venu d'ailleurs,* in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vols. II and III, Research Notes in Mathematics 60 and 70, Pitman, London, 1982, pp. 93–138, 154–178.

[7] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, Tome 1, contrôlabilité exacte,* Rech. Math. Appl. 8, Masson, Paris, 1988.

[8] ———, *Remarks on approximate controllability,* J. Analyse Math., 59 (1992), pp. 103–116.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 51–58, January 1994

004

DENSITY RESULTS FOR PROPER EFFICIENCIES*

DEMING ZHUANG†

Abstract. Several density results are established for different notions of proper efficiency in vector optimization without the requirement of ordering cones being boundedly based. For a compact set, the set of Henig proper efficient points is shown to be dense in the set of all efficient points. Density theorems for Borwein's proper efficiency and for positive scalarizable efficiency follow immediately with appropriate assumptions.

Key words. vector optimization, proper efficiency, density results, bases of ordering convex cones

AMS subject classifications. 49A27, 46A40, 90C31

1. Introduction and preliminaries. Vector optimization problems originated from decision-making problems appearing in economics, management science, and social science, where it is often required that decision making be based on optimizing several criteria. A vector optimization problem therefore involves finding all efficient points in some vector partial order. However, some efficient points (see Example 2.1) exhibit certain abnormal properties: They may cause arbitrarily large marginal trade off or may not be expressed as a solution of an appropriate linear scalar optimization problem. Therefore, various concepts of proper efficiency have been introduced to eliminate such anomalous efficient points. Any reasonable concept of proper efficiency should retain most of the efficient points and exclude only anomalous ones. Thus, when a new proper efficiency is introduced, it is important to study density results—sufficient conditions to guarantee that the closure of the set of proper efficient points contains all efficient points. Density results for various proper efficiencies have been considered by many authors (e.g., see [1]–[4], [11]–[15], [17], [21], [24]). Most previous results required compactness, or at least boundedness, of the base of the ordering cones. While those results have proved to be important in applications, one notable limitation is that many natural ordering cones in infinite-dimensional normed linear spaces do not have bases that are bounded, let alone compact. Recently, Dauer and Gallagher [12] established some interesting density results without the requirement of ordering cones being boundedly based. In this note, we show density results for several kinds of proper efficiencies without such a requirement. The machinery developed here also enables us to derive much stronger results when the boundedness of the base of the ordering cone is present. This is demonstrated in [9]; see also [25].

For the sake of simplicity, we make the following assumptions (unless specifically stated otherwise). Throughout the note, X will always be a partially ordered real normed linear space and a subset C of X is always assumed to be nonempty. The partial ordering cone S of X is always assumed to be closed, convex ($S+S \subset S$), and pointed ($S \cap -S\{0\}$). We associate a *dual cone* with S, denoted by S^+, in X^* (the norm dual of X)

$$S^+ := \{\phi \in X^* | \phi(s) \geq 0 \quad \forall s \in S\}.$$

Then S^+ is a convex cone and is closed in $\sigma(X^*, X)$, the weak-star topology.

Recall that a *base* of a cone S is a convex subset Θ of S such that

$$S = \{\lambda\theta | \lambda \geq 0 \text{ and } \theta \in \Theta\} \quad \text{and} \quad 0 \notin \text{cl}(\Theta).$$

* Received by the editors September 25, 1989; accepted for publication (in revised form) July 28, 1992.

† Department of Mathematics, Mount Saint Vincent University, Halifax, Nova Scotia, Canada, B3M 2J6. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada and Mount Saint Vincent University internal research grants.

Of course, S is pointed if S has a base. We also use the following notation:

(1) cone(A) denotes the cone generated by the set A, i.e.,

$$\text{cone}(A) := \cup\{tA | t \geq 0\};$$

while cl[cone(A)] denotes the closure of cone(A);

(2) S^{+i} denotes the set of all strictly positive linear functionals in S^+, that is,

$$S^{+i} := \{f \in X^* | f(s) > 0 \text{ for all } s \text{ in } S \text{ and } s \neq 0\}.$$

It follows directly from the Hahn–Banach theorem that S^{+i} is nonempty exactly when S has a base.

2. Proper efficiencies. Efficiency (Pareto minimality) is a fundamental concept in vector optimization.

DEFINITION 2.0. A point x_0 in C is said to be an *efficient point* of C with respect to S, written as $x_0 \in E(C, S)$, if

$$(C - \{x_0\}) \cap -S = \{0\}. \tag{2.1}$$

For simplicity, we often write (2.1) as

$$(C - x_0) \cap -S = 0.$$

Note that, as S is assumed to be convex and pointed, we have

$$E(C, S) = E(C + S, S). \tag{2.2}$$

Some efficient points exhibit abnormal behavior.

Example 2.1. Consider

$$\min_S \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 = 1, y \leq 0\},$$

where S is the nonnegative orthant in $\mathbb{R}^2$. Then $(-1, 0)$ is an efficient point with respect to S. However, when x is near -1, y increases as x decreases; moreover,

$$\frac{y - 0}{-1 - x} \to \infty \quad \text{as } x \to -1,$$

i.e., the *marginal trade off* is arbitrarily large. This is an undesirable property [13]. We may also note that the point $(-1, 0)$ is not scalarizable by a strictly positive linear functional.

To eliminate such anomalous types of efficient points, Geoffrion [13] introduced the notion of proper efficiency in $\mathbb{R}^n$ with respect to the usual ordering cone, the nonnegative orthant $\mathbb{R}^n_+$, which extends Kuhn and Tucker's definition of proper efficiency [22].

DEFINITION 2.2. Let C be a subset of $\mathbb{R}^n$. A point $x_0 = (x_0(1), x_0(2), \ldots x_0(n))$ in C is said to be a *Geoffrion proper efficient point* of C, written as $x_0 \in GE(C, \mathbb{R}^n_+)$, if x_0 is efficient and if there is $M > 0$ with the property that whenever $x \in C$ with $x_0(i) > x(i)$ for some $i \in \{1, 2, \ldots, n\}$, we can find some $j \in \{1, 2, \ldots n\} \backslash \{i\}$ such that $x_0(j) < x(j)$ and

$$\frac{x(i) - x_0(i)}{x_0(j) - x(j)} \leq M.$$

According to this definition, the point $(-1, 0)$ in Example 2.1 is not proper efficient.

Geoffrion's notion of proper efficiency was extended to the setting of a locally convex topological linear space with partial ordering induced by a closed convex cone by Borwein in [3].

DEFINITION 2.3. Let X be a normed space and let C be a subset of X. A point x_0 in C is said to be a *Borwein proper efficient* point of C with respect to S, written as $x_0 \in BE(C, S)$, if

$$\text{cl}[\text{cone}(C - x_0)] \cap -S = 0. \tag{2.3}$$

It is clear from these definitions that $BE(C, S) \subset E(C, S)$. The containment can be strict as is demonstrated by an example of the closed Euclidean unit ball in $\mathbb{R}^2$ with respect to $\mathbb{R}^2_+$. In this case $(0, -1)$ and $(-1, 0)$ are efficient but not proper efficient.

Borwein's proper efficiency coincides with Geoffrion's proper efficiency in finite-dimensional settings [3]. Existence results and various scalarization properties of Borwein's proper efficiency can be found in [3], [4], [6], [11], and are surveyed in [8], [10], [19], [20], and elsewhere.

When the ordering cone S has a base Θ, Henig defined another kind of proper efficiency which refines Borwein's proper efficiency in this setting [15].

DEFINITION 2.4. Let X be a normed space and let C be a subset of X. A point x_0 in C is said to be a *Henig proper efficient point* (or Henig point) of C with respect to S (more properly, with respect to Θ), denoted by

$$x_0 \in HE(C, S) \quad \text{or} \quad x_0 \in HE(C, \Theta),$$

if there exists some $\varepsilon > 0$ such that

$$\text{cl}[\text{cone}(C - x_0)] \cap -S_\varepsilon(\Theta) = 0, \tag{2.4}$$

where $S_\varepsilon(\Theta) := \text{cl}[\text{cone}(\Theta + \varepsilon B)]$ and B is the closed unit ball of X. We call $S_\varepsilon(\Theta)$ the *Henig dilating cone*. Note that (2.4) says $x_0 \in BE(C, S_\varepsilon(\Theta))$. In other words, $x_0 \in HE(C, \Theta)$ if and only if, for some $\varepsilon > 0, x_0 \in BE(C, S_\varepsilon(\Theta))$. It is clear that if the ordering cone S has a base Θ, then $HE(C, \Theta) \subset BE(C, S)$, since $S \subset S_\varepsilon(\Theta)$ for all $\varepsilon > 0$. The following example shows that, in general, the containment is strict.

Example 2.5. Let X be $l_1(\mathbb{N})$ and S be the natural ordering cone $l_1+ (\mathbb{N})$ that has a bounded base:

$$\Theta = \{x \in X / \Sigma x(n) = 1 \text{ and } x(n) \geq 0, n = 0, 1, 2, \ldots\}.$$

Let

$$C := \{-e_n + 2^{-(n-1)} e_1 | n = 2, 3, 4, \ldots\} \cup \{0\},$$

where e_i is the ith basis element in $l_1(\mathbb{N})$, that is, $e_i = \{0, \ldots, 1, 0 \ldots)$ with 1 on the ith coordinate.

Then because $\text{cl}[\text{cone}(C)] = \text{cone}(C)$ and $\text{cone}(C) \cap -S = 0$, we have $0 \in BE(C, S)$. However, for $n = 2, 3, \ldots$

$$x_n := -e_n + 2^{-(n-1)} e_1 \in C \cap (2^{-(n-1)} B - \Theta).$$

Hence $0 \notin HE(C, \Theta)$.

3. Density theorem for Henig proper efficiency. We have seen that, in general, the set of Henig proper efficient points is strictly contained in the set of efficient points. In this section, we show that with (weak) compactness of the set C, the set of Henig proper efficient points of C is (weakly dense) norm dense in the set of efficient points.

The following lemma and proposition on properties of Henig dilating cones are keys to the proof of the main theorem. They are partial results in [9]. For the reader's convenience, we present the proof here again.

LEMMA 3.1. *Let X be a normed linear space, $S \subset X$ a closed and convex cone with a closed base Θ. Let $\delta := \inf\{\|\theta\| \,|\theta \in \Theta\} > 0$. Define, for $0 < \varepsilon < \delta$, the Henig dilating cone*

$$S_\varepsilon(\Theta) := \mathrm{cl}[\mathrm{cone}(\Theta + \varepsilon B)].$$

Then $S_\varepsilon(\Theta) = \mathrm{cone}[\mathrm{cl}(\Theta + \varepsilon B)]$ and $S_\varepsilon(\Theta)$ is pointed.

Proof. Let

$$\Theta_\varepsilon := \mathrm{cl}(\Theta + \varepsilon B).$$

Since $S_\varepsilon(\Theta)$ is closed, $\mathrm{cl}[\mathrm{cone}(\Theta_\varepsilon] = S_\varepsilon(\Theta)$. Then since $0 \notin \Theta_\varepsilon$ for $0 < \varepsilon < \delta$,

$$\mathrm{cl}[\mathrm{cone}(\Theta_\varepsilon)] = \mathrm{cone}(\Theta_\varepsilon) \cup R(\Theta_\varepsilon), \tag{3.1}$$

where $R(\Theta_\varepsilon) := \{\lim t_\alpha \theta_\alpha^\varepsilon | t_\alpha \to 0, t_\alpha \geq 0, \theta_\alpha^\varepsilon \in \Theta_\varepsilon\}$.

Indeed, x in $\mathrm{cl}[\mathrm{cone}(\Theta_\varepsilon)]$ implies that x is either in $\mathrm{cone}(\Theta_\varepsilon)$ or

$$x = \lim t_\alpha \theta_\alpha^\varepsilon \quad \text{for } t_\alpha \geq 0 \quad \text{and} \quad \theta_\alpha^\varepsilon \in \Theta_\varepsilon.$$

Without loss of generality (taking subnets if necessary), we may assume that t_α converges. If t_α tends to infinity, then t_α^{-1} tends to zero, which implies that $\theta_\alpha^\varepsilon$ tends to zero. This is impossible as 0 is not in the closure of Θ_ε. Hence, t_α converges to t for some $t < \infty$. If $t = 0$, then x is in $R(\Theta_\varepsilon)$ by the definition. If $t \neq 0$, then as $\Theta_\varepsilon := \mathrm{cl}(\Theta + \varepsilon B)$,

$$\theta_\alpha^\varepsilon \to t^{-1} x \in \Theta_\varepsilon.$$

Hence, $x \in \mathrm{cone}(\Theta_\varepsilon)$. Therefore, (3.1) is verified.

Note that it is easy to check that

$$R(\Theta_\varepsilon) = R(\Theta) \subset S \subset \mathrm{cone}(\Theta_\varepsilon)$$

because Θ is a base for S and S is closed. Whence $\mathrm{cone}(\Theta_\varepsilon) = S_\varepsilon(\Theta)$.

Now, $S_\varepsilon(\Theta) = \mathrm{cone}(\Theta_\varepsilon)$ and has a closed base Θ_ε; hence $S_\varepsilon(\Theta)$ is pointed. □

PROPOSITION 3.2. *Let X be a normed space and suppose the ordering cone S has a closed base Θ. Let*

$$\delta := \inf\{\|\theta\| \,|\theta \in \Theta\}.$$

Then for any set C in X the following are equivalent:

(1) $x_0 \in E(C, S_\varepsilon(\Theta))$ *for some* $0 < \varepsilon < \delta$;

(2) $x_0 \in HE(C, \Theta)$.

Proof. From the definitions,

$$HE(C, \Theta) \subset E(C, S_\varepsilon(\Theta))$$

is clear for any sufficiently small $\varepsilon > 0$. On the other hand, x_0 is in $E(C, S_\varepsilon(\Theta))$ if and only if

$$\operatorname{cone}(C - x_0) \cap -S_\varepsilon(\Theta) = 0.$$

By Lemma 3.1, for $0 < \varepsilon < \delta$, $\operatorname{cl}(\varepsilon B - \Theta)$ is a closed base and

$$-S_\varepsilon(\Theta) = \operatorname{cone}[\operatorname{cl}(\varepsilon B - \Theta)].$$

Thus, for any $0 < \varepsilon' < \varepsilon$, $\operatorname{cone}(C - x_0) \cap (\varepsilon' B - \Theta) = \emptyset$. This implies that

$$\operatorname{cone}(C - x_0) \cap [(2^{-1}\varepsilon' B - \Theta) + 2^{-1}\varepsilon' \operatorname{int}(B)] = \emptyset.$$

Noting that the second set is open, we have

$$\operatorname{cl}[\operatorname{cone}(C - x_0)] \cap \operatorname{cl}(2^{-1}\varepsilon' B - \Theta) = \emptyset.$$

Therefore, by Lemma 3.1 again,

$$\operatorname{cl}[\operatorname{cone}(C - x_0)] \cap \operatorname{cone}[\operatorname{cl}(2^{-1}\varepsilon' B - \Theta)] = \operatorname{cl}[\operatorname{cone}(C - x_0)] \cap -S_{\varepsilon''}(\Theta) = \{0\}$$

for $\varepsilon'' := 2^{-1}\varepsilon'$. This proves that x_0 is in $BE(C, S_{\varepsilon''}(\Theta))$, i.e., $x_0 \in HE(C, \Theta)$. Hence (1) implies (2). □

Now we present our main density theorem. Note that for a fixed closed base Θ, we always have

$$HE(C + S, \Theta) \subset HE(C, \Theta).$$

THEOREM 3.3. *Let X be a normed linear space, let S be a closed ordering cone with a closed base Θ, and let C be a nonempty subset of X.*

(1) *If C is weakly compact, then $HE(C + S, \Theta)$ is weakly dense in $E(C, S)$.*

(2) *If C is norm compact, then $HE(C + S, \Theta)$ is norm dense in $E(C, S)$.*

Proof. (1) Let x_0 be in $E(C, S)$. Without loss of generality, we may assume that $x_0 = 0$. Let

$$\delta := \inf\{\|\theta\| \mid \theta \in \Theta\}.$$

We claim that for any given weak neighbourhood W of 0, we can find $0 < \varepsilon < \delta$ so that

$$C_\varepsilon := C \cap -S_\varepsilon(\Theta) \subset W. \tag{3.2}$$

Indeed, suppose (3.2) were not true; then there would be a continuous linear functional $f \neq 0$ such that for n so large that $1/n < \delta$, we could find c_n in C with the property

$$c_n \in -S_{1/n}(\Theta), \qquad |f(c_n)| \geq 1.$$

Since, by Lemma 3.1,

$$-S_{1/n}(\Theta) = -\operatorname{cone}[\operatorname{cl}(\Theta + n^{-1}B)] \subset -\operatorname{cone}[\Theta + (n-1)^{-1}B],$$

there would be b_n in B, θ_n in Θ, and $t_n > 0$ such that

$$-c_n = t_n[\theta_n + (n-1)^{-1}b_n].$$

Since C is weakly compact we may (on extracting a subnet if necessary) assume that c_n tends to c_0 weakly and t_n tends to $\mu \in [0, \infty]$.

(i) If $\mu = \infty$, then, as $-c_n$ is bounded, $t_n^{-1}(-c_n) = \theta_n + (n-1)^{-1}b_n$ tends to zero. This contradicts the fact $0 \notin \Theta$.

(ii) If $\mu > 0$, we have

$$w - \lim(-c_n) = \mu\theta$$

for some $\theta \in \Theta$. This violates the fact that $0 \in E(C, S)$.

(iii) If $\mu = 0$, then because for all n,

$$-c_n = t_n[\theta_n + (n-1)^{-1}b_n] \geq_S t_n b_n/(n-1),$$

we have $0 \geq_S c_0$, which implies that $0 = c_0$ as $0 \in E(C, S)$. Then for large n,

$$\inf|f(c_n)| \geq 1$$

is clearly impossible. Therefore (3.2) is verified.

Now C_ε is weakly compact, so by Theorem 2.1 in [6], $E(C_\varepsilon, K)$ is not empty for every closed ordering cone K. In particular,

$$E(C_\varepsilon, S_\varepsilon(\Theta)) \neq \emptyset.$$

Let $e \in E(C_\varepsilon, S_\varepsilon(\Theta))$, and we check that $e \in E(C, S_\varepsilon(\Theta))$ and e is in W. Therefore, by Proposition 3.2,

$$E(C, S_\varepsilon(\Theta)) = E(C + S, S_\varepsilon(\Theta)) \subset HE(C + S, \Theta).$$

The first equality holds because $S_\varepsilon(\Theta)$ is convex and pointed. Since both x_0 and W are arbitrarily chosen, we see that $HE(C + S, \Theta)$ is weakly dense in $E(C, S)$.

(2) When C is norm compact then by (1), given x_0 in $E(C, S)$ we can find a net c_α in $HE(C + S, \Theta)$ converging weakly to x_0 and, from the proof, each c_α is in C. Since C is compact in this case, c_α may be assumed to be norm convergent. Therefore $HE(C+S, \Theta)$ is norm dense in $E(C, S)$. □

4. Density results for other proper efficiencies. In this section we derive density results for various other proper efficiencies. These results are established by exploring the relationship between Henig proper efficiency and other proper efficiencies and then applying Theorem 3.3. First, as an easy consequence of Theorem 3.3 we prove a density theorem for Borwein's proper efficiency.

COROLLARY 4.1. *Let X be a normed linear space, let S be a closed ordering cone with a closed base Θ, and let C be a nonempty subset of X.*

(1) *If C is weakly compact, then $BE(C + S, S)$ is weakly dense in $E(C, S)$.*

(2) *If C is norm compact, then $BE(C + S, S)$ is norm dense in $E(C, S)$.*

Proof. When S has a base, $HE(C + S, \Theta) \subset BE(C + S, S)$. □

Let us recall another class of proper efficient points.

DEFINITION 4.2. If S is based and C is closed, x is said to be *positive scalarizable*, denoted by $x \in \text{Pos}(C, S)$ if there is some $f \in S^{+i}$ such that $f(C - x) \geq 0$.

Scalarization is a fundamental principle in vector optimization theory. Thus positive scalarizable points form an important class of proper efficient points. It is easy to see that every positive scalarizable point is Borwein proper efficient. But any Borwein proper efficient point in $l_2(\mathbb{R})$ with ordering cone $l_2^+(\mathbb{R})$ is not positive scalarizable, as there is

no strict positive linear functional on the space. Example 2.5 shows also that positive scalarizable points may not be Henig proper efficient.

Example 4.3. Let X, S, Θ, and C be as in Example 2.5. We have seen that $0 \notin HE(C, \Theta)$. To see 0 is in $\text{Pos}(C, S)$, we observe that

$$\phi = (2^{-1}, 2^{-2}, \ldots, 2^{-n}, \ldots) \in S^{+i} := \{y \in l_\infty(\mathbb{N}) | y(n) > 0, n = 1, 2, \ldots\}$$

and $\phi(x_n) = 0$ for $x_n := -e_n + 2^{-(n-1)}e_1 \in C$.

When the set C is convex and the ordering cone S is based, however, a positive scalarizable point is Henig proper efficient. Moreover, any Henig point is also positive scalarizable, as the following proposition shows.

PROPOSITION 4.4. *Let X be a normed linear space, let S be a closed ordering cone with a closed base Θ, and let*

$$\delta := \inf\{\|\theta\| \,|\theta \in \Theta\}.$$

Suppose that C is convex. Then x in $BE(C, S_\varepsilon(\Theta))$ for some $0 < \varepsilon < \delta$ implies that x is in $\text{Pos}(C, S)$. *On the other hand, x in* $\text{Pos}(C, S)$ *implies that there is a closed base Θ' of S such that x is in $HE(C, \Theta')$.*

Proof. If x is in $BE(C, S_\varepsilon(\Theta))$ for some $0 < \varepsilon < \delta$, then

$$\text{cl}[\text{cone}(C - x_0)] \cap \text{cl}[\text{cone}(\varepsilon B - \Theta)] = 0.$$

By Lemma 3.1, this implies

$$\text{cone}(C - x_0) \cap \text{int}[\text{cone}(\varepsilon B - \Theta)] = \emptyset.$$

As $\text{cone}(C - x_0)$ and $\text{cone}(\varepsilon B - \Theta)$ are convex and the latter has nonempty interior, the Hahn–Banach theorem [16, p. 15] provides a nonzero linear functional ϕ in X^* and b in $\mathbb{R}$ such that

$$\phi[\text{cone}(C - x_0)] \geq b \geq \phi[\text{cone}(\varepsilon B - \Theta)].$$

Note that 0 in $\text{cone}(C - x_0)$ implies that $b \leq 0$. Moreover,

$$0 \geq b \geq \phi[\text{cone}(\varepsilon B - \Theta)]$$

implies that $b = 0$. Hence for all $\|x\| \leq 1$, and each θ in Θ,

$$\phi(\theta) \geq \phi(\varepsilon 2^{-1} x).$$

Thus $\phi(\theta) \geq \varepsilon 2^{-1}\|\phi\| > 0$. Thus, $\phi \in S^{+i}$, and $\phi(C - x) \geq 0$. This proves that x_0 is in $\text{Pos}(C, S)$.

Conversely, suppose that x_0 is in $\text{Pos}(C, S)$. Let ϕ be in S^{+i} such that $\|\phi\| = 1$ and $\phi(C - x_0) \geq 0$. Let $\Theta' = \phi^{-1}(1) \cap S$. Then Θ' is a closed base for S and for $0 < \varepsilon < 1$,

$$\phi(\Theta' + \varepsilon B) > 0.$$

So, $\phi(x) < 0$ for all x in $-S_\varepsilon(\Theta') \backslash \{0\}$. Hence,

$$\text{cl}[\text{cone}(C - x_0)] \cap -S_\varepsilon(\Theta') = 0.$$

Therefore, x_0 is in $BE(C, S_\varepsilon(\Theta'))$. □

COROLLARY 4.5. *Let X be a normed linear space, let S be a closed ordering cone with a closed base Θ, and let C be a nonempty convex subset of X.*

(1) *If C is weakly compact, then $\text{Pos}(C + S, S)$ is weakly dense in $E(C, S)$.*

(2) *If C is norm compact, then $\text{Pos}(C + S, S)$ is norm dense in $E(C, S)$.*

Proof. The corollary follows because in this case,

$$HE(C + S, \Theta) \subset \text{Pos}(C + S, S). \qquad \square$$

Acknowledgments. This research was conducted while the author was working on his Ph.D. under the supervision of Professor J.M. Borwein, whose guidance and valuable suggestions are gratefully appreciated. The author would also like to thank referees for their helpful comments.

REFERENCES

[1] K. J. Arrow, E. W. Barankin, and D. Blackwell, *Admissible points of convex sets,* in Contribution to the Theory of Games, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1953, pp. 87–92.

[2] H. P. Benson, *An improved definition of proper efficiency for vector maximization with respect to cones*, J. Math. Anal. Appl., 71 (1979), pp. 232–241.

[3] J. M. Borwein, *Proper efficient points for maximizations with respect to cones*, SIAM J. Control Optim., 15 (1977), pp. 57–63.

[4] ———, *The geometry of Pareto optimality*, Math. Oper. Statist., 11 (1980), pp. 235–248.

[5] ———, *Continuity and differentiability of convex operators*, Proc. London Math. Soc., 44 (1982), pp. 420–444.

[6] ———, *On the existence of Pareto efficient points*, Math. Oper. Res., 9 (1983), pp. 64–73.

[7] ———, *Norm duality for convex processes and applications*, J. Optim. Theory Appl., 48 (1986), pp. 22–29.

[8] ———, *Convex cones, minimality notions and consequences*, in Recent Advances and Historical Development of Vector Optimization, J. Jahn and W. Krabs, eds., Springer-Verlag, New York, 1987, pp. 64–73.

[9] J. M. Borwein and D. Zhuang, *Super efficiency in vector optimization*, Trans. Amer. Math. Soc., to appear.

[10] J. P. Dauer and W. Stadler, *A survey of vector optimization in infinite dimensional spaces, Part* II, J. Optim. Theory Appl., 51 (1986), pp. 205–241.

[11] J. P. Dauer and O. A. Saleh, *A characterization of proper minimal points as solutions of sublinear optimization problems*, to appear.

[12] J. P. Dauer and R. J. Gallagher, *Positive proper efficient points and related cone results in vector optimization theory*, SIAM J. Control Optim., 28 (1990), pp. 158–172.

[13] A. M. Geoffrion, *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl., 22 (1968), pp. 618–630.

[14] R. Hartley, *On cone-efficiency, cone-convexity, and cone-compactness*, SIAM J. Appl. Math., 34 (1978), pp. 211– 222.

[15] M. I. Henig, *Proper efficiency with respect to cones*, J. Optim. Theory Appl., 36 (1982), pp. 387–407.

[16] R. B. Holmes, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, New York, 1975.

[17] L. Hurwicz, *Programming in linear spaces*, in Studies in Linear and Nonlinear Programming, K. Arrow, L. Hurwicz, and Uzawa, eds., Standford University Press, Standford, CA, 1958, pp. 38–102.

[18] G. Jameson, *Ordered linear spaces*, in Lecture Notes in Mathmatics 141, Springer-Verlag, Berlin, 1970.

[19] J. Jahn, *Mathematical Vector Optimization in Partially Ordered Linear Spaces*, Verlag Peter Lang, Frankfurt am Main, 1986.

[20] ———, *Scalarization in vector optimization*, Math. Programming, 29 (1984), pp. 203–218.

[21] ———, *A generalization of a theorem of Arrow, Barankin, and Blackwell*, SIAM J. Control Optim., 26 (1988), pp. 999–1005.

[22] H. W. Kuhn and A. W. Tucker, *Nonlinear programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, CA, 1951, pp. 481–492.

[23] A. L. Peressini, *Ordered Topological Vector Spaces*, Harper & Row, New York, 1967.

[24] M. Petschke, *On a theorem of Arrow, Barankin, and Blackwell*, SIAM J. Control Optim., 28 (1990), pp. 395–401.

[25] D. Zhuang, *Regularity and minimality properties of set-valued structures in optimization*, Ph.D. thesis, Dalhousie University, Halifax, 1989.

[26] ———, *Bases of convex cones and proper efficiency*, J. Optim. Theory Appl., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 59–85, January 1994

005

CONSUMPTION-INVESTMENT MODELS WITH CONSTRAINTS*

THALEIA ZARIPHOPOULOU†

Abstract. The paper examines a general investment and consumption problem for a single agent who consumes and invests in a riskless asset and a risky one. The objective is to maximize the total expected discounted utility of consumption. Trading constraints, limited borrowing, and no bankruptcy are binding, and the optimization problem is formulated as a stochastic control problem with state and control constraints. It is shown that the value function is the unique smooth the associated Hamilton–Jacobi–Bellman equation and the optimal consumption and portfolios are provided in feedback form.

Key words. dynamic programming, Bellman equation, viscosity solutions, state constraints, mathematical finance, investment and consumption models

AMS subject classifications. 49L20, 49L25, 90A09, 60H10

Introduction. This paper treats a general consumption and investment problem for a single agent. The investor consumes wealth X_t at a nonnegative rate C_t and distributes it between two assets continuously in time. One asset is a *bond*, i.e., a riskless security with instantaneous rate of return r. The other asset is a *stock* whose value is driven by a Wiener process.

The objective is to maximize the *total expected (discounted) utility from consumption* over an infinite trading horizon and the *total expected utility both from consumption and terminal wealth* in the case of finite horizon. The investor faces the following *trading constraints:* Wealth must stay nonnegative, i.e., bankruptcy never occurs, moreover, the amount π_t invested in stock must not exceed an exogenous function $f(X_t)$ of the wealth at any time t. The function f represents general borrowing constraints, which are frequently binding in practice, such as in portfolio insurance models with prespecified liability flow, models with nontraded assets, stochastic income and/or uninsurable risks, etc. The possibility of imposing short-selling constraints, which amounts to requiring $g(x_t) \leq \pi_t$ for some exogenous function g, is addressed in detail in §1. Finally, the agent is a "small investor," in that his or her decisions do not affect the asset prices and he or she does not pay transaction fees when trading.

This financial model gives rise to a stochastic control problem with control variables *consumption rate* C_t and *portfolio vector* (π_t^0, π_t), where π_t^0 and π_t are the amount of wealth invested in bond and stock, respectively. The state variable X_t is the total wealth at time t. Finally, the value function is the maximum total expected discounted utility.

The goal of this paper is to determine the value functions of these control problems, to examine how smooth they are, and to characterize the optimal policies. The basic tools come from the theory of partial differential equations, in particular the theory of viscosity solutions for second-order partial differential equations and elliptic regularity. We first show that the value functions are the unique constrained viscosity solutions of the associated Hamilton–Jacobi–Bellman (HJB) equation. Then we prove that viscosity solutions of these equations are smooth. Finally, we obtain an explicit feedback form for the optimal policies (C^*, π^*).

The paper is organized as follows: In §1 we describe the model and we give a summary of the history of consumption—investment models in continuous-time finance. Sections 2–5 deal with the infinite horizon model. More precisely, in §2 we describe basic properties of the value function, and in §3 we characterize the value function as a constrained viscosity

* Received by the editors September 9, 1991; accepted for publication (in revised form) May 21, 1992. This work was partially supported by National Science Foundation grant DMS-9009310.

† Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts 01609.

solution of the HJB equation. Moreover, in §4 we prove that the value function is the unique constrained solution of the HJB equation. In §5, we show that the value function is also a smooth solution of this equation and we provide the optimal policies. Finally, in §6 we state results for the finite horizon model.

1. We consider a market with two assets: A *bond* and a *stock*. The price P_t^0 of the bond is given by

$$\begin{aligned} dP_t^0 &= rP_t^0\,dt \qquad (t \geq 0) \\ P_0^0 &= p_0, \qquad (p_0 > 0), \end{aligned} \tag{1.1}$$

where $r > 0$ is the *interest rate*. The price P_t of the stock satisfies

$$\begin{aligned} dP_t &= bP_t\,dt + \sigma P_t\,dW_t \qquad (t \geq 0) \\ P_0 &= p, \qquad (p > 0), \end{aligned} \tag{1.2}$$

where b is the *mean rate of return*, σ is the *dispersion coefficient* and the process $W_\cdot$, which represents the source of uncertainty in the market, is a standard Brownian motion defined on the underlying probability space (Ω, F, P). We will denote by F_t the augmentation under P of $F_t^W = \sigma(W_s : 0 \leq s \leq t)$ for $0 < t < +\infty$. The interest rate r, the mean rate of return b, and the dispersion coefficient σ are assumed to be constant with $\sigma \neq 0$ and $b > r > 0$.

The total current *wealth* $X_t = \pi_t^0 + \pi_t$ is the state variable and π_t^0 and π_t are the amount of wealth invested in bond and stock, respectively; X_t evolves (see [40]) according to the equation

$$\begin{aligned} dX_t &= rX_t\,dt + (b-r)\pi_r\,dt - C_t\,dt + \sigma\pi_t\,dW_t \qquad (t \geq 0) \\ X_0 &= x, \qquad (x \in [0, +\infty)) \end{aligned} \tag{1.3}$$

where x is the *initial endowment* of the investor.

The *control process* are the consumption rate C_t and the portfolio π_t. To state their properties we introduce the following sets:

$$\begin{aligned} \mathcal{L}_+ = \Bigg\{ & z_t : z_t \text{ is } F_t\text{-progressively measurable process, } z_t \geq 0 \text{ a.s. } \forall t \geq 0 \\ & \text{and } \int_o^t z_s\,ds < +\infty \text{ a.s. } \forall t \geq 0 \Bigg\} \\ \mathcal{M} = \Bigg\{ & z_t : z_t \text{ is } F_t\text{-progressively measurable process} \\ & \text{and } \int_0^t z_s^2\,ds < +\infty \text{ a.s. } \forall t \geq 0 \Bigg\}. \end{aligned}$$

The set $\mathcal{A}_x$ of *admissible controls* for $x \in [0, +\infty)$ consists of all pairs (C, π) such that:

(i) $C \in \mathcal{L}_+$,

(ii) $\pi \in \mathcal{M}$.

Moreover, $\pi_t \leq f(X_t)$ almost surely for all $t \geq 0$, where the function $f : [0, +\infty) \to [0, +\infty)$ has the following properties:

$$\begin{aligned} & f \text{ is increasing, concave, } \quad f(0) \geq 0 \quad \text{and} \\ & |f(x) - f(y)| \leq K|x - y| \quad \forall x, y \geq 0 \end{aligned} \tag{1.4}$$

(iii) $X_t \geq 0$ almost surely for all $t \geq 0$, where X_t is the trajectory given by the state equation (1.3) using the controls (C, π).

The function f represents the borrowing constraints that the investor must meet; these constraints are present in models with prespecified liabilities such as problems of management of funds as well as in models with uninsurable risks. The possibility of short-selling constraints, i.e., $g(x) \leq \pi$, is not examined in this paper for the following reasons: First, if $g \leq 0$, the short-selling constraints can be removed because the model is of constant coefficients with $b > r$ (see, for example, [40] and [8]). Second, if $0 < g(x) \leq \pi$ this only facilitates the analysis presented here and therefore this case is not discussed.

All the results in this paper hold for the case $f \equiv \infty$, which was studied in [18], provided that some of the arguments in what follows are slightly modified. We will not pursue this any further in this paper unless it is necessary for the study of the $f \equiv \infty$ case. On the other hand, we will occasionally use some results of [18] only to facilitate the presentation and avoid lengthy arguments.

The *total expected discounted utility* J coming from consumption is given by

$$J(x, C, \pi) = E \int_0^{+\infty} e^{-\beta t} U(C_t)\, dt$$

with $(C, \pi) \in \mathcal{A}_x$, where Eg denotes the expectation of g with respect to the probability measure P, $\beta > 0$ is a *discount factor* such that

$$\beta > r, \tag{1.5}$$

and U is the *utility function,* which is assumed to have the following properties:

$$\begin{aligned} &U \text{ is a strictly increasing, concave } C^2(0, +\infty) \text{ function such that} \\ &U(c) \leq M(1 + c)^\gamma \quad \text{with } 0 < \gamma < 1 \quad \text{and} \quad M > 0, \\ &U(0) \geq 0, \quad \lim_{c \to 0} U'(c) = +\infty, \quad \lim_{c \to \infty} U'(c) = 0. \end{aligned} \tag{1.6}$$

The *value function* is given by

$$v(x) = \sup_{\mathcal{A}_x} E \int_0^{+\infty} e^{-\beta t} U(C_t)\, dt. \tag{1.7}$$

To guarantee that the value function is well defined when U is unbounded, we assume that

$$\beta > r\gamma + \gamma(b - r)/\sigma^2(1 - \gamma).$$

The above condition yields that the value function which corresponds to $f \equiv +\infty$ and $U(c) = M(1 + c)^\gamma$, and thereby all value functions, are finite (see [18]).

The goal is to characterize v as a classical solution of the HJB equation, associated with the control problem, and use the regularity of v to provide the optimal policies.

We now state the main results.

THEOREM 1.1. *The value function v is the unique $C^2((0, +\infty)) \cap C([0, +\infty))$ solution of*

$$\beta v = \max_{\pi \leq f(x)} \left[\tfrac{1}{2}\sigma^2\pi^2 v_{xx} + (b - r)\pi v_x\right] + \max_{c \geq 0}[-cv_x + U(c)] + rxv_x \tag{1.8}$$

in the class of concave functions.

THEOREM 1.2. *The optimal policies C_t^* and π_t^* are given in the feedback form $C_t^* = c^*(X_t), \pi_t^* = \pi^*(X_t)$ where*

$$c^*(x) = (U')^{-1}(v_x(x)) \quad \textit{and} \quad \pi^*(x) = \min\left\{f(x), -\frac{b - r}{\sigma^2}\frac{v_x(x)}{v_{xx}(x)}\right\}.$$

We continue with a brief discussion of the history of the model.

The single agent consumption-portfolio problem was first investigated by Merton in 1969 and 1971 ([28], [29]). He assumed that the returns of asset prices in perfect markets satisfy the "geometric Brownian motion" hypothesis and he considered utility functions belonging to the hyperbolic absolute risk aversion (HARA) family, i.e., $U(c) = 1 - \gamma/\gamma[\beta c/1 - \gamma + \eta]^{\gamma}$. Under these assumptions, he found explicit formulae for the optimal consumption and portfolio in both the finite and infinite horizon case. Moreover, he showed that the optimal policies are linear functions of the current wealth if and only if the utility function belongs to the HARA family.

In Merton's work, the portfolio is unconstrained, which means that unlimited borrowing and short selling are allowed. Moreover, the consumption process has to stay nonnegative and bankruptcy should never occur. Extra restrictions on the parameters β, γ, and η were later imposed by Merton [30] and Sethi and Taksar [34] to meet the above feasibility conditions.

Another important contribution is the work of Karatzas et al. [18], which is a continuation of work initiated by Lehoczky, Sethi, and Shreve [25]. Reference [18] examines a model with constant coefficients when borrowing and short selling are allowed (i.e., $f \equiv \infty$) and provides solutions of the Bellman equation in closed form. The possibility of bankruptcy is treated in this paper as well as in Sethi and Taksar [33]. The special case of a finite horizon model with constant market coefficients is examined by the same authors in [19]. The fact that borrowing and short selling are allowed is used strongly in [19] (see also [4]) to "linearize" the fully nonlinear Bellman equation to get a system of two linear parabolic equations. Solving these linear equations, they obtain a closed-form solution of the HJB equation.

The Bellman equation can be also linearized when only short-selling constraints are imposed; such a model was studied by Shreve and Xu [35], [36] and Xu [39] in a finite horizon setting in incomplete markets. Such linearization cannot be done if general borrowing constraints are imposed, which is the case we treat in this paper.

A different approach to studying investment-consumption problems with constraints in continuous-time finance was introduced by the author in [40], which studies an investment consumption model with borrowing and short-selling constraints, i.e., $0 \leq \pi_t \leq X_t$. This new approach is based on the theory of viscosity solutions of nonlinear first- and second-order partial differential equations and appears to be flexible enough to handle a wide variety of problems with constraints and related asymptotic problems, e.g., convergence of numerical schemes, asymptotic behavior, etc.

The asymptotic behavior of the value function and the optimal policies for the model with constraints and different interest rates were examined by Fleming and Zariphopoulou in [13]. Moreover, numerical results for the optimal policies and the value function were obtained by Fitzpatrick and Fleming in [10]. A consumption-investment model with leverage constraints (i.e., $f(x) = x + L, L > 0$) was examined by Vila and Zariphopoulou in [38].

Finally, a martingale representation technology has been used by Pliska [32], Cox and Huang [4], Pages [31], and Karatzas, Lehoczky, and Shreve [19] to study optimal portfolio and consumption policies in models with general market coefficients. Moreover, the case of incomplete markets with short-selling constraints in the finite horizon setting has been examined by He and Pearson [15], Xu [39], Shreve and Xu [35], [36], and in the absence of constraints by Karatzas et al. [20].

After this paper was submitted, the author received a paper by Cvitanic and Karatzas

[7]. This paper uses martingale and convex duality methods to study a finite horizon model with nonconstant coefficients and constrained portfolio policies but with utility functions which are more restrictive than the ones used in this paper; in particular, they only consider the case of utility functions with Arrow–Pratt index less than one.

2. In this section we derive some basic properties of the value function.

PROPOSITION 2.1. *The value function v is concave and strictly increasing.*

Proof. The concavity of v is an immediate consequence of the concavity of the utility function U and the fact that if $(C^1, \pi^1) \in \mathcal{A}_{x_1}$, $(C^2, \pi^2) \in \mathcal{A}_{x_2}$, and $\lambda \in (0,1)$, then $(\lambda C^1 + (1-\lambda)C^2, \lambda\pi^1 + (1-\lambda)\pi^2) \in \mathcal{A}_{\lambda x_1 + (1-\lambda)x_2}$; the latter follows from the linear dependence of the dynamics (1.3) with respect to the controls and the state variable.

That v is increasing follows from the observation that $\mathcal{A}_{x_1} \subset \mathcal{A}_{x_2}$ if $x_1 \leqq x_2$. If v is not strictly increasing, then it must be constant on an interval, which, by concavity, has to be of the form $[x_0, \infty)$ for some $x_0 \geqq 0$, i.e., there must exist $x_0 \in [0, +\infty)$ such that $v(x) = v(x_0)$, for all $x \geqq x_0$. In this case, fix $\epsilon > 0$ and choose $(C^\epsilon, \pi^\epsilon) \in \mathcal{A}_{x_0}$ such that

$$v(x_0) \leqq E \int_0^{+\infty} e^{-\beta t} U(C_t^\epsilon)\, dt + \epsilon.$$

If

$$x_1 > \max\left(x_0, \frac{U^{-1}\left[\beta\left(E\int_0^{+\infty} e^{-\beta t}U(C_t^\epsilon)\,dt + \epsilon\right)\right]}{r} \right),$$

the policy $(\overline{C}, \overline{\pi}) = (rx_1, 0)$ is in $\mathcal{A}_{x_1}$. Therefore

$$v(x_0) < \frac{1}{\beta}U(rx_1) = E\int_0^{+\infty} e^{-\beta t}U(rx_1)\,dt \leqq v(x_1),$$

which contradicts our assumption. □

PROPOSITION 2.2. *The value function v is uniformly continuous on $\overline{\Omega} = [0, \infty)$ and $v(0) = U(0)/\beta$.*

Proof. Since $(0,0) \in \mathcal{A}_0$, $v(0) \geqq U(0)/\beta$. On the other hand, $v \leqq u$ in $[0, +\infty)$, where u is the value function with $f \equiv +\infty$ studied in [16]. Since (cf. [18]) $u(0) = U(0)/\beta$ and $u \in C([0, +\infty))$, it follows that $v(0) = U(0)/\beta$ and v is continuous at $x = 0$. The continuity of v in $(0, +\infty)$ follows from concavity.

Finally, since v is uniformly continuous on compact subsets of $\overline{\Omega}$, we remark that its uniform continuity on $\overline{\Omega}$ follows from the fact that, by concavity, v is Lipschitz continuous in $[a, +\infty)$ with Lipschitz constant of order $1/a$ for every $a > 0$. □

PROPOSITION 2.3. *The value function satisfies $v(x) \leq 0(x^\gamma)$ as $x \to +\infty$.*

Proof. Since $v \leq u$ on $\overline{\Omega}$, where u is the value function with $f \equiv +\infty$ and $U(c) = M(1+x)^\gamma$, we only need to check this upper bound for u.

On the other hand, a direct modification of the proof of Theorem 4.5 in [13] yields that if $U \sim c^\gamma$ (as $c \to \infty$), then $u \sim x^\gamma$ (as $x \to \infty$). □

We conclude this section by stating (for a proof see [1], [26]) a fundamental property of the value function known as the *Dynamic Programming Principle.*

PROPOSITION 2.4. *If θ is a stopping time (i.e., a nonnegative $\mathcal{F}$-measurable random variable) then*

$$v(x) = \sup_{\mathcal{A}_x} E\left[\int_0^\theta e^{-\beta t}U(C_t)\,dt + e^{-\beta\theta}v(X_\theta)\right] (x \in \overline{\Omega}). \tag{2.1}$$

3. In this section we show that the value function v is a constrained viscosity solution of the HJB equation associated with the underlying stochastic control problem. The characterization of v as a constrained viscosity solution is natural because of the presence of the state $(X_t \geq 0)$ and control $(\pi_t \leq f(X_t))$ constraints.

The notion of viscosity solution was introduced by Crandall and Lions [6] for first-order and by Lions [27] for second-order equations. For a general overview of the theory we refer to the *User's Guide* by Crandall, Ishii, and Lions [5].

Next we recall the notion of contrained viscosity solutions, which was introduced by Soner [37] and Capuzzo-Dolcetta and Lions [3] for first-order equations (see also Ishii and Lions [16] and Katsoulakis [21]). To this end, consider a nonlinear second-order partial differential equation of the form

$$F(x, u, u_x, u_{xx}) = 0 \quad \text{in } \Omega, \tag{3.1}$$

where Ω is an open subset of $\mathbb{R}$ and $F : \Omega \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is continuous and (degenerate) elliptic, i.e.,

$$F(x, t, p, X + Y) \leq F(x, t, p, X) \quad \text{if } Y \geqq 0.$$

DEFINITION 3.1. A continuous function $u : \overline{\Omega} \to \mathbb{R}$ is a *constrained viscosity solution* of (3.1) if and only if

(i) u is a *viscosity subsolution* of (3.1) on $\overline{\Omega}$, i.e., if for any $\varphi \in C^2(\overline{\Omega})$ and any maximum point $x_0 \in \Omega$ of $u - \varphi$,

$$F(x_0, u(x_0), \varphi_x(x_0), \varphi_{xx}(x_0)) \leq 0;$$

and

(ii) u is a *viscosity supersolution* of (3.1) in Ω, i.e., if for any $\varphi \in C^2(\overline{\Omega})$ and any minimum point $x_0 \in \Omega$ of $u - \varphi$,

$$F(x_0, u(x_0), \varphi_x(x_0), \varphi_{xx}(x_0)) \geq 0.$$

Remark 1. We say that $u \in C(\overline{\Omega})$ is a viscosity solution of (3.1) in Ω if and only if it is both sub- and supersolution in Ω.

Remark 2. As a matter of fact, we can extend the definition of viscosity subsolutions (respectively, supersolutions) for upper-semicontinuous (respectively, lower-semicontinuous) functions.

THEOREM 3.1. *The value function v is a constrained viscosity solution of* (1.8) *on* $\overline{\Omega}$.

The fact that, in general, value functions of control problems and differential games turn out to be viscosity solutions of the associated partial differential equations is a direct consequence of the principle of dynamic programming and the definition of viscosity solutions (see, for example, Lions [26], Evans and Souganidis [9], Fleming and Souganidis [12], etc.). The main difficulty, however, in the problem at hand is that the consumption rates and the portfolios are not uniformly bounded. This gives rise to some serious complications in the proofs of the results of the aforementioned papers. To overcome these difficulties we need to introduce a number of approximations of the original problem and make use repeatedly of the *stability* properties of viscosity solutions.

Proof of Theorem 3.1. We first show that v is a viscosity supersolution of (1.8) in Ω.

Let $\varphi \in C^2(\overline{\Omega})$ and $x_0 \in \Omega$ be a minimum of $v - \varphi$; without any loss of generality, we may assume that

$$v(x_0) = \varphi(x_0) \quad \text{and} \quad v \geq \varphi \quad \text{in } \Omega. \tag{3.2}$$

We need to show that

$$\beta v(x_0) \geq \max_{\pi \leq f(x_0)} \left[\tfrac{1}{2}\sigma^2\pi^2\varphi_{xx}(x_0) + (b-r)\pi\varphi_x(x_0)\right] + \max_{c \geq o}[-c\varphi_x(x_0) + U(c)] + rx_0\varphi_x(x_0). \tag{3.3}$$

To this end, at $(C, \pi) \in \mathcal{A}_{x_0}$ such that $C_t = C_0, \pi_t = \pi_0 \leq f(x_0)$, for all $t \geq 0$. The dynamic programming principle, together with (3.2), yields

$$v(x_0) \geq E\left[\int_0^\theta e^{-\beta t}U(C_0)\,dt + e^{-\beta\theta}\varphi(X_\theta)\right] \tag{3.4}$$

where $X_\cdot$ is the trajectory given by (1.3) using the controls (C_0, π_0) and starting at x_0 and $\theta = \min(\tau, \frac{1}{n})$, with $n > 0$ and $\tau_\cdot = \inf\{t \geq 0 : X_t = 0\}$.

On the other hand, applying Itô's lemma to $g(t, X_t) = e^{-\beta t}\varphi(X_t)$, we get

$$E[e^{-\beta\theta}\varphi(X_\theta)] = v(x_0) + E\int_0^\theta e^{-\beta t}\left[-\beta\varphi(X_t) + \frac{1}{2}\sigma^2\pi_0^2\varphi_{xx}(X_t) + (b-r)\pi_0\varphi_x(X_t) - C_0\varphi_x(X_t) + rX_t\varphi_x(X_t)\right]dt.$$

Combining the above equality with (3.4) and using standard estimates from the theory of stochastic differential equations (see [14]), we get

$$E\int_0^\theta \left[-\beta v(x_0) + \frac{1}{2}\sigma^2\pi_0^2\varphi_{xx}(x_0) + (b-r)\pi_0\varphi_x(x_0) - C_0\varphi_x(x_0) + U(C_0) + rx_0\varphi_x(x_0)\right]ds + E\int_0^\theta h(s)\,ds \leq 0,$$

where $h(s) = 0(s)$. Dividing both sides by $E(\theta)$ and passing to the limit as $n \to \infty$ yields

$$\beta v(x_0) \geq \left[\tfrac{1}{2}\sigma^2\pi_0^2\varphi_{xx}(x_0) + (b-r)\pi_0\varphi_x(x_0)\right] + [-C_0\varphi_x(x_0) + U(C_0)] + rx_0\varphi_x(x_0),$$

for every pair of constant controls $(C_0, \pi_0), C_0 \geq 0$, and $\pi_0 \leq f(x_0)$; inequality (3.3) then follows easily.

We next show that v is a viscosity subsolution of (1.8) on $\overline{\Omega}$.

We first approximate v by a sequence of functions $(v_\epsilon^{N,n})$ defined by

$$v_\epsilon^{N,n}(x) = \sup_{\mathcal{A}_{N,n}} E\int_0^{+\infty} e^{-\beta t}\left[U(C_t) - \frac{1}{\epsilon}p(X_t)\right]dt, \qquad (x \in \mathbb{R})$$

for $\epsilon > \imath, N > 0, n > 0$, and $p(x) = \max(0, -x)$. The set of admissible policies $\mathcal{A}_{N,n}$ consists of all pairs (C, π) such that

(i) $C \in \mathcal{L}_+$ and $C_t \leq N$ almost surely for all $t \geq 0$,

(ii) $\pi \in \mathcal{M}$ and $-n \leq \pi_t \leq \hat{f}(X_t)$ almost surely for all $t \geq 0, n > 0$ where the function $\hat{f} : \mathbb{R} \to \mathbb{R}$ (denoted for convenience in the sequel by f) satisfies (1.4) and coincides with f on $[0, +\infty)$;

(iii) X_t is the trajectory given by the state equation (1.3) using the controls (C, π) and starting at $x \in \mathbb{R}$.

It follows from the dynamic programming principle and the definition of viscosity solution (see [27]), that $v_\epsilon^{N,n}$ is a viscosity solution of

$$\beta v_\epsilon^{N,n} = \max_{-n \leq \pi \leq f(x)}\left[\tfrac{1}{2}\sigma^2\pi^2 v_{\epsilon,xx}^{N,n} + (b-r)\pi v_{\epsilon,x}^{N,n}\right] + \max_{0 \leq c \leq N}[-cv_{\epsilon,x}^{N,n} + U(c)] + rxv_{\epsilon,x}^{N,n} - \frac{1}{\epsilon}p(x) \qquad (x \in \mathbb{R}).$$

We next observe that as $n \to \infty$,

$$v_\epsilon^{N,n} \to v_\epsilon^N, \text{ locally uniformly in } \mathbb{R},$$

(see [22, Chap. 6]) where

$$v_\epsilon^{N(x)} = \sup_{\mathcal{A}_N} E \int_0^{+\infty} e^{-\beta t}\left[U(C_t) - \frac{1}{\epsilon}p(X_t)\right] dt \qquad (x \in \mathbb{R})$$

and the set $\mathcal{A}_N$ of admissible policies is defined in the same way as $\mathcal{A}_{N,n}$, but without a lower bound on π.

It is immediate that

$$v_\epsilon^N \leq \frac{U(N)}{\beta} \quad \text{in } \mathbb{R} \tag{3.5}$$

and

$$v^N \leq v_\epsilon^N \quad \text{on } [0, +\infty), \tag{3.6}$$

where, for $x \in [0, +\infty)$,

$$v^N(X) = \sup_{\mathcal{A}_{x,N}} E \int_0^{+\infty} e^{-\beta t} U(C_t)\, dt \tag{3.7}$$

and

$$\mathcal{A}_{x,N} = \{(C, \pi) \in \mathcal{A}_x : C_t \leq N \text{ a.s. } \forall t \geq 0\}.$$

Moreover, the v_ϵ^N's are increasing and concave with respect to x. Both properties follow as in Proposition 2.1.

Finally, the stability property of viscosity solutions (see [27, Prop. I.3]) yields that v_ϵ^N is a viscosity solution of

$$\begin{aligned} \beta v_\epsilon^N = \max_{\pi \leq f(x)} \left[\tfrac{1}{2}\sigma^2\pi^2 v_{\epsilon,xx}^N + (b-r)\pi v_{\epsilon,x}^N\right] + \max_{0 \leq c \leq N}[-c v_{\epsilon,x}^N + U(c)] \\ + rx v_{\epsilon,x}^N - \tfrac{1}{\epsilon}p(x) \qquad (x \in \mathbb{R}). \end{aligned} \tag{3.8}$$

In the sequel we look at the behavior of the v_ϵ^N's on $[0, +\infty)$ as $\epsilon \to 0$. Since the only available bounds on the v_ϵ^N's are the ones stated above, we employ the limsup operation introduced by Barles and Perthame [2]. To this end, we define

$$v^{N,*}(x) = \limsup_{y \to x, \epsilon \to 0} v_\epsilon^N(y) \qquad (x \in [0, +\infty))$$

and we claim that

(i) $v^{N,*}$ is an upper semicontinuous viscosity subsolution on $\overline{\Omega}$ of

$$\begin{aligned} \beta v^{N,*} = \max_{\pi \leq f(x)} \left[\tfrac{1}{2}\sigma^2\pi^2 v_{xx}^{N,*} + (b-r)\pi v_x^{N,*}\right] \\ + \max_{0 \leq c \leq N}[-c v_x^{N,*} + U(c)] + rx v_x^{N,*} \qquad (x \in [0, +\infty)); \end{aligned} \tag{3.9}$$

and

(ii) $v^{N,*} = v^N$ on $\overline{\Omega}$.

We first observe that $v^{N,*}$ is increasing and concave on $\overline{\Omega}$. The first property is an immediate consequence of the definition. For the concavity we argue as follows: The concavity of $v^{N,*}$ in Ω follows from the fact that, since the v_ϵ^N's are concave in Ω and uniformly bounded on $\overline{\Omega}$, they converge, as $\epsilon \to 0$, locally uniformly to a concave function which actually coincides with $v^{N,*}$. It remains to show that

$$v^{N,*}((1-\lambda)x) \geq \lambda v^{N,*}(0) + (1-\lambda)v^{N,*}(x) \tag{3.10}$$

for $\lambda \in (0,1)$ and $x > 0$.

Let (ϵ_n) and $(y_n) \in \mathbb{R}$ be sequences such that, as $n \to \infty, \epsilon_n \to 0, y_n \to 0$, and $v^{N,*}(0) = \limsup_{y_n \to 0, \epsilon \to 0} v_\epsilon^N(y_n)$. The concavity of v_ϵ^N yields

$$v_{\epsilon_n}^N(\lambda y_n + (1-\lambda)x) \geq \lambda v_{\epsilon_n}^N(y_n) + (1-\lambda)v_{\epsilon_n}^N(x). \tag{3.11}$$

On the other hand,

$$v^{N,*}(x) = \lim_{\epsilon \to 0} v_\epsilon^N(x) \qquad (x \in (0, +\infty)). \tag{3.12}$$

Indeed, let $x \in [x_1, x_2]$ with $x_1 > 0$. The concavity of v_ϵ^N's and (3.5) yields that the v_ϵ^N are locally Lipschitz on $[x_1, x_2]$ with Lipschitz constant L independent of ϵ, i.e.,

$$v_\epsilon^N(y) \leq v_\epsilon^N(x) + L|y - x| \qquad (x, y \in [x_1, x_2]);$$

therefore

$$\limsup_{\epsilon \to 0, y \to x} v_\epsilon^N(y) \leq \limsup_{\epsilon \to 0} v_\epsilon^N(x).$$

Moreover,

$$\limsup_{\epsilon \to 0} v_\epsilon^N(x) = \lim_{\epsilon \to 0} v_\epsilon^N(x)$$

since the v_ϵ^N's are increasing in ϵ. Combining the last inequalities we get

$$v^{N,*}(x) \leq \lim_{\epsilon \to 0} v_\epsilon^N(x)$$

which, together with the definition of $v^{N,*}$ yields (3.12).

We now observe that, for n large enough, $\lambda y_n + (1-\lambda)x > a$, for some $a > 0$. Sending $n \to \infty$ in (3.11) and using the properties of $(\epsilon_n), (y_n)$, and (3.12) we conclude.

We continue with the proof of (3.9). We need to examine the following cases.

Case 1. $f \equiv \infty$.

Let $\varphi \in C^2(\overline{\Omega})$ and assume that $v^{N,*} - \varphi$ has a maximum at 0, which can be assumed to be strict. We need to show

$$\beta v^{N,*}(0) \leq \max_{\pi} \left[\tfrac{1}{2}\sigma^2\pi^2\varphi_{xx}(0) + (b-r)\pi\varphi_x(0)\right] + \max_{0 \leq c \leq N}[-c\varphi_x(0) + U(c)]. \tag{3.13}$$

First observe that the concavity and monotonicity of $v^{N,*}$ imply $\varphi_x(0) > 0$. Inequality (3.5), along with the fact that the max with respect to π in (3.13) is unconstrained, implies

that (3.13) holds if $\varphi_{xx}(0) \geqq 0$. It remains to prove (3.13) if $\varphi_{xx}(0) < 0$. To this end, we first extend φ to $\mathbb{R}^-$ in $C^2(\mathbb{R})$ so that for some $\alpha > 0$,

$$\varphi_{xx}(x) < 0 \qquad (-\alpha \leq x \leq 0)$$

and

$$v_\epsilon^N(-\alpha) \leq v_\epsilon^N(0) - \varphi(0) + \varphi(-\alpha) - \alpha.$$

Let x_ϵ be a maximum point of $v_\epsilon^N - \varphi$ over $[-\alpha, \alpha]$. If $x_\epsilon = -\alpha$, the choice of φ together with (3.6) yields

$$v_\epsilon^N(-\alpha) - \varphi(-\alpha) \leq v_\epsilon^N(0) - \varphi(0) - \alpha$$

which is a contradiction. Moreover, 0 being a strict maximum of $v^{N,*} - \varphi$ yields $x_\epsilon \neq \alpha$ for ϵ small enough. Since v_ϵ^N is viscosity solution of (3.8) we have

$$\begin{aligned} \frac{1}{\epsilon} p(x_\epsilon) \leq & \max_{\pi} \left[\tfrac{1}{2}\sigma^2\pi^2\varphi_{xx}(x_\epsilon) + (b-r)\pi\varphi_x(x_\epsilon) \right] \\ & + \max_{0\leq c\leq N} [-c\varphi_x(x_\epsilon) + U(c)] + rx_\epsilon\varphi_x(x_\epsilon) - \beta v_\epsilon^N(x_\epsilon). \end{aligned} \tag{3.14}$$

We next observe that the right-hand side of the above inequality is finite since $\varphi_{xx}(x_\epsilon) < 0, \varphi_x(x_\epsilon) > 0$ and $-v_\epsilon^N(x_\epsilon) < +\infty$, where the latter follows from

$$v_\epsilon^N(x_\epsilon) - \varphi(x_\epsilon) \geq v_\epsilon^N(0) - \varphi(0) \geq v^N(0) - \varphi(0).$$

Let $\overline{x}$ be a limit (along subsequence) of the x_ϵ's. The definitions of p and (3.14) yield $\overline{x} \geq 0$. Actually, $\overline{x} = 0$.

Indeed,

$$v_\epsilon^N(x_\epsilon) - \varphi(x_\epsilon) \geq v_\epsilon^N(0) - \varphi(0)$$

and therefore

$$v^{N,*}(\overline{x}) - \varphi(\overline{x}) \geq v^{N,*}(0) - \varphi(0)$$

which yields $\overline{x} = 0$, since 0 is a strict maximum.

Moreover,

$$\lim_{\epsilon\to 0} v_\epsilon^N(x_\epsilon) = v^{N,*}(0).$$

Indeed, $\limsup_{\epsilon\to 0} v_\epsilon^N(x_\epsilon) \leq v^{N,*}(0)$. On the other hand, if $\limsup_{\epsilon\to 0} v_\epsilon^N(x_\epsilon) < v^{N,*}(0)$, then $v^{N,*}(0) - \varphi(0) > \limsup_{\epsilon\to 0}[v_\epsilon^N(x_\epsilon) - \varphi(x_\epsilon)]$, which is a contradiction. Finally, passing to the limit in (3.14) as $\epsilon \to 0$, we get (3.13).

Working similarly, we show that $v^{N,*}$ is a viscosity subsolution of (1.8) in $(0, +\infty)$.

It remains to show that

$$v^{N,*} = v^N \quad \text{on } [0, +\infty). \tag{3.15}$$

Since $v^{N,*}$ and v^N are, respectively, viscosity subsolution of (1.8) on $[0, +\infty)$ and supersolution in $(0, +\infty)$, a comparison result similar to Theorem 4.1 (easily modified for the case the consumption rates are uniformly bounded) implies

$$v^{N,*} \leqq v^N \quad \text{on } [0, +\infty) \tag{3.16}$$

which together with (3.6) yields (3.15) in $(0, +\infty)$.

Finally, the upper semicontinuity of $v^{N,*}$ implies $v^{N,*}(0) = v^N(0)$.

Case 2. $f < +\infty$.

In view of the analysis above, we only have to examine the case $\varphi_{xx}(0) = 0$, i.e., we need to show

$$\beta v^{N,*}(0) \leq (b-r)f(0)\varphi_x(0) + \max_{0 \leq c \leq N} [-c\varphi_x(0) + U(c)]$$

where we used that $\varphi_x(0) > 0$. We first observe

$$v^{N,*}(0) = \frac{U(0)}{\beta}.$$

This follows from the fact that $U(0)/\beta \leq v^{N,*}(0) \leq v^N_\infty(0)$ and $v^N_\infty(0) \leq u(0) = U(0)/\beta$, where v^N_∞ is given by (3.7) for $f \equiv \infty$.

Using that $\max_{0 \leq c \leq N}[-c\varphi_x(0) + U(c)] \geq U(0)$, (3.17), and $\varphi_x(0) > 0$, we conclude.

We now conclude the proof of the theorem.

In view of the stability properties of viscosity solutions, to conclude the proof of the theorem we only need to establish that as $N \to \infty$,

$$v^N \to v, \text{ locally uniformly on } \overline{\Omega}.$$

To this end, fix $x \in \overline{\Omega}, \epsilon > 0$, and choose $(C^\epsilon, \pi^\epsilon) \in \mathcal{A}_x$ such that

$$v(x) \leq E \int_0^{+\infty} e^{-\beta t} U(C^\epsilon_t)\, dt + \epsilon. \tag{3.17}$$

From the definitions of $\mathcal{A}_{x,N}$ and $\mathcal{A}_x$ we have that $(C^\epsilon \wedge N, \pi) \in \mathcal{A}_{x,N}$. Moreover, since U is increasing and nonnegative, the monotone convergence theorem yields

$$\lim_{N \to \infty} E \int_0^{+\infty} e^{-\beta t} U(C^\epsilon_t \wedge N)\, dt = E \int_0^{+\infty} e^{-\beta t} U(C^\epsilon_t)\, dt$$

which, combined with (3.17) and the definitions of v^N and v, gives

$$v^N(x) \leq v(x) \leq E \int_0^{+\infty} e^{-\beta t} U(C^\epsilon_t \wedge N)\, dt + 2\epsilon \leq v^N(x) + 2\epsilon \quad \text{for } N \geq N(\epsilon).$$

Therefore, $v^N \to v$ as $N \to \infty$, for each $x \in \overline{\Omega}$. On the other hand, since v^N increases with respect to N and v is continuous, Dini's theorem implies that $v^N \to v$ locally uniformly on $\overline{\Omega}$. □

4. In this section we present a comparison result for constrained viscosity solutions of (1.8). Comparison results for a large class of boundary problems were given by Ishii and Lions [16]. The equation on hand, however, does not satisfy some of the assumptions in [16], in view of the fact that the controls are not uniformly bounded. It is therefore necessary to modify some of the arguments of Theorem II.2 of [16] to take care of these difficulties. For completeness we present the whole proof; we rely, however, on some basic facts which are analyzed in [14].

THEOREM 4.1. *If u is an upper-semicontinuous concave viscosity subsolution of* (1.8) *on $\overline{\Omega}$ and v is a bounded from below, sublinearly growing, uniformly continuous on $\overline{\Omega}$, and locally Lipschitz in Ω supersolution of* (1.8) *on Ω, then $u \leq v$ on $\overline{\Omega}$.*

Before we begin with the proof of the theorem, we observe that (1.8) can be written as

$$\beta u = G(x, u_x, u_{xx}) \tag{4.1}$$

where $G : \overline{\Omega} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by

$$G(x,p,A) = \max_{\pi \le f(x)} \left[\tfrac{1}{2}\sigma^2\pi^2 A + (b-r)\pi p\right] + \max_{c \ge 0}[-cp + U(c)] + rxp.$$

An important ingredient of the proof of Theorem 4.1 is the sup- and inf-*convolution* approximation of u and v, respectively. Next we recall their definitions and summarize their main properties. For a general discussion about sup- and inf-convolution as well as their use in proving comparison results for second-order PDEs we refer to Lasry and Lions [24], Jensen, Lions, and Souganidis [17], and Ishii and Lions [16].

For $\epsilon > 0$ the ϵ sup-convolution of u is defined by

$$u^\epsilon(x) = \sup_{y \in \Omega}\left\{u(y) - \frac{1}{\epsilon}|x-y|^2\right\} \quad \forall x \in \overline{\Omega}, \tag{4.2}$$

and, similarly, the ϵ inf-convolution of v by

$$v_\epsilon(x) = \inf_{z \in \Omega}\left\{v(z) + \frac{1}{\epsilon}|x-z|^2\right\} \quad \forall x \in \overline{\Omega}. \tag{4.3}$$

It follows that the sup and inf in the definitions of u^ϵ and v_ϵ are actually taken for

$$|x-y| \le C\sqrt{\epsilon} \quad \text{and} \quad |x-z| \le C\sqrt{\epsilon}, \tag{4.4}$$

where $C = C(x)$ depends on the coefficients of the sublinear growth of u and v.

Moreover

(i) u^ϵ is a viscosity subsolution of

$$F'_\epsilon(x, u^\epsilon, u^\epsilon_x, u^\epsilon_{xx}) = 0 \quad \text{in } \Omega_\epsilon,$$

where $F^\epsilon(x,t,p,A) = \min\{\beta t - G(y,p,A) : |y-x| \le C\sqrt{\epsilon}\}$ and $\overline{\Omega}_\epsilon = \{x \in \Omega : x \ge C\sqrt{\epsilon}\}$; and

(ii) v_ϵ is a viscosity supersolution of

$$F^\epsilon(x, v_\epsilon, v_{\epsilon,x}, v_{\epsilon,xx}) = 0 \quad \text{in } \Omega_\epsilon,$$

where

$$F^\epsilon(x,t,p,A) = \max\{\beta t - G(y,p,A) : |y-x| \le C\sqrt{\epsilon}\}.$$

Proof of Theorem 4.1. We present the proof of the theorem for the case $f < +\infty$. The case $f \equiv \infty$ is discussed at the end of the section.

We argue by contradiction, i.e., we assume that

$$\sup_{x \in \overline{\Omega}}[u(x) - v(x)] > 0. \tag{4.5}$$

Then for sufficiently small $\theta > 0$

$$\sup_{x\in\overline{\Omega}}[u(x) - v(x) - \theta x] > 0. \tag{4.6}$$

Indeed, if not, there would be a sequence $\theta_n \downarrow 0$ such that $\sup_{x\in\overline{\Omega}}[u(x) - v(x) - \theta_n x] \leq 0$, which in turn would yield $\sup_{x\in\overline{\Omega}}[u(x) - v(x)] \leq 0$, contradicting (4.5).

Since u has, by concavity, sublinear growth and v is bounded from below, there exists $\overline{x} \in \overline{\Omega}$ such that

$$\sup_{x\in\overline{\Omega}}[u(x) - v(x) - \theta x] = u(\overline{x}) - v(\overline{x}) - \theta\overline{x}. \tag{4.7}$$

Next, for $\delta > 0$ and $\eta > 0$ we define $\varphi : \overline{\Omega} \times \overline{\Omega} \to \mathbb{R}$ by

$$\varphi(x, y) = u(x) - v(y) - \left|\frac{y - x}{\delta} - 4\eta\right|^4 - \theta x$$

and observe that for each fixed η, φ attains its maximum at a point (x_0, y_0) such that for δ small and some $l = l(\theta) > 0$,

$$|y_0 - x_0| \leq l\delta. \tag{4.8}$$

Indeed, φ is bounded and

$$\sup_{\overline{\Omega}\times\overline{\Omega}} \varphi(x, y) \geq \varphi(x, x + 4\eta\delta) \geq u(\overline{x}) - v(\overline{x}) - \theta\overline{x} - \omega_v(k\eta\delta) \tag{4.9}$$

where ω_v is the modulus of continuity of v and $k > 0$. Using (4.6) and (4.7) we get

$$\sup_{\overline{\Omega}\times\overline{\Omega}} \varphi(x, y) > 0 \tag{4.10}$$

for δ and η sufficiently small.

Next, let (x_n, y_n) be a maximizing sequence for φ and observe that

$$u(x_n) - v(y_n) - \theta x_n \geq \left|\frac{y_n - x_n}{\delta} - 4\eta\right|^4 \quad \text{as } n \to \infty.$$

The last inequality, combined with the fact that u has sublinear growth, implies (4.8).

On the other hand, the choice of (x_n, y_n) and (4.10) yields that the sequence (x_n) and, in view of the above observation, (y_n) are bounded as $n \to \infty$. Hence, along subsequences, (x_n, y_n) converges to a maximum point of φ, which we denote by (x_0, y_0).

We next fix δ small enough and we consider for $\epsilon \in (0, 1)$ the function

$$\varphi^\epsilon(x, y) = u^\epsilon(x) - v_\epsilon(y) - \left|\frac{y - x}{\delta} - 4\eta\right|^4 - \theta x$$

where u^ϵ and v_ϵ are, respectively, the ϵ sup- and inf-convolutions of u and v given by (4.3) and (4.4).

In the sequel we need to study separately the cases $\overline{x} > 0$ and $\overline{x} = 0$.

Case A. $\overline{x} > 0$.

If δ is small enough it follows that the point (x_0, y_0) lies in a fixed compact subset of $\Omega \times \Omega$. Moreover, the function φ^ϵ achieves its maximum at a point that we denote by

$(\overline{x}_\epsilon, \overline{y}_\epsilon)$, which lies in $\Omega_\epsilon \times \Omega_\epsilon$ (see [16]). Since $\beta > 0$ we can apply Proposition II.3 of [14], according to which there exist $X_\epsilon, Y_\epsilon \in \mathbb{R}$ such that

$$F_\epsilon(\overline{x}_\epsilon, u^\epsilon(\overline{x}_\epsilon), w_x(\overline{x}_\epsilon), \overline{y}_\epsilon) + \theta, X_\epsilon) \leq 0, \qquad F^\epsilon(\overline{y}_\epsilon, v_\epsilon(\overline{y}_\epsilon), -w_y(\overline{x}_\epsilon, \overline{y}_\epsilon), -Y_\epsilon) \geq 0$$

and

$$\begin{pmatrix} X_\epsilon & 0 \\ 0 & Y_\epsilon \end{pmatrix} \leq \begin{pmatrix} w_{xx}(\overline{x}_\epsilon, \overline{y}_\epsilon) & w_{xy}(\overline{x}_\epsilon, \overline{y}_\epsilon) \\ w_{yx}(\overline{x}_\epsilon, \overline{y}_\epsilon) & w_{yy}(\overline{x}_\epsilon, \overline{y}_\epsilon) \end{pmatrix}$$

where $w(x, y) = |(y - x)/\delta - 4\eta|^4$. Therefore

$$F_\epsilon\left(\overline{x}_\epsilon, u^\epsilon(\overline{x}_\epsilon), -\frac{4}{\delta}\left(\frac{\overline{y}_\epsilon - \overline{x}_\epsilon}{\delta} - 4\eta\right)^3 + \theta, X_\epsilon\right) \leq 0, \tag{4.11}$$

$$F^\epsilon\left(\overline{y}_\epsilon, v_\epsilon(\overline{y}_\epsilon), -\frac{4}{\delta}\left(\frac{\overline{y}_\epsilon - \overline{x}_\epsilon}{\delta} - 4\eta\right)^3, -Y_\epsilon\right) \geq 0. \tag{4.12}$$

Also,

$$\begin{pmatrix} X_\epsilon & 0 \\ 0 & Y_\epsilon \end{pmatrix} \leq \frac{12}{\delta^2}\left(\frac{\overline{y}_\epsilon - \overline{x}_\epsilon}{\delta} - 4\eta\right)^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

and therefore

$$X_\epsilon + Y_\epsilon \leq 0. \tag{4.13}$$

We next observe that there exists a constant $c \geq 0$ such that

$$Y_\epsilon \geq c. \tag{4.14}$$

We argue by contradiction. Let us assume that there exists a subsequence (Y_{ϵ_n}) such that $\lim_{\epsilon_n \to 0} Y\epsilon_n = Y < 0$. From the definition of F_ϵ we have

$$\begin{aligned} \beta v_\epsilon(\overline{y}\epsilon) \geq \max_{\pi \leq f(\widehat{y}_\epsilon)} \left[-\tfrac{1}{2}\sigma^2\pi^2 Y_{\epsilon_n} - (b - r)\pi w_y(\overline{x}_\epsilon, \overline{y}_\epsilon)\right] \\ + \max_{c \geq 0}[cw_y(\overline{x}_\epsilon, \overline{y}_\epsilon) + U(c)] - r\,\widehat{y}_\epsilon w_y(\overline{x}_\epsilon, \overline{y}_\epsilon), \end{aligned}$$

for some $\widehat{y}_\epsilon \in \Omega$ such that $|\overline{y}_\epsilon - \widehat{y}_\epsilon| \leq C\sqrt{\epsilon}$. Sending $\epsilon_n \downarrow 0$ and using that $v_\epsilon(\overline{y}_\epsilon) \leq U(N)/\beta$, we get a contradiction.

Therefore, there exists $Y \in \mathbb{R}_0^+$ or $Y = +\infty$ such that $\lim_{\epsilon \to 0} Y_\epsilon = Y$ (along a subsequence). Moreover, (4.13) and (4.14) imply that there exists $X \in \mathbb{R}_0^-$ or $X = -\infty$ such that $\lim_{\epsilon \to 0} X_\epsilon = X$ (along a subsequence).

Sending $\epsilon \to 0$, inequalities (4.11) and (4.12) yield (see [16])

$$\begin{aligned} \beta u(x_0) \leq \max_{\pi \leq f(x_0)} \left[\tfrac{1}{2}\sigma^2\pi^2 X + (b - r)\pi(w_x(x_0, y_0) + \theta)\right] \\ + g(w_x(x_0, y_0) + \theta) + rx_0(w_x(x_0, y_0) + \theta) \end{aligned} \tag{4.15}$$

and

$$\begin{aligned} \beta v(y_0) \geq \max_{\pi \leq f(y_0)} \left[-\tfrac{1}{2}\sigma^2\pi^2 Y + (b - r)\pi w_x(x_0, y_0)\right] \\ + g(w_x(x_0, y_0)) + ry_0 w_x(x_0, y_0) \end{aligned} \tag{4.16}$$

where we used that $w_x(x_0, y_0) = -w_y(x_0, y_0)$ and $g(p) = \max_{c \geq 0}[-cp + U(c)]$.

We now look at the following cases.

Case (i). $X = -\infty$. Inequalities (4.15) and (4.16) yield

$$\beta u(x_0) \leq g(w_x(x_0, y_0) + \theta) + rx_0(w_x(x_0, y_0) + \theta)$$

and

$$\beta v(y_0) \geq g(w_x(x_0, y_0)) + ry_0 w_x(x_0, y_0).$$

Therefore,

$$\beta(u(x_0) - v(y_0) - \theta x_0) \leq r(x_0 - y_o)w_x(x_0, y_0) \tag{4.17}$$

where we used that g is a decreasing function and (1.5).

Case (ii). $X \leq 0$. From (1.4), (4.8), and (4.15) we get

$$\begin{aligned}\beta u(x_0) \leq \max_{\pi \leq f(y_0)+Kl(\theta)\delta} \left[\tfrac{1}{2}\sigma^2\pi^2 X + (b-r)\pi w_x(x_0, y_0)\right] + g(w_x(x_0, y_0) + \theta) \\ + rx_0(w_x(x_0, y_0) + \theta) + (b-r)\theta f(x_0)\end{aligned}$$

which, combined with (4.16), gives

$$\begin{aligned}\beta[u(x_0) - v(y_0) - \theta_{x_0}] \leq & \max_{\pi \leq f(y_0)+Kl(\theta)\delta} \left[\tfrac{1}{2}\sigma^2\pi^2 X + (b-r)\pi w_x(x_0, y_0)\right] \\ & - \max_{\pi \leq f(y_0)} \left[\tfrac{1}{2}\sigma^2\pi^2 y + \pi w_x(x_0, y_0)\right] \\ & + r(x_0 - y_0)w_x(x_0, y_0) + (b-r)\theta f(x_0).\end{aligned} \tag{4.18}$$

In the sequel we will need the following two lemmas.

LEMMA 4.1. *Let $p > 0$ and $X \leq 0, Y \geq 0$ be such that*

$$\begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} \leq A \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{4.19}$$

with $A > 0$. Then

$$\begin{aligned}\max_{\pi \leq a_1} \left[\tfrac{1}{2}\sigma^2\pi^2 X + (b-r)\pi p\right] - \max_{\pi \leq a_2} \left[-\tfrac{1}{2}\sigma^2\pi^2 Y + (b-r)\pi p\right] \\ \leq \omega((a_1 - a_2)^2 A + (a_1 - a_2)p)\end{aligned} \tag{4.20}$$

where $a_1 > a_2$ and $\omega : [0, +\infty) \to [0, +\infty)$ is uniformly continuous with $\omega(0) = 0$.

LEMMA 4.2. *For fixed $\eta > 0$ and $\theta > 0$, the following holds:*

$$\lim_{\delta \downarrow 0} \left| \frac{y_0 - x_0}{\delta} - 4\eta \right| = 0. \tag{4.21}$$

Moreover,

$$\lim_{\theta \downarrow 0} \lim_{\delta \downarrow 0} \theta f(x_0(\theta, \delta)) = 0. \tag{4.22}$$

We now conclude the proof of the theorem and next give the proof of the lemmas. First observe that (4.17) gives

$$\beta(u(\overline{x}) - v(\overline{x}) - \theta\overline{x}) \leq \left[-4r\left(\frac{y_0 - x_0}{\delta}\right)\left(\frac{y_0 - x_0}{\delta} - 4\eta\right)^3 + \omega_v(k\eta\delta) + \left(\frac{y_0 - x_0}{\delta} - 4\eta\right)\right]^4.$$

Sending $\delta \to 0$ and using (4.21) we contradict (4.6).

Next, we use (4.18) and Lemma 4.2 with

$$A = \frac{12}{\delta^2}\left(\frac{y_0 - x_0}{\delta} - 4\eta\right)^2, \quad a_1 = f(y_0) + Kl(\theta)\delta, \quad a_2 = f(y_0)$$

and

$$p = -\frac{4}{\delta}\left(\frac{y_0 - x_0}{\delta} - 4\eta\right)^3.$$

Note that from the definition of g and (4.16) we must have $p = w_x(x_0, y_0) > 0$. We get

$$\begin{aligned}\beta[u(\overline{x}) - v(\overline{x}) - \theta\overline{x}] \leq \omega\left(12K^2l^2(\theta)\left(\frac{y_0 - x_0}{\delta} - 4\eta\right)^2 + 4Kl(\theta)\left|\frac{y_0 - x_0}{\delta} - 4\eta\right|^3\right) \\ + 4r\left|\frac{y_0 - x_0}{\delta}\right|\left|\frac{y_0 - x_0}{\delta} - 4\eta\right|^3 + (b - r)\theta f(x_0) + \omega_v(k\eta\delta) + \left(\frac{y_0 - x_0}{\delta} - 4\eta\right)^4.\end{aligned} \tag{4.23}$$

We now use (4.21) and (4.22) and we send first $\delta \downarrow 0$, then $\theta \downarrow 0$, and last $\eta \downarrow 0$ to contradict (4.3).

Case B. $\overline{x} = 0$.

Since the proof follows along the lines of Theorem VI.5 in [16], modified with arguments similar to the ones in Case A, we only present the main steps.

First, we work as in Theorem VI.5 in [16], with $h \equiv -\infty$ and w as before, to get the existence of $X_\epsilon, Y_\epsilon \in \mathbb{R}$ such that

$$\beta u^\epsilon(\overline{x}_\epsilon) \leq G(\widehat{x}_\epsilon, w_x(\overline{x}_\epsilon, \overline{y}_\epsilon) + \theta, X_\epsilon),$$

$$\beta v_\epsilon(\overline{y}_\epsilon) \geq G(\widehat{y}_\epsilon, -w_y(\overline{x}_\epsilon, \overline{y}_\epsilon), -y_\epsilon),$$

and

$$\begin{bmatrix} X_\epsilon & 0 \\ 0 & Y_\epsilon \end{bmatrix} \leq \begin{bmatrix} w_{xx}(\overline{x}_\epsilon, \overline{y}_\epsilon) & w_{xy}(\overline{x}_\epsilon, \overline{y}_\epsilon) \\ w_{yx}(\overline{x}_\epsilon, \overline{y}_\epsilon) & w_{yy}(\overline{x}_\epsilon, \overline{y}_\epsilon) \end{bmatrix}$$

for some $\widehat{x}_\epsilon, \widehat{y}_\epsilon \in \mathbb{R}$, where $|\widehat{x}_\epsilon - \overline{x}_\epsilon| \leq C\sqrt{\epsilon}$ and $|\widehat{y}_\epsilon - y_\epsilon| \leq C\sqrt{\epsilon}$ for some positive constant C independent of ϵ, δ, and η.

Next, working as in Case A we pass to the limit as $\epsilon \downarrow 0$ and using Lemma 3.2 we derive (4.23) for $\overline{x} = 0$. Finally, we use (4.21) and (4.22) and we send first $\theta \downarrow 0$, then $\delta \downarrow 0$ and last $\eta \downarrow 0$ to conclude.

Proof of Lemma 4.1. We first observe that (4.18) yields

$$X + Y \leq 0. \tag{4.24}$$

Let π_1^* and π_2^*, satisfying $\pi_1^* \leq a_1$ and $\pi_2^* \leq a_2$, be the points where the constrained maxima of $[\frac{1}{2}\sigma^2\pi^2 X + (b - r)\pi p]$ and $[-\frac{1}{2}\sigma^2\pi^2 Y + (b - r)\pi p]$ are, respectively, achieved. We look at the following cases.

Case 1. $\pi_1^* = a_1$ and $\pi_2^* = a_2$. Then

$$\begin{aligned}&\max_{\pi \leq a_1}\left[\tfrac{1}{2}\sigma^2\pi^2 X + (b - r)\pi p\right] - \max_{\pi \leq a_2}\left[-\tfrac{1}{2}\sigma^2\pi^2 Y + (b - r)\pi p\right] \\ &\quad = \tfrac{1}{2}\sigma^2(a_1^2 X + a_2^2 Y) + (b - r)p(a_1 - a_2).\end{aligned} \tag{4.25}$$

Multiplying both sides of (4.19) by the positive definite matrix

$$\Sigma = \begin{bmatrix} a_1^2 & a_1 a_2 \\ a_1 a_2 & a_2^2 \end{bmatrix},$$

and taking the trace, yields

$$\tfrac{1}{2}(a_1^2 X + a_2^2 Y) \leq \tfrac{1}{2}(a_1 - a_2)^2 A$$

which, combined with (4.25), gives (4.20).

Case 2.

$$\pi_1^* = -\frac{(b-r)p}{\sigma^2 X} \qquad \text{and} \qquad \pi_2^* = \frac{(b-r)p}{\sigma^2 Y}.$$

Then, the maxima are unconstrained and we easily get

$$\max_{\pi}\left[\tfrac{1}{2}\sigma^2\pi^2 X + (b-r)\pi p\right] - \max_{\pi}\left[-\tfrac{1}{2}\sigma^2\pi^2 Y + (b-r)\pi p\right] \leq \max_{\pi}\left[\tfrac{1}{2}\sigma^2\pi^2(X+Y)\right] = 0$$

where we used (4.24).

Case 3. $\pi_1^* = -(b-r)p/\sigma^2 X < a_1$ and $\pi_2^* = a_2$. Then

$$\max_{\pi \leq a_1}\left[\frac{1}{2}\sigma^2\pi^2 X + (b-r)\pi p\right] = -\frac{(b-r)^2 p^2}{2\sigma^2 X} < \frac{(b-r)a_1}{2}p$$

and

$$\max_{\pi \leq a_2}\left[-\frac{1}{2}\sigma^2\pi^2 Y + (b-r)\pi p\right] = -\frac{1}{2}\sigma^2 a_2^2 Y + (b-r)a_2 p.$$

If $Y = 0$, (4.20) follows immediately. If $Y > 0$, then by assumption, $(b-r)p/\sigma^2 Y > a_2$. Therefore,

$$\begin{aligned} \frac{(b-r)}{2}a_1 p - \left[-\frac{1}{2}\sigma^2 a_2^2 Y + (b-r)a_2 p\right] &= \frac{(b-r)}{2}p(a_1 - a_2) + \frac{1}{2}\sigma^2 a_2^2 Y - \frac{(b-r)}{2}a_2 p \\ &= \frac{(b-r)}{2}p(a_1 - a_2) + \frac{1}{2}a_2[\sigma^2 a_2 Y - (b-r)p] \\ &< \frac{(b-r)}{2}p(a_1 - a_2). \end{aligned}$$

Case 4. $\pi_1^* = a_1$ and $\pi_2^* = (b-r)p/\sigma^2 Y$. This case is easily reduced to Case 2. □

Proof of Lemma 4.2. Relation (4.21) follows directly from (4.7) and (4.8). To show (4.22), since f is Lipschitz, it suffices to show $\lim_{\theta \downarrow 0} \lim_{\delta \downarrow 0} \theta x_0(\theta, \delta) = 0$. Indeed, from (4.7) we have

$$u(x_0) - v(x_0) - \theta x_0(\theta, \delta) \geq [u(\overline{x}) - v(\overline{x}) - \theta\overline{x}] - \omega_v(kl(\theta)\delta) - \omega_v(k\eta\delta) \tag{4.26}$$

which in turn implies (for fixed θ) $\sup_{\delta > 0} x_0(\theta, \delta) < +\infty$.

Therefore, there exists $\hat{x}_0(\theta)$ such that $\lim_{\delta \to 0} x_0(\theta, \delta) = \hat{x}_0(\theta)$. The limit here is taken along subsequences which, to simplify notation, we denote the same way as the whole family. By sending $\delta \to 0$, (4.26) combined with (4.5) implies

$$u(\overline{x}_0) - v(\overline{x}_0) - \theta\overline{x}_0 \geq [u(x) - v(x) - \theta x] \quad \forall x \in \overline{\Omega}. \tag{4.27}$$

We now send $\theta \to 0$. If $\lim_{\theta \downarrow 0} \theta \hat{x}_0 = \alpha \neq 0$, again along subsequences, (4.27) yields $\sup_{\overline{\Omega}}[u - v] - \alpha \geq \sup_{\overline{\Omega}}[u - v]$ which implies $\sup_{\overline{\Omega}}[u - v] \leq 0$, which contradicts (4.5). □

Remark. In the case $f \equiv \infty$, we assume

$$\sup[u(x) - v(x) - \theta x^{\delta}] > 0$$

for some $\delta \in (\gamma, 1)$, and we argue as before.

5. In this section we show that the value function is a smooth solution of the Hamilton–Jacobi–Bellman equation and we characterize the optimal policies.

THEOREM 5.1. *The value function v is the unique continuous on $[0, +\infty)$ and twice continuously differentiable in $(0, +\infty)$ solution of* (1.8) *in the class of concave functions.*

Before we go into the details of the proof of the theorem we describe the main ideas. We will work in intervals $(x_1, x_2) \subset [0, +\infty)$ and show that v solves a uniformly elliptic HJB equation in (x_1, x_2) with boundary conditions $v(x_1)$ and $v(x_2)$. Standard elliptic regularity theory (cf. Krylov [23]) and the uniqueness result about viscosity solutions will yield that v is smooth in (x_1, x_2).

We next explain how we come up with the uniformly elliptic HJB equation. Formally, according to the constraints, the optimal π^* is either $f(x)$, if

$$-\frac{b-r}{\sigma^2}\frac{v_x(x)}{v_{xx}(x)} \geq f(x) \quad \text{or} \quad -\frac{b-r}{\sigma^2}\frac{v_x(x)}{v_{xx}(x)}, \quad \text{if} \quad -\frac{b-r}{\sigma^2}\frac{v_x(x)}{v_{xx}(x)} \leq f(x).$$

In the second case, we want to get a positive lower bound of π^* in $[x_1, x_2]$. Since v_x is nonincreasing and strictly positive, it is bounded from below away from zero. Therefore, it suffices to find a lower bound for v_{xx}.

An important feature of the proof is the approximation of v by a family of smooth functions (v^{ϵ}) which are solutions of a suitably regularized equation. Next, we define the v^{ϵ}'s and discuss their main properties.

Let W_t^1 be a Wiener process which is independent of W_t and is defined on some probability space (Ω^1, F^1, P^1). We consider the process $\overline{W}_t = (W_t, W_t^1)$, which is a Wiener process, on $(\Omega \times \Omega^1, \overline{F}, P \times P^1)$ and $\overline{F} = \sigma(F \times F^1)$ where $\sigma(F \times F^1)$ is the smallest σ-algebra which contains $F \times F^1$. Let ϵ be a positive number. A real process $(C^{\epsilon}, \pi^{\epsilon})$ which is F_t-progressively measurable is called an *admissible* policy if:

(i) $C_t^{\epsilon} \geq 0$ a.s. $\forall t \geq 0$ and $\int_0^{+\infty} C_s^{\epsilon} ds < +\infty$ a.s.;

(ii) $\int_0^{+\infty} (\pi_s^{\epsilon})^2 ds < +\infty$ a.s. and $\pi_t^{\epsilon} \leq f(X_t^{\epsilon})$ a.s. $\forall t \geq 0$ where f satisfies (1.4);

(iii) $X_t^{\epsilon} \geq 0$ a.s. $\forall t \geq 0$, where X_t^{ϵ} is the trajectory given by the state equation

$$\begin{cases} dX_t^{\epsilon} = [rX_t^{\epsilon} + (b-r)\pi_t^{\epsilon} - C_t^{\epsilon}]dt + \sigma\pi_t^{\epsilon} dW_t + \sigma\epsilon X_t^{\epsilon} dW_t^1 \ (t > 0) \\ X_0^{\epsilon} = x \qquad (x \epsilon [0, +\infty)) \end{cases}$$

using the controls $(C^{\epsilon}, \pi^{\epsilon})$.

We denote by $\mathcal{A}_x^{\epsilon}$ the set of admissible policies. We define the value function v^{ϵ} by

$$v^{\epsilon}(x) = \sup_{\mathcal{A}_x^{\epsilon}} E \int_0^{+\infty} e^{-\beta t} U(C_t^{\epsilon}) dt,$$

where U is the usual utility function. Using arguments similar to those in Propositions 2.1 and 2.2 we can prove that v^ϵ is concave, strictly increasing in x, and uniformly continuous on $\overline{\Omega}$.

Using a variation of Theorems 3.1 and 4.1, we have that the value function v^ϵ is the unique constrained viscosity solution on $\overline{\Omega}$ of the equation

$$\beta v^\epsilon = \max_{\pi \le f(x)} \left[\frac{1}{2}\sigma^2(\pi^2 + \epsilon^2 x^2)v^\epsilon_{xx} + (b-r)\pi v^\epsilon_x\right] + \max_{c\ge 0}[-cv^\epsilon_x + U(c)] + rxv^\epsilon_x. \tag{5.1}$$

We next consider a sequence (v^ϵ_L) with

$$v^\epsilon_L(x) = \sup_{\mathcal{A}^{\epsilon,L}} E\int_0^{+\infty} e^{-\beta t}U(C_t)dt \qquad (x \in \mathbb{P}).$$

The set $\mathcal{A}^{\epsilon,L}$ of admissible policies consists of pairs $(C^{\epsilon,L}, \pi^{\epsilon,L})$ such that $C^{\epsilon,L}, \pi^{\epsilon,L}$ are F_t-progressively measurable satisfying (i), (ii), and also $-L \le \pi^{\epsilon,L}_t$ almost surely for all $t \ge 0$.

Working as in Theorems 3.1 and 4.1, we get that v^ϵ_L is the unique constrained viscosity solution of $\overline{\Omega}$ of

$$\begin{aligned}\beta v^\epsilon_L = &\max_{-L\le\pi\le f(x)}\left[\tfrac{1}{2}\sigma^2\pi^2 v^\epsilon_{L,xx} + (b-r)\pi v^\epsilon_{L,x}\right]\\ &+\max_{c\ge 0}[-cv^\epsilon_{L,x} + U(c)] + rxv^\epsilon_{L,x}\end{aligned} \tag{5.2}$$

and, also, the unique viscosity solution (see [16]) of

$$\begin{aligned}\beta u = &\max_{-L\le\pi\le f(x)}\left[\tfrac{1}{2}\sigma^2\pi^2 u_{xx} + (b-r)\pi u_x\right]\\ &+\max_{c\ge 0}[-cu_x + U(c)] + rxu_x(x\epsilon[x_1, x_2])\\ u(x_1) = v^\epsilon_L(x_1), &\qquad u(x_2) = v^\epsilon_L(x_2).\end{aligned} \tag{5.3}$$

On the other hand, (5.3) admits a unique smooth solution u which is also the unique viscosity solution; therefore, v^ϵ_L is smooth which, together with the fact that v^ϵ_L is increasing and concave, yields that v^ϵ_L is also smooth solution of

$$\begin{aligned}\beta u = &\max_{0\le\pi\le f(x)}\left[\tfrac{1}{2}\sigma^2\pi^2 u_{xx} + (b-r)\pi u_x\right]\\ &+\max_{c\ge 0}[-cu_x + U(c)] + rxu_x\\ u(x_1) = v^\epsilon_L(x_1), &\qquad u(x_2) = v^\epsilon_L(x_2).\end{aligned}$$

We next observe that there exists w concave such that $v^\epsilon_L \to w$, as $L \to \infty$, locally uniformly in $\overline{\Omega}$. Therefore, w is a constrained viscosity solution of (5.1) and, by uniqueness of viscosity solution, it coincides with v^ϵ. This also implies that v^ϵ is a viscosity solution of

$$\begin{aligned}\beta u = &\max_{0\le\pi\le f(x)}\left[\tfrac{1}{2}\sigma^2\pi^2 u_{xx} + (b-r)\pi u_x\right]\\ &+\max_{c\ge 0}[-cu_x + U(c)] + rxu_x\\ u(x_1) = v^\epsilon(x_1), &\quad u(x_2) = v^\epsilon(x_2).\end{aligned} \tag{5.4}$$

Equation (5.4) admits a unique smooth solution which is the unique viscosity solution. Therefore v^ϵ is smooth.

Proof. Consider an interval $[x_1, x_2] \subset [0, +\infty)$ and let $[\overline{x}_1, \overline{x}_2]$ and $[X_1, X_2]$ with $X_1 > 0$ be such that $[x_1, x_2] \subset [\overline{x}_1, \overline{x}_2] \subset [X_1, X_2]$. Since v is concave and increasing, its first and second derivatives exist almost everywhere. Without any loss of generality, we can assume that $v_x(X_1)$ and $v_x(X_2)$ exist. The reason for this will become apparent in the sequel.

We are now going to prove that the optimal portfolio π^ϵ of the approximating problem is bounded from below by a positive number which is independent of ϵ (of course it may depend on (x_1, x_2)). To this end, it suffices to show that

$$-\frac{b-r}{\sigma^2}\frac{v^\epsilon_x}{v^\epsilon_{xx}} \geq c > 0 \quad \text{on } [x_1, x_2].$$

We first show that $v^\epsilon \to v$ locally uniformly on $\overline{\Omega}$. Indeed, v^ϵ and v^ϵ_x are bounded locally by u and u/x where u is the value function with $f(x) = +\infty$ and $\epsilon = 0$. This follows from the fact that $v^\epsilon \leq u^\epsilon \leq u$ (where u^ϵ is the value function with $f(x) = +\infty$ and $\epsilon > 0$), which can be proved by using a similar comparison result as in Theorem 4.1. Therefore, there exists a subsequence $\{v^{\epsilon_n}\}$ and a function w such that $v^{\epsilon_n} \to w$ locally uniformly in $\overline{\Omega}$. Moreover, an argument similar to the proof of Proposition 2.2 yields that $\lim_{x\to 0} v^\epsilon(x) = U(0)/\beta$ uniformly in ϵ, therefore $w(0) = U(0)/\beta$.

Using a variation of Proposition I.3 in [27] we get that w is a constrained viscosity of (1.8). Moreover, since w is concave, using Theorem 4.1 we get that it is the unique constrained viscosity solution of (1.8). Therefore, we conclude that all subsequences have the same limit which coincides with v. Using that $v^\epsilon \to v$ locally uniformly and the fact that v^ϵ and v are concave, we get that $\lim_{\epsilon\to 0} v^\epsilon_x(x_0) = v_x(x_0)$ at any point x_0 where $v_x(x_0)$ exists. Taking into account that v^ϵ_x is nondecreasing in $[X_1, X_2]$, we conclude that there exist positive constants $R_1 = R_1([X_1, X_2])$ and $R_2 = R_2([X_1, X_2])$ such that

$$R_1 \leq v^\epsilon_x(x) \leq R_2 \quad \text{on } [x_1, x_2]. \tag{5.5}$$

We next show that there exists a constant $R = R([X_1, X_2])$ such that

$$|v^\epsilon_{xx}(x)| \geq R \quad \text{on } [x_1, x_2]. \tag{5.6}$$

To this end, let $\zeta : \mathbb{R}^+ \to \mathbb{R}^+$ be as follows:

(i) $\zeta \in C^\infty_0$ (i.e., ζ is a smooth function with compact support);

(ii) $\zeta \equiv 1$ on $[x_1, x_2]$, $\zeta \equiv 0$ on $\mathbb{R}\backslash[\overline{x}_1, \overline{x}_2]$, and $0 \leq \zeta \leq 1$ otherwise;

(iii) $|\zeta_x| \leq M\zeta^p$, $|\zeta_{xx}| \leq M\zeta^p$ with $0 < p < 1$ and $M > 0$.

From now on, for simplicity we suppress the ϵ-notation. We next consider a function $Z : [X_1, X_2] \to \mathbb{R}$ given by $Z(x) = \zeta^2 v^2_{xx} + \lambda v^2_x - \mu v$, where λ and μ are positive constants to be chosen later. We are interested in looking at the maximum of Z on $[X_1, X_2]$. The following cases can happen.

Case 1. The function Z attains its maximum at a point $x_0 \notin \text{supp } \zeta$. Then using (ii) and that $v > 0, v_x > 0$, we get

$$v^2_{xx}(x) \leq \lambda v^2_x(x_0) + \mu v(x) \quad \forall x \in [x_1, x_2]$$

which implies (5.6).

Case 2. The function Z attains its maximum at the point $x_0 \in \text{supp } \zeta$. In this case we have

$$Z_x(x_0) = 0 \quad \text{and} \quad Z_{xx}(x_0) \leq 0 \tag{5.7}$$

where

$$Z_x = 2\zeta\zeta_x v_{xx}^2 + 2\zeta^2 v_{xx}v_{xxx} + 2\lambda v_x v_{xx} - \mu v_x \tag{5.8}$$

and

$$\begin{aligned} Z_{xx} = {} & 2\zeta_x^2 v_{xx}^2 + 2\zeta\zeta_{xx}v_{xx}^2 + 8\zeta\zeta_x v_{xx}v_{xxx} + 2\zeta^2 v_{xxx}^2 \\ & + 2\zeta^2 v_{xx}v_{xxxx} + 2\lambda v_{xx}^2 + 2\lambda v_x v_{xxx} - \mu v_{xx}. \end{aligned} \tag{5.9}$$

We examine each of the following cases separately.

Case 2(a).

$$x_0 \in A_1 = \left\{ x \in [x_1, x_2] : -\frac{b-r}{\sigma^2}\frac{v_x^\epsilon(x)}{v_{xx}^\epsilon(x)} < f(x) \right\}.$$

In this case the Bellman equation has the form

$$\beta v = -\gamma\frac{v_x^2}{v_{xx}} + \frac{1}{2}\epsilon^2\sigma^2 x^2 v_{xx} - v_x I(v_x) + U(I(v_x)) + rxv_x \tag{5.10}$$

with $\gamma = (b-r)^2/2\sigma^2$. Here we used that $\max_{c\geq 0}[-cp + U(c)] = -pI(p) + U(I(p))$ with $I = (U')^{-1}$. After differentiating (5.10) twice and rearranging the terms we get:

$$\begin{aligned} -\gamma\frac{v_x^2 v_{xxxx}}{v_{xx}^2} - \frac{1}{2}\epsilon^2\sigma^2 x^2 v_{xxx} - \gamma\frac{v_x^2 v_{xxx}^2}{v_{xx}^3} = {} & -\beta v_{xx} \\ & + v_{xx}[2r - 2\gamma + \epsilon^2\sigma^2] + 2\gamma\frac{v_x v_{xxx}}{v_{xx}} - 3\gamma\frac{v_x^2 v_{xxx}^2}{v_{xx}^3} \\ & + v_{xxx}[rx - I(v_x) + 2\epsilon^2\sigma^2 x] - v_{xx}^2 I'(v_x). \end{aligned} \tag{5.11}$$

On the other hand, (5.7) yields

$$-\left[\gamma\frac{v_x^2(x_0)}{v_{xx}^2(x_0)} + \frac{1}{2}\epsilon^2\sigma^2 x_0^2\right] Z_{xx}(x_0) \geq 0$$

which, combined with (5.9), yields

$$\begin{aligned} -\left[\gamma\frac{v_x^2}{v_{xx}} + \frac{1}{2}\epsilon^2\sigma^2 x^2\right] Z_{xx} = {} & 2\zeta^2 v_{xx}\left[-\gamma\frac{v_x^2 v_{xxxx}}{v_{xx}^2} - \frac{1}{2}\epsilon^2\sigma^2 x^2 v_{xxxx} - \gamma\frac{v_x^2 v_{xxx}^2}{v_{xx}^3}\right] \\ & + 2\lambda v_x\left[-\gamma v_x - \frac{1}{2}\epsilon^2\sigma^2 x^2 v_{xx} - \gamma\frac{v_x^2 v_{xxx}}{v_{xx}^2}\right] + \mu\left[\gamma\frac{v_x^2}{v_{xx}} + \frac{1}{2}\epsilon^2\sigma^2 x^2 v_{xx}\right] \\ & - 2\gamma\zeta_x^2 v_x^2 - \epsilon^2\sigma^2 x^2\zeta_x^2 v_{xx}^2 - 2\gamma v_x^2\zeta\zeta_{xx} - x^2\epsilon^2\sigma^2\zeta\zeta_{xx}v_{xx}^2 - 8\gamma\zeta\zeta_x\frac{v_x^2 v_{xxx}}{v_{xx}} \\ & - 4x^2\epsilon^2\sigma^2\zeta\zeta_x v_{xx}v_{xxx} - \epsilon^2\sigma^2 x^2\zeta^2 v_{xxx}^2 - \lambda x^2\epsilon^2\sigma^2 v_{xx}^2 \geq 0, \end{aligned} \tag{5.12}$$

where all the expressions are evaluated at x_0.

Using (5.10), (5.11), and

$$\begin{aligned} -\gamma v_x - \frac{1}{2}\epsilon^2\sigma^2 x^2 v_{xxx} - \gamma\frac{v_x^2 v_{xxx}}{v_{xx}^2} = {} & v_{xx}[rx - I(v_x) + \epsilon^2\sigma^2 x] \\ & + v_x(r - 3\gamma) - \beta v_x \end{aligned} \tag{5.13}$$

which follows from differentiating (5.10) once and rearranging terms, we obtain from (5.12):

$$
\begin{aligned}
&-2\beta\zeta^2 v_{xx}^2 + 2\zeta^2(2r - 2\gamma + \epsilon^2\sigma^2)v_{xx}^2 + 4\gamma\zeta^2 v_x v_{xxx} - 6\gamma\zeta^2\frac{v_x^2 v_{xxx}^2}{v_{xx}^2} \\
&\quad + 2\zeta^2 v_{xx} v_{xxx}[rx - I(v_x) + 2\epsilon^2\sigma^2 x] - 2\zeta^2 v_{xx}^3 I'(v_x) \\
&\quad + 2\lambda v_x v_{xx}[rx - I(v_x) + \epsilon^2\sigma^2 x] + 2\lambda(r - 3\gamma)v_x^2 - 2\lambda\beta v_x^2 + 2\gamma\mu\frac{v_x^2}{v_{xx}} \\
&\quad + \mu v_x[-rx + I(v_x)] - \mu U(I(v_x)) \\
&\quad + \beta\mu v - 2\gamma\zeta_x^2 v_x^2 - \epsilon^2\sigma^2 x^2\zeta_x^2 v_{xx}^2 - 2\gamma v_x^2\zeta\zeta_{xx} - \epsilon^2\sigma^2 x^2\zeta\zeta_{xx}v_{xx}^2 - 8\gamma\zeta\zeta_x\frac{v_x^2 v_{xxx}}{v_{xx}} \\
&\quad - 4x^2\epsilon^2\sigma^2\zeta\zeta_x v_{xx}v_{xxx} - \epsilon^2\sigma^2 x^2\zeta_x^2 v_{xxx}^2 - \lambda x^2\epsilon^2\sigma^2 v_{xx}^2 \geq 0.
\end{aligned}
$$

A further calculation yields that at x_0,

$$
\begin{aligned}
&(rx - I(v_x) + \epsilon^2\sigma^2 x)(2\zeta^2 v_{xx}v_{xxx} + 2\lambda v_x v_{xx} - \mu v_x) + 2\zeta^2\epsilon^2\sigma^2 x v_{xx}v_{xxx} \\
&\quad + \epsilon^2\sigma^2\mu x v_x - 2\beta\zeta^2 v_{xx}^2 - 4\left(\gamma - r - \frac{\epsilon^2\sigma^2}{2}\right)\zeta^2 v_{xx}^2 + 4\gamma\zeta^2 v_x v_{xxx} \\
&\quad - 6K\zeta^2\frac{v_x^2 v_{xxx}^2}{v_{xx}^2} - 2\lambda\beta v_x^2 - 2\lambda(3\gamma - r)v_x^2 + \mu\beta v + 2\gamma\mu\frac{v_x^2}{v_{xx}} - \mu U(I(v_x)) \\
&\quad - 2\gamma\zeta_x^2 v_x^2 - 2\gamma\zeta\zeta_{xx}v_x^2 - 8\gamma\zeta\zeta_x\frac{v_x^2 v_{xxx}}{v_{xx}} + \epsilon^2\sigma^2 x^2[4\zeta\zeta_x|v_{xx}|v_{xxx} - \zeta^2 v_{xxx}^2 \\
&\quad - \lambda v_{xx}^2] - \epsilon^2\sigma^2 x^2\zeta_x^2 v_{xx}^2 - \epsilon^2\sigma^2 x^2\zeta\zeta_{xx}v_{xx}^2 \geq 2\zeta^2 v_{xx}^3 I'(v_x).
\end{aligned}
$$

(5.14)

If $A(x) = 2\zeta^2 v_{xx}v_{xxx} + 2\lambda v_x v_{xx} - \mu v_{x'}$ (5.7) and (5.8) imply

$$A(x_0) = -2\zeta(x_0)\zeta_x(x_0)v_{xx}^2(x_0).$$

Then (5.14) becomes

$$
\begin{aligned}
&-2\zeta\zeta_x v_{xx}^2(rx - I(v_x) + \epsilon^2\sigma^2 x) - 2\beta\zeta^2 v_{xx}^2 - 2\lambda\beta v_x^2 + \mu\beta v - 2\lambda(3\gamma - r)v_x^2 \\
&\quad -2\gamma\mu\frac{v_x^2}{|v_{xx}|} - \mu U(I(v_x)) - 2\gamma\zeta_x^2 v_x^2 - \epsilon^2\sigma^2 x^2\zeta_x^2 v_{xx}^2 - \epsilon^2\sigma^2 x^2\zeta\zeta_{xx}v_{xx}^2 - 2\gamma\zeta\zeta_{xx}v_x^2 \\
&\quad + \left[-4\left(\gamma - r - \frac{\epsilon^2\sigma^2}{2}\right)\zeta^2 v_{xx}^2 + 4\gamma\zeta^2 v_x v_{xxx} - 6\gamma\zeta^2\frac{v_x^2 v_{xxx}^2}{v_{xx}^2} + 8\gamma\zeta\zeta_x\frac{v_x^2 v_{xxx}}{|v_{xx}|}\right] \\
&\quad +\epsilon^2\sigma^2 x^2\left[\left(4\zeta\zeta_x - \frac{2\zeta^2}{x}\right)|v_{xx}|v_{xxx} - \zeta^2 v_{xxx}^2 - \lambda v_{xx}^2\right] + \epsilon^2\sigma^2\mu x v_x \\
&\quad \geq 2\zeta^2 v_{xx}^3 I'(v_x).
\end{aligned}
$$

(5.15)

Let

$$B(x) = -4\left(\gamma - r - \frac{\epsilon^2\sigma^2}{2}\right)\zeta^2 v_{xx}^2 + 4\gamma\zeta^2 v_x v_{xxx} - 6\gamma\zeta^2\frac{v_x^2 v_{xxx}^2}{v_{xx}^2} + 8\gamma\zeta\zeta_x\frac{v_x^2 v_{xxx}}{|v_{xx}|}$$

and

$$C(x) = \left(4\zeta\zeta_x - \frac{2\zeta^2}{x}\right)|v_{xx}|v_{xxx} - \zeta^2 v_{xxx}^2 - \lambda v_{xx}^2.$$

Let ϵ sufficiently small and $\theta \in (0, \frac{3}{4})$. Using the Cauchy–Schwartz inequality we get

$$B(x_0) \leq C\left(\frac{\zeta_x^2(x_0)}{\theta}v_x^2(x_0) + \zeta^2(x_0)v_{xx}^2(x_0)\right) \tag{5.16}$$

for some positive constant C.

A similar argument yields

$$C(x_0) \leq v_{xx}^2(x_0)\left[\left(2\zeta_x(x_0) - \frac{\zeta(x_0)}{x_0}\right)^2 - \lambda\right]. \tag{5.17}$$

We next choose λ so that $\lambda > 4\zeta_x^2 + (2/x_1^2)$ on $[\overline{x}_1, \overline{x}_2]$. Then $C(x_0) \leq 0$. If we leave out all the negative terms in (5.15) and use (5.16) and (5.17) we get

$$\begin{aligned} & v_{xx}^2(x_0)[-2\zeta(x_0)\zeta_x(x_0)[rx_0 - I(v_x(x_0))) + 2\epsilon^2\sigma^2 x_0] \\ & \quad -\epsilon^2 x_0^2\sigma^2\zeta(x_0)\zeta_{xx}(x_0) + C\zeta^2(x_0)] \\ & \quad +v_x^2(x_0)\left[C\frac{\zeta_x^2(x_0)}{\theta} - 2\gamma\zeta(x_0)\zeta_{xx}(x_0)\right] \\ & \quad +\mu[\beta v(x_0) + 2\epsilon\sigma^2 x_0 v_x(x_0))] \\ & \quad \geq 2\zeta(x_0)^2 v_{xx}^3(x_0) I'(v_x(x_0)). \end{aligned} \tag{5.18}$$

We now return to the ϵ-notation. From (5.5) we have

$$I(v_x^\epsilon(x_0)) \leq I(R_1) \tag{5.19}$$

and

$$|I'(v_x^\epsilon(x_0))| \leq |I'(R_2)|. \tag{5.20}$$

From (5.18), (5.19), and (5.20) we get that for some constants k_1, k_2, k_3, and k_4,

$$v_{xx}^\epsilon(x_0)^2[k_1\zeta(x_0)|\zeta_x(x_0)| + k_2\zeta(x_0)|\zeta_{xx}(x_0)| + C\zeta^2(x_0)] + k_4 \geq k_3\zeta(x_0)^2|v_{xx}^\epsilon(x_0)|^3.$$

Using property (iii) of ζ with $p = \frac{1}{3}$ we obtain

$$C_1 v_{xx}^\epsilon(x_0)^2[\zeta(x_0)^2 + \zeta(x_0)^{1+1/3}] + k \geq k_3\zeta(x_0)^2|v_{xx}^\epsilon(x_0)|^3 \tag{5.21}$$

for some $C_1 > 0$. Now if $w(x_0) = \zeta^2(x_0)|v_{xx}^\epsilon(x_0)|^3$, then $w(x_0)^{2/3} = \zeta(x_0)^{4/3}|v_{xx}^\epsilon(x_0)|^2$. In view of property (ii) of ζ, (5.21) yields

$$2C_1 w(x_0)^{2/3} + \overline{k} \geq k_3 w(x_0),$$

for some $\overline{k} > 0$ and, therefore,

$$w(x_0) \leq N$$

where $N = N(C_1, \overline{K}, K_3)$ is independent of ϵ.

Thus

$$(v_{xx}^\epsilon)^2 + \lambda(v_x^\epsilon)^2 - \mu v^\epsilon \leq N^{2/3} + \lambda(v_x^\epsilon(x_0))^2 - \mu v^\epsilon(x_0) \text{ on } [x_1, x_2],$$

i.e., there exists a constant L_1, independent of ϵ, such that

$$|v_{xx}^\epsilon| \leq L_1 \text{ on } [x_1, x_2].$$

Case 2b.

$$x_0 \epsilon A_2 = \left\{x \epsilon [x_1, x_2] : -\frac{b-r}{\sigma^2}\frac{v_x^\epsilon(x)}{v_{xx}^\epsilon(x)} \geq f(x)\right\}.$$

In this case,

$$|v^\epsilon_{xx}(x_0)| \leq L_2 \quad \text{on } [x_1, x_2] \quad \text{where } L_2 = \frac{b-r}{\sigma^2}\frac{R_2}{f(x_1)}.$$

Therefore,

$$|v^\epsilon_{xx}| \leq R \quad \text{on } [x_1, x_2]. \tag{5.22}$$

where $R = \max(L_1, L_2)$, independent of ϵ.

Combining (5.5) and (5.22) we see that

$$-\frac{b-r}{\sigma^2}\frac{v^\epsilon_x}{v^\epsilon_{xx}} \geq B > 0 \quad \text{on } [x_1, x_2] \quad \text{with } B = \frac{b-r}{\sigma^2}\frac{R_1}{R}.$$

Let us now consider the equation

$$\begin{aligned} \beta u = &\max_{B\leq\pi\leq f(x)}\left[\tfrac{1}{2}\sigma^2(\pi^2+\epsilon^2x^2)u_{xx} + (b-r)\pi u_x\right] \\ &+ \max_{c\geq 0}[-cu_x + U(c)] + rxu_x \\ u(x_1) = v^\epsilon(x_1), &\quad u(x_2) = v^\epsilon(x_2) \quad (x\epsilon[x_1, x_2]). \end{aligned} \tag{5.23}$$

In view of the above analysis, we know that v^ϵ solves (5.23). Let $\epsilon \to 0$. Since $v^\epsilon \to v$, locally uniformly, v is a viscosity solution of

$$\begin{aligned} \beta u = &\max_{B\leq\pi\leq f(x)}\left[\tfrac{1}{2}\sigma^2\pi^2u_{xx} + (b-r)\pi u_x\right] \\ &+ \max_{c\geq 0}[-cu_x + U(c)] + rxu_x (x\epsilon[x_1, x_2]) \\ u(x_1) = v(x_1), &\quad u(x_2) = v(x_2). \end{aligned} \tag{5.24}$$

On the other hand, (4.21) admits a unique smooth solution u (see [23]) which is the unique viscosity solution (see [16, Thm. II.2]); therefore v is smooth. □

THEOREM 5.2. *The feedback optimal controls C^* and π^* are given by*

$$c^*(x) = I(v_x(x)) \quad \textit{and} \quad \pi^*(x) = \min\left\{-\frac{b-r}{\sigma^2}\frac{v_x(x)}{v_{xx}(x)}, f(x)\right\} \quad \textit{for } x > 0.$$

The state equation (1.3) *has a strong unique solution X^*_t, corresponding to $C^*_t = c^*(X^*_t)$ and $\pi^*_t = \pi^*(X^*_t)$ and starting at $x > 0$ at $t = 0$, which is unique in probability law up to the first time τ such that $X^*_\tau = 0$.*

Proof. The formulae for π^* and C^* follow from a standard verification theorem (see [11]) and the equation. We now show that π^* and C^* are locally Lipschitz functions of x. It is clear that v_x is locally Lipschitz because in any compact set K there exists a constant $C = C(K)$ such that $|v_{xx}| \leq C(K), (x\epsilon K)$. Therefore C^* is locally Lipschitz. Moreover, from the Bellman equation we have that

$$v_{xx} = H(x, v, v_x)$$

where

$$H(x, v, v_x) = \frac{2[\beta v - (b-r)f(x)v_x + v_x I(v_x) - U(I(v_x)) + rxv_x]}{\sigma^2 f(x)^2}$$

$$\text{if} \quad -\frac{b-r}{\sigma^2}\frac{v_x}{v_{xx}} \geq f(x)$$

and

$$H(x,v,v_x) = -\frac{b-r}{2\sigma^2}\frac{v_x^2}{\beta v - rxv_x + v_x I(v_x) - U(I(v_x))} \quad \text{if} \quad -\frac{b-r}{\sigma^2}\frac{v_x}{v_{xx}} < f(x).$$

Since v_x is locally Lipschitz we get that v_{xx} is also locally Lipschitz. Therefore (see Gikhman and Skorohod [14]) equation (1.3) has a unique strong solution X_t^* in probability law up to the first time τ such that $X_\tau^* = 0$. □

6. In this section we discuss the finite horizon model and we state results about the value function and the optimal policies.

The investor starts at time $t \in [0,T)$ with an initial endowment x, consumes at rate C_s and invests π_s^0 (respectively, π_s) amount of money in bond (respectively, in stock) for $t \le s \le T$. The prices of the bond and the stock satisfy the same equations as in the infinite horizon case. The wealth of the investor $X_s = \pi_s^0 + \pi_s (t \le s \le T)$ satisfies the state equation

$$\begin{aligned} &dX_s = rX_s ds + (b-r)\pi_s ds - C_s ds + \sigma\pi_s dW_s \qquad (t \le s \le T) \\ &X_t = x \qquad (x > 0). \end{aligned} \tag{6.1}$$

The agent faces the same constraints as in the infinite horizon case. In other words, the wealth must stay nonnegative; the agent cannot consume at a negative rate and must meet borrowing constraints ($\pi_s \le f(X_s)$ for $t \le s \le T$).

The *total utility* coming both from consumption and terminal wealth is

$$J(x,t,C,\pi) = E\left[\int_t^T U(C_s)ds + \Phi(X_T)\right]$$

where U is the usual utility function and Φ is the *bequest* function which is typically concave, increasing, and smooth.

The *value function* is

$$\mathcal{A}(x,t) = \sup_{\mathcal{A}_{(x,t)}} J(x,t,C,\pi)$$

where $\mathcal{A}_{(x,t)}$ is the set of admissible controls.

In the sequel we state the main theorems. Since the proofs are modifications of the ones given in the previous sections they are not presented here.

THEOREM 6.1. *The value function is the unique continuous on* $\Omega \times [t,T]$ *and* $C^{2,1}(\Omega \times (t,T))$ *solution of*

$$\begin{aligned} &v_t + \max_{\pi \le f(x)}\left[\tfrac{1}{2}\sigma^2\pi^2 v_{xx} + (b-r)\pi v_x\right] + \max_{c \ge 0}[-cv_x + U(c)] + rxv_x = 0 \\ &v(x,T) = \Phi((x) \end{aligned} \tag{6.2}$$

in the class of concave (with respect to the space variable x) functions.

THEOREM 6.2. *The feedback optimal controls* C^* *and* π^* *are given by* $C_t^* = c^*(X_t^*,t)$ *and* $\pi_t^* = \pi^*(X_t^*,t)$ *where* $C^*(x,t) = (U')^{-1}(v_x(x,t))$ *and*

$$\pi^*(x,t) = \min\left\{-\frac{b-r}{\sigma^2}\frac{v_x(x,t)}{v_{xx}(x,t)}, f(x)\right\} \quad \textit{for } x > 0.$$

REFERENCES

[1] R. S. Bellman, *Dynamic Programming,* Princeton University Press, Princeton, NJ, 1957.

[2] G. Barles and B. Perthame, *Exit time problems in optimal control and the vanishing viscosity method,* SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[3] I. Capuzzo–Dolcetta and P.-L. Lions, *Hamilton–Jacobi equations and state-constraints problems,* preprint.

[4] J. Cox and C. Huang, *Optimal consumption and portfolio policies when asset prices follow a diffusion process,* J. Economic Theory, 49 (1989), pp. 33–83.

[5] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations,* preprint.

[6] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton–Jacobi equations,* Trans. AMS, 277 (1983), pp. 1–42.

[7] J. Cvitanic and I. Karatzas, *Convex duality in constrained portfolio optimization,* Department of Statistics, Columbia University, New York, preprint.

[8] D. Duffie, W. Fleming, and T. Zariphopoulou, *Hedging in incomplete markets with HARA utility,* Business School, Stanford University, Stanford, CA, preprint.

[9] L. C. Evans and P.E. Souganidis, *Differential games and representation for solutions of Hamilton–Jacobi–Isaacs equations,* Indiana U. Math. J., 33 (1984), pp. 773–797.

[10] B. G. Fitzpatrick and W.H. Fleming, *Numerical methods for an optimal investment–consumption model,* Math. Oper. Res., 16 (1991), pp. 823–841.

[11] W. H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control,* Springer-Verlag, Berlin, New York, 1975.

[12] W. H. Fleming and P.E. Souganidis, *On the existence of value functions of two player, zero-sum stochastic differential games,* Indiana U. Math. J., 38 (1989), pp. 293–314.

[13] W. H. Fleming and T. Zariphopoulou, *An optimal investment-consumption model with borrowing,* Math. Oper. Res., 16 (1991), pp. 802–822.

[14] I. I. Gikhman and A.V. Skorohod, *Stochastic Differential Equations,* Springer-Verlag, Berlin, New York, 1972.

[15] H. He and N.D. Pearson, *Consumption and portfolio policies with incomplete markets and short-sale constraints: The infinite dimensional case,* J. Econom. Theory, 54 (1991), 259–304.

[16] H. Ishii and P.-L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations,* J. Differential Equations, 83 (1990), pp. 26–78.

[17] R. Jensen, P.-L. Lions, and P.E. Souganidis, *A uniqueness result for viscosity solutions of second order fully nonlinear partial differential equations,* Proc. AMS, 102 (1988), pp. 975–978.

[18] I. Karatzas, J. Lehoczky, S. Sethi, and S. Shreve, *Explicit solution of a general consumption-investment problem,* Math. Oper. Res., 11 (1986), pp. 261–294.

[19] I. Karatzas, J. Lehoczky, and S. Shreve, *Optimal portfolio and consumption decisions for a small investor on a finite horizon,* SIAM J. Control Optim., 25 (1987), pp. 1557–1586.

[20] I. Karatzas, J. Lehoczky, S. Shreve, and G.-L. Xu, *Martingale and duality methods for utility maximization in an incomplete market,* SIAM J. Control Optim., 29 (1991), pp. 702–730.

[21] M. Katsoulakis, *State-constraints problems for second order fully nonlinear degenerate partial differential equations,* preprint.

[22] N. V. Krylov, *Controlled Diffusion Processes,* Springer-Verlag, Berlin, New York, 1980.

[23] ———, *Nonlinear Elliptic and Parabolic Equations of the Second Order,* D. Reidel, Dordrecht, Holland, 1987.

[24] J. M. Lasry and P.-L. Lions, *A remark on regularization in Hilbert spaces,* Israel J. Math., 55 (1986), pp. 257–266.

[25] J. Lehoczky, S. Sethi, and S. Shreve, *Optimal consumption and investment policies allowing consumption constraints and bankruptcy,* Math. Oper. Res., 8 (1983), pp. 613–636.

[26] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations, Part 1,* Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174.

[27] ———, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations, Part 2,* Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.

[28] R. C. Merton, *Lifetime portfolio selection under uncertainty: The continuous-time case,* Rev. Econom. Statist., 51 (1969), pp. 247–257.

[29] ———, *Optimum consumption and portfolio rules in a continuous-time model,* J. Econom. Theory 3 (1971), 373–413.

[30] ———, *Erratum,* J. Econom. Theory, 6 (1973), pp. 213–214.

[31] H. Pages, *Optimal consumption and portfolio policies when markets are incomplete,* MIT mimeo, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[32] S. A. Pliska, *Stochastic calculus model of continuous trading: Optimal portfolios,* Math. Oper. Res., 11 (1986), pp. 371–382.

[33] S. Sethi and M. Taksar, *Infinite horizon investment consumption model with a nonterminal bankruptcy,* FSU Stat. Report No. M-750,

[34] ———, *A Note on Merton's optimum consumption and portfolio rules in a continuous-time model,* FSU Statistics Report No. M746,

[35] S. Shreve and G.-L. Xu, *A duality method for optimal consumption and investment under short-selling prohibition, Part 1: General market coefficients,* Ann. Appl. Probability, 1 (1991), pp. 87–112.

[36] ———, *A duality method for optimal consumption and investment under short-selling prohibition, Part 2: Constant market coefficients,* Ann. Appl. Probability 1 (1991), pp. 52–69.

[37] H. M Soner, *Optimal Control with state-space constraints,* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[38] J.-L. Vila and T. Zariphopoulou, *Optimal consumption and portfolio choice with borrowing constraints,* J. Econom. Theory, submitted.

[39] G. Xu, *A duality approach to stochastic portfolio/consumption decision problem in a continuous time market with short-selling restriction,* Ph. thesis, Carnegie Mellon University, Pittsburgh, PA, 1988.

[40] T. Zariphopoulou, *Optimal investment-consumption models with constraints,* Ph. thesis, Brown University, Providence, RI, 1988.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 86–115, January 1994

006

A SUPER-OPTIMIZATION METHOD FOR FOUR-BLOCK PROBLEMS*

P.-O. NYMAN†

Abstract. This paper deals with a discrete time frequency domain approach to super-optimization of four-block problems via a dimension reduction and diagonalization technique based on the so-called equalizer principle, an optimality criterion associated with the polynomial approach to $\mathcal{H}^\infty$-optimization by Kwakernaak. The reduction technique provides a hierarchical decomposition of the super-optimization problem into successive ordinary $\mathcal{H}^\infty$-optimization problems, one for each singular value to be minimized. The procedure allows computation of super-optimal solutions for a large class of four-block problems.

Key words. super-optimization, $\mathcal{H}^\infty$-optimal control, equalizer principle

AMS subject classifications. 93B36, 93B50, 93B35

1. Introduction. In ordinary $\mathcal{H}^\infty$-optimization of a multivariable control system, the compensator is chosen to minimize the largest singular value of the transfer function. Such a controller is, in general, far from unique, and a substantial amount of freedom is left which could be used to optimize some additional performance criterion. Among many possible choices a most natural and appealing one is to pursue a compensator minimizing not only the largest singular value, but also the second largest, the third largest, etc., with respect to lexicographic ordering. This strengthened $\mathcal{H}^\infty$-optimality criterion was proposed by Young in [27] where, in the context of an one-block problem, an algorithm was given and the existence and uniqueness of the optimal solution was proved. Strengthened $\mathcal{H}^\infty$-optimization has also received attention in the control theoretic community (see, e.g., [3], [8], [9], [11], [15], [16], [18], [21], [24], [26]) and has been called *super optimization*. Similar to ordinary $\mathcal{H}^\infty$-optimization, super-optimization is motivated by robustness requirements. In fact, simple examples suggest that a super-optimal compensator may improve the robustness properties achieved by an arbitrary $\mathcal{H}^\infty$-optimal compensator. Algorithms appearing in the literature mostly restrict themselves to one- and two-block problems. The two-block method in [9], based on a state-space descriptor representation of all $\mathcal{H}^\infty$-optimal solutions is, however, reported to have an extension to the four-block problem.

In this paper we propose a frequency domain approach to super-optimization of the four-block problem. The setting will be discrete time, but with appropriate modifications the method may be used for continuous time as well. As in [27], a dimension reducing and diagonalizing technique is used to decompose the problem into $\mathcal{H}^\infty$-minimizations of individual singular values in a hierarchical fashion. The diagonalization transformations are derived via the so-called equalizer principle, an optimality criterion associated with the polynomial $\mathcal{H}^\infty$-optimization approach by Kwakernaak (cf., e.g., [13]). Previous use of the equalizer principle in the context of super-optimization may be found in [11], [15], and [18]. By using a polynomial matrix fraction representation of data, the computational steps may be adapted for numerical calculations based on polynomial routines of the type considered in [14]. This certainly holds for the minimizations of individual singular values, where the polynomial $\mathcal{H}^\infty$-optimization approach by Kwakernaak is available.

The paper is organized as follows. Section 2 collects a few preliminaries. Section 3 gives some useful characterizations of the conditions of equalizer principle. In §4 these are used to establish the dimension reduction and diagonalization technique. The spectral density associated with the equalizer principle plays an essential role here. In §5 the

* Received by the editors August 5, 1991; accepted for publication (in revised form) August 11, 1992.
† University of Twente, Department of Applied Mathematics, P.O. Box 217, 7500 AE Enschede, the Netherlands.

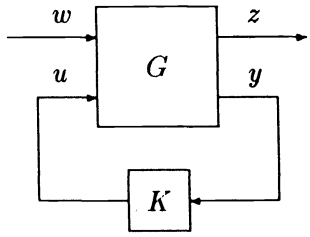


FIG. 1. *The standard configuration.*

reduction technique and the polynomial $\mathcal{H}^\infty$-optimizations approach are conjoined into a super-optimization procedure. Section 6, which logically may precede §4, gives a method for computing the spectral density needed for the dimension reduction. In §7 conditions for the existence of the spectral density are given in terms of a four-block operator. Finally, in §8 the super-optimization procedure is illustrated with a numerical four-block example.

2. Preliminaries. We first recall some well-known facts about the rich class of interesting control theoretic $\mathcal{H}^\infty$-optimization problems described by the so-called standard problem formulation, illustrated by the block diagram in Fig. 1. The block G is a generalized plant which, in addition to the actual plant, also includes weighting functions. The block K is a feedback compensator. Both G and K are supposed to be representable by rational transfer matrices, also denoted G and K. The signal w contains all external inputs to the system. The output z to be controlled may be thought of as a control error. The signals u and y are the control input and the measured output, respectively. Let

$$G = \begin{matrix} \\ m \\ q \end{matrix} \overset{\begin{matrix} n & p \end{matrix}}{\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}}$$

be a partition of the plant done conformly with its input $(w^{\mathrm{T}}, u^{\mathrm{T}})^{\mathrm{T}}$ and output $(z^{\mathrm{T}}, y^{\mathrm{T}})^{\mathrm{T}}$. The transfer matrix from the input w to the output z may then be written

$$H_K = G_{11} + G_{12}K(I - G_{22}K)^{-1}G_{21}. \tag{1}$$

The standard problem of $\mathcal{H}^\infty$ control consists of finding a compensator K stabilizing the plant G, and making the $\mathcal{H}^\infty$ norm of H_K as small as possible. For a precise definition of what is meant by a stabilizing compensator we refer to [6]. Assuming G to be stabilizable (i.e., that G admits a stabilizing compensator), we have at our disposal the so-called Youla parametrization, which states that a compensator K stabilizes G if and only if

$$K = -(U_l + D_r Q)(V_l - N_r Q)^{-1} = -(V_r - QN_l)^{-1}(U_r + QD_l), \tag{2}$$

where the rational matrices $U_l, V_l, N_l, D_l U_r, V_r, N_r, D_r$ are obtained from a doubly coprime factorization

$$G_{22} = N_r D_r^{-1} = D_l^{-1} N_l, \qquad \begin{bmatrix} V_r & U_r \\ -N_l & D_l \end{bmatrix} \begin{bmatrix} D_r & -U_l \\ N_r & V_l \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix}. \tag{3}$$

For a proof see [6] and [25]. Using (2) and (3) we may write $K(I - G_{22}K)^{-1} = -(U_l + D_r Q)D_l$ which, substituted into (1), gives

$$H_K = F + BQC,$$

where $F = G_{11} - G_{12}U_l D_l G_{21} \in \mathcal{RH}^\infty_{m\times n}$, $B = -G_{12}D_r \in \mathcal{RH}^\infty_{m\times p}$, and $C = D_l G_{21} \in \mathcal{RH}^\infty_{q\times n}$. The standard problem therefore is equivalent to the model-matching problem of finding a $Q \in \mathcal{RH}^\infty_{p\times q}$ such that $\|F + BQC\|_\infty$ is minimized. Conversely, any model-matching problem is easily expressed as a standard problem. As the basic structure for $\mathcal{H}^\infty$-optimization problems we therefore take, instead of the standard configuration, the model-matching configuration; that is, a coset in $\mathcal{H}^\infty_{m\times n}$ defined by

$$\Sigma = F + B\mathcal{H}^\infty_{p\times q}C = \{F + BQC | Q \in \mathcal{H}^\infty_{p\times q}\}, \tag{4}$$

where $F \in \mathcal{H}^\infty_{m\times n}$, $B \in \mathcal{H}^\infty_{m\times p}$, $C \in \mathcal{H}^\infty_{q\times n}$ are rational matrices, with B tall and of full column rank, and C wide and of full row rank. To guarantee that $\mathcal{H}^\infty$-optimal solutions exist we assume that $B(e^{i\theta})$ and $C(e^{i\theta})$ have full rank for each $e^{i\theta}$ on the unit circle (cf. [6]).

$\mathcal{H}^\infty$-optimization over the coset Σ is also known as the *four-block problem.* More precisely, it may be shown to be equivalent to $\mathcal{H}^\infty$-optimization over

$$\begin{bmatrix} A_{11} + M\mathcal{H}^\infty_{p\times q} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{5}$$

where M is a rational square inner matrix which, together with the four rational matrix blocks $A_{ij}, i, j \in \{1, 2\}$, is obtained from the data F, B, and C. If one of B and C is square we call Σ a *two-block problem.* This corresponds in (5) to disappearance of A_{22} together with one of the blocks A_{12}, A_{21}. If both B and C are square Σ is called a *one-block problem.* This corresponds to disappearance of all three blocks A_{22}, A_{12}, and A_{21}.

When desired, we may, via an inner-outer factorization of B and a co-inner/co-outer factorization of C, together with subsequent absorption of the outer and co-outer factors in the parameter $Q \in \mathcal{H}^\infty_{p\times q}$, take B inner and C co-inner; that is, $B^*B = I$ and $CC^* = I$ on the unit circle.

2.1. The polynomial $\mathcal{H}^\infty$-optimization approach. Polynomial methods have come to play a significant role in frequency domain design of optimal control. One reason for this is their capability of providing numerically reliable techniques. A manifestation of this is the so-called *polynomial approach* to $\mathcal{H}^\infty$-optimization by Kwakernaak [13]. The method allows us to compute not only suboptimal, but also strictly optimal solutions to the standard problem. Of central interest in the polynomial approach is a parametrization of all stabilizing compensators K such that $\|H_K\|_\infty \leq \lambda$, where λ is greater or equal to the minimal $\mathcal{H}^\infty$ norm λ_o. For $\lambda > \lambda_o$ the parametrization is obtained via two polynomial J-spectral factorizations, one for the denominator and one for the numerator of a certain para-Hermitian rational matrix Π_λ that depends only on the plant G and the performance level λ. This results in a rational J-spectral factor Z_λ of Π_λ such that K stabilizes G and achieves $\|H_K\|_\infty \leq \lambda$ if and only if for some stable U with $\|U\|_\infty \leq 1$ we have

$$K = YX^{-1}, \qquad \begin{bmatrix} X \\ Y \end{bmatrix} = Z_\lambda^{-1} \begin{bmatrix} I \\ U \end{bmatrix}. \tag{6}$$

By decreasing λ, compensators with performance arbitrarily close to λ_o may be computed. Moreover, the numerical value of λ_o may be delimited to desired accuracy.

As λ reaches the optimal value λ_o, the numerator and denominator of the compensator acquire a common factor that may be canceled. This results in a so-called *reduced degree solution* K_o that is $\mathcal{H}^\infty$-optimal, that is, $\|H_{K_o}\|_\infty = \lambda_o$. A second phenomenon that takes place as λ reaches λ_o is that some of the coefficients of Z_λ approach infinity. Due to this fact, (6) is not suitable for parametrization of strictly optimal solutions. However, if the

numerator of Π_{λ_o} is only partially J-spectral factorized (i.e., with the signature matrix J replaced by a certain para-Hermitian polynomial matrix), then a partial rational J-spectral factor $\hat{Z}_{\lambda_o}$ of Π_{λ_o} is obtained which gives rise to a parametrization of all strictly optimal compensators. In the simplest case, where the problem has a solution with largest singular value of multiplicity one, the parametrization may be expressed in the following way.

A compensator K stabilizes G if and only if for some stable rational matrix U with $\|U\|_\infty \leq 1$ we have

$$K = YX^{-1}, \qquad \begin{bmatrix} X \\ Y \end{bmatrix} = \hat{Z}_\lambda^{-1} \begin{bmatrix} I & 0 \\ 0 & \alpha \\ 0 & \beta \\ U & 0 \end{bmatrix}, \tag{7}$$

where α and β are some fixed constants.

Comparing (7) and (6) we see that when passing from suboptimality to optimality there is a reduction in the dimension of the free parameter U. We also see that the optimal compensators contain a common part that is given by the column $[0 \quad \alpha \quad \beta \quad 0]^{\mathrm{T}}$ of (7). In the scalar case only, this part is present and determines the unique optimal compensator.

We point out that the technique for computing $\hat{Z}_\lambda$ is numerically feasible and leads to compensators having the correct reduced degree structure. For a detailed description of the algorithm we refer to [13]. For a MATLAB implementation the reader may consult [14].

A reduced degree solution admits a spectral density Φ such that the following sufficient condition for optimality holds.

Equalizer principle. Suppose that the compensator K_o is *equalizing*, that is,

$$H_{K_o}^* H_{K_o} = \lambda_o^2 I - L_o^* L_o \tag{8}$$

for some nonnegative constant λ_o and some rational rank deficient matrix L_o, and that K_o and L_o minimize

$$\frac{1}{2\pi} \int_0^{2\pi} tr\{[H_K^*(e^{i\theta}) H_K(e^{i\theta}) + L^*(e^{i\theta}) L(e^{i\theta})] \Phi(e^{i\theta})\} d\theta \tag{9}$$

with respect to all stabilizing compensators K and all rank deficient matrices L. Then K_o minimizes $\|H_K\|_\infty$ with respect to all stabilizing compensators, and $\|H_{K_o}\|_\infty = \lambda_o$.

The equalizer principle is closely associated with the polynomial approach. For an account of the use of this criterion in proving optimality of reduced solutions we refer to [10], [11], [1], [12], and [13].

Besides its usefulness as an optimality criterion for ordinary $\mathcal{H}^\infty$-optimization, the equalizer principle provides a bridge to super-optimization. Via the spectral density of the equalizer principle, the $\mathcal{H}^\infty$-optimal solutions, that is, those optimal with respect to the largest singular value, may be diagonalized, and a similar but smaller super-optimization problem for the remaining singular values may be derived. The super-optimization algorithm considered in this paper is based on this fact. It is much inspired by that given in [27], but is applicable to the four-block problem. For the one-block problem a similar method was studied in [18]. The equalizer principle also plays a central role in [15] and [11], where super-optimization of a mixed sensitivity problem is studied.

2.2. Notation. The notation we use is fairly standard. By the jth *singular value* of a complex matrix A, denoted $s_j(A)$, we mean the jth largest eigenvalue of $(A^*A)^{1/2}$, where A^* is the conjugate transpose of A. A corresponding eigenvector of $(AA^*)^{1/2}$ is called a (*right*) *singular vector* of A associated with $s_j(A)$. For a matrix valued function F on the

unit circle $\mathbb{T}$ we mean by F^* the function on $\mathbb{T}$ defined by $F^*(e^{i\theta}) = F(e^{i\theta})^*$. Consider on $\mathbb{C}^{m\times n}$ the Schatten p-norms $|\cdot|_p$, that is,

$$|A|_p = \begin{cases} [\sum_{j=1}^n s_j(A)^p]^{1/p}, & 1 \le p < \infty, \\ |A|_\infty = s_1(A), & p = \infty. \end{cases}$$

Note that $p = \infty, p = 2$, and $p = 1$ yield the spectral norm, the Frobenius norm, and trace norm, respectively. By $\mathcal{L}^p_{m\times n}$ we mean the space of (equivalence classes of) $\mathbb{C}^{m\times n}$ valued Lebesgue measurable functions F on $\mathbb{T}$ for which $\|F\|_p < \infty$, where the norm $\|\cdot\|_p$ is defined by

$$\|F\|_p = \begin{cases} [\frac{1}{2\pi}\int_0^{2\pi} |F(e^{i\theta})|_p^p d\theta]^{1/p}, & 1 \le p < \infty, \\ \|F\|_\infty = \operatorname{ess\,sup}\{|F(e^{i\theta})|_\infty : \theta \in [0, 2\pi)\}, & p = \infty. \end{cases}$$

The inner product in the Hilbert space $\mathcal{L}^2_{m\times n}$ is given by

$$(F, G) = \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}\{G(e^{i\theta})^* F(e^{i\theta})\} d\theta.$$

By the Hardy space $\mathcal{H}^p_{m\times n}$, $1 \le p \le \infty$ we mean the closed subspace of $\mathcal{L}^p_{m\times n}$ of functions F whose Poisson integral $\tilde{F}$ is analytic in the open unit disk $\mathbb{D}$, or equivalently, whose Fourier coefficients $(2\pi)^{-1}\int_0^{2\pi} F(e^{i\theta})e^{-ik\theta}d\theta$ vanish for $k < 0$. As in the scalar valued case, the correspondence between F and $\tilde{F}$ is an isomorphism. When so desired we identify F and $\tilde{F}$. By $\mathcal{RL}^p_{m\times n}$ and $\mathcal{RH}^p_{m\times n}$ we mean the subspaces of rational functions in $\mathcal{L}^p_{m\times n}$ and $\mathcal{H}^p_{m\times n}$, respectively. Note that regarded as sets is $\mathcal{RL}^p_{m\times n}$ equal to $\mathcal{RL}^q_{m\times n}$ for all $1 \le p, q \le \infty$. The same holds for $\mathcal{RH}^p_{m\times n}$.

A function $F \in \mathcal{H}^2_{m\times n}$ is *inner* if $F^*F = I$ almost everywhere on $\mathbb{T}$. It is *outer* if there exists a scalar valued outer function $g \in \mathcal{H}^\infty$ such that $gF \in \mathcal{H}^\infty_{m\times n}$ and $gF\mathcal{H}^2_n$ is dense in $\mathcal{H}^2_m$ (cf. [22]). F is said to be *co-inner* (*co-outer*) if the transpose of F is inner (outer). For any $F \in \mathcal{H}^2_{m\times n}$ there exists an integer $k \le m, n$, an inner function $F_i \in \mathcal{H}^\infty_{m\times k}$, and an outer function $F_o \in \mathcal{H}^2_{k\times n}$ such that F has the *inner-outer factorization* $F = F_iF_o$ (cf. [5], [22]).

Define for any $F \in \mathcal{L}^\infty_{m\times n}$ a bounded linear functional $\langle\cdot, F\rangle$ on $\mathcal{L}^1_{n\times m}$ by

$$\langle G, F\rangle = \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}\{F(e^{i\theta})G(e^{i\theta})\} d\theta = \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}\{G(e^{i\theta})F(e^{i\theta})\} d\theta. \tag{10}$$

The mapping $F \to \langle\cdot, F\rangle$ is an isometric isomorphism from $\mathcal{L}^\infty_{m\times n}$ onto $(\mathcal{L}^1_{n\times m})^*$ (cf. [23], [27]). As a convenient abbreviation for the right-hand side of (10) we take the liberty of using the notation $\langle G, F\rangle$ even when the arguments do not belong to the aforementioned spaces, the only requirement being that $\operatorname{tr}\{F(e^{i\theta})G(e^{i\theta})\}$ is integrable.

A function $\Phi \in \mathcal{L}^1_{n\times n}$ is called a *spectral density* if $\Phi \ne 0$ and $\Phi(e^{i\theta}) \ge 0$ for every $\theta \in [0, 2\pi)$. Note that $\|\Phi\|_1 = (2\pi)^{-1}\int_0^{2\pi} \operatorname{tr}\{\Phi(e^{i\theta})\} d\theta$.

Similarly to $\|F\|_\infty$, we introduce bounds for each singular value $s_j\{F(e^{i\theta})\}, 1 \le j \le n$, by defining

$$s_j^\infty(F) = \operatorname{ess\,sup}\{s_j\{F(e^{i\theta})\} : \theta \in [0, 2\pi)\}.$$

Let $\le_k$ be the partial ordering on $\mathbb{R}^n$ which is lexicographic with respect to the first k positions and ignores the remaining ones. More precisely,

$$(a_1, \ldots, a_k, a_{k+1}, \ldots, a_n) \le_k (b_1, \ldots, b_k, b_{k+1}, \ldots, b_n)$$

if and only either $a_i = b_i$ for all $i = 1, 2, \ldots, k$, or else $a_i < b_i$ at the smallest $1 \le i \le k$ where $a_i \ne b_i$. Let $\mathcal{F}$ be any subset of $\mathcal{L}^\infty_{m\times n}$. We then say that $F_o \in \mathcal{F}$ is *class k optimal* if

$$(s_1^\infty(F_o), s_2^\infty(F_o), \ldots, s_n^\infty(F_o)) \le_k (s_1^\infty(F), s_2^\infty(F), \ldots, s_n^\infty(F))$$

for all $F \in \mathcal{F}$. When $k = 1$ we also say that F_o is $\mathcal{L}^\infty$*-optimal*, or $\mathcal{H}^\infty$*-optimal* in case $\mathcal{F} \subseteq \mathcal{H}^\infty_{n\times n}$. When $k = n$ we call F_o *super-optimal.*

3. Conditions for $\mathcal{H}^\infty$-optimality. In this section, we collect some basic results concerning the spectral density associated with the equalizer principle and the $\mathcal{H}^\infty$-optimal solutions satisfying this criterion. By taking $\mathcal{Z}$ equal to the set of all functions $Z_{K,L} = H_K^* H_K + L^* L$ appearing in the integrand of (9), the equalizer principle may be regarded as a case of the following lemma (see [18]).

LEMMA 1. *Suppose we are given a nonempty subset $\mathcal{Z}$ of $\mathcal{L}^\infty_{m\times n}$ and a $\Phi \in \mathcal{L}^1_{n\times m}$ with $\Phi \ne 0$. Let $\mathcal{Z}_o$ be the set of all $Z_o \in \mathcal{Z}$ such that*

$$|\langle \Phi, Z_o\rangle| \le |\langle \Phi, Z\rangle| \quad \textit{for all } Z \in \mathcal{Z}, \tag{11}$$

$$|\langle \Phi, Z_o\rangle| = \|Z_o\|_\infty \|\Phi\|_1. \tag{12}$$

Then $\mathcal{Z}_o$ either is empty or equals the set of $\mathcal{L}^\infty$-optimal functions in $\mathcal{Z}$.

Proof. Owing to isometry we have $\|Z\|_\infty = \|\langle \cdot, Z\rangle\|$. Thus $\|Z_o\|_\infty \|\Phi\|_1 = |\langle \Phi, Z_o\rangle| \le |\langle \Phi, Z\rangle| \le \|Z\|_\infty \|\Phi\|_1$. Since $\|\Phi\|_1 > 0$, the conclusions follow. □

If the set $\mathcal{Z}_o$ is nonempty, that is, if Φ allows a Z_o such that (11) and (12) are satisfied, then Φ is said to *distinguish* the $\mathcal{L}^\infty$-optimal functions in $\mathcal{Z}$, or to be *distinguishing* for $\mathcal{Z}$. The reason behind this terminology is that either Φ allows no Z_o to satisfy conditions (11) and (12), or else they are satisfied precisely by the $\mathcal{L}^\infty$-optimal functions in $\mathcal{Z}$. We remark, however, that a set $\mathcal{Z}$ may very well have $\mathcal{L}^\infty$-optimal elements without admitting any distinguishing Φ.

Note that if all $Z \in \mathcal{Z}$, together with Φ, take only nonnegative semidefinite values on $\mathbb{T}$, then the absolute value signs in (11) and (12) may be omitted. This certainly holds in the setting of the equalizer principle, and more generally if, for some subset Σ of $\mathcal{L}^\infty_{m\times n}$, we have $\mathcal{Z} = \mathcal{S}(\Sigma)$, where $\mathcal{S}(\Sigma)$ denotes the set $\{H^* H : H \in \Sigma\}$. All our applications of Lemma 1 will be special cases of this, where we in fact will take $\Sigma = F + \mathcal{U}$ for some given function $F \in \mathcal{H}^\infty_{m\times n}$ and some subspace $\mathcal{U}$ of $\mathcal{H}^\infty_{m\times n}$, that is, Σ is a coset of $\mathcal{U}$ in $\mathcal{H}^\infty_{m\times n}$. The following lemma applies to this situation and characterizes the second condition of Lemma 1.

LEMMA 2. *Suppose that $Z \in \mathcal{L}^\infty_{m\times n}$ with $Z(e^{i\theta}) \ge 0$ for each $\theta \in [0, 2\pi)$. Let $\Phi \in \mathcal{L}^1_{n\times m}$ be a spectral density. Then*

$$\langle \Phi, Z\rangle = \|Z\|_\infty \|\Phi\|_1 \tag{13}$$

if and only if

$$Z(e^{i\theta})\Phi(e^{i\theta}) = \|Z\|_\infty \Phi(e^{i\theta}) \quad \textit{for almost all } \theta \in [0, 2\pi). \tag{14}$$

Note that condition (14) means that for almost all $\theta \in [0, 2\pi)$, each nonzero vector in the range of $\Phi(e^{i\theta})$ is an eigenvector of $Z(e^{i\theta})$ with the corresponding eigenvalue equal to $s_1^\infty(Z) = \|Z\|_\infty$. A similar result holds even if Z and Φ are not required to have positive

semidefinite values. For a proof of this and of Lemma 2 see [18]. Note also that if Z and Φ are continuous, then (14) must hold, not only almost everywhere, but for all $\theta \in [0, 2\pi)$.

For a given spectral density $\Phi \in \mathcal{L}^{\infty}_{n\times n}$, we denote by $\mathcal{A}(\Sigma; \Phi)$ the set of $H \in \Sigma$ such that

$$\langle \Phi, H^*H\rangle \leq \langle \Phi, G^*G\rangle \quad \text{for all} \quad G \in \Sigma;$$

that is, such that the first condition of Lemma 1 holds for H^*H. The following lemma gives a useful characterization of this set.

LEMMA 3. *Let $H_o \in F + \mathcal{U}$. Suppose $\Phi \in \mathcal{L}^1_{n\times n}$ is a spectral density. Then:*

(i) *The function H_o belongs to $\mathcal{A}(F+\mathcal{U}; \Phi)$ if and only if*

$$\langle \Phi H_o^*, U\rangle = 0 \quad \textit{for all} \quad U \in \mathcal{U}. \tag{15}$$

(ii) *Suppose that $H_o \in \mathcal{A}(F+\mathcal{U}; \Phi)$ and that Φ has a factorization $\Phi = \phi\phi^*$, where ϕ belongs to some $\mathcal{L}^2_{n\times k}$. Then $H \in \mathcal{A}(F+\mathcal{U}; \Phi)$, if and only if $H\phi = H_o\phi$.*

Proof. (i) Suppose that $H_o = F + U_o \in \mathcal{A}(F+\mathcal{U}; \Phi)$. Let $Z_o = H_o^*H_o$ and consider an arbitrary $Z = (F+U)^*(F+U) \in \mathcal{S}(F+\mathcal{U})$. With $\Delta = U - U_o$ we then have $Z = Z_o + H_o^*\Delta + \Delta^*H_o + \Delta^*\Delta$. Hence

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(Z\Phi)d\theta = \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(Z_o\Phi)d\theta + 2\operatorname{Re}\left\{\frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(H_o^*\Delta\Phi)d\theta\right\} \\ + \frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(\Delta^*\Delta\Phi)d\theta.\end{aligned} \tag{16}$$

In the middle term the expression within braces is precisely $\langle \Phi H_o^*, \Delta\rangle$, which we wish to prove equal to zero for all $\Delta \in \mathcal{U}$. Suppose $(2\pi)^{-1}\int_0^{2\pi} \operatorname{tr}(H_o^*\Delta\Phi)d\theta$ would differ from zero for some Δ. Replacing Δ by $\epsilon\,\Delta$, where ϵ is a suitable complex constant of sufficiently small modulus, we may then assume that

$$\frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(\Delta^*\Delta\Phi)d\theta < -2\operatorname{Re}\left\{\frac{1}{2\pi}\int_0^{2\pi} \operatorname{tr}(H_0^*\Delta\Phi)d\theta\right\}.$$

But this would mean that $(2\pi)^{-1}\int_0^{2\pi} \operatorname{tr}(Z\Phi)d\theta < (2\pi)^{-1}\int_0^{2\pi} \operatorname{tr}(Z_o\Phi)d\theta$, which contradicts our assumptions about H_o. Thus (15) must hold. The converse is immediate from (15) and (16).

(ii) Suppose that $H_o = F + U_o$ and $H = F + U$ are in $\mathcal{A}(F+\mathcal{U}; \Phi)$. From (15) we then have $\langle \Phi(U_o - U)^*, W\rangle = 0$ for each $W \in \mathcal{U}$. Taking $W = U_o - U$ and using the factorization of Φ we get

$$\langle \phi^*(U_o - U)^*, (U_o - U)\phi\rangle = ((U_o - U)\phi, (U_o - U)\phi) = 0.$$

Hence $(U_o - U)\phi = 0$ and thus also $H_o\phi = H\phi$. The converse is obvious. □

3.1. Rational data. The preceding two lemmas are concerned with a structure that is somewhat more general than what we need. Consider therefore a coset of the form (4). Thus $\Sigma = F + B\mathcal{H}^{\infty}_{p\times q}C$, where the "data" $F \in \mathcal{H}^{\infty}_{m\times n}$, $B \in \mathcal{H}^{\infty}_{m\times p}$, and $C \in \mathcal{H}^{\infty}_{q\times n}$ are rational matrices, with $B(e^{i\theta})$ having full column rank and $C(e^{i\theta})$ having full row rank for each $e^{i\theta}$ on the unit circle. For such a Σ it is plausible that there are, among the $\mathcal{H}^{\infty}$-optimal functions, also rational ones. Moreover, it is then reasonable to believe that a distinguishing spectral density for $\mathcal{S}(\Sigma)$, if it exists, also may be taken rational.

We will assume something slightly weaker, however, namely, that the spectral density has a particular kind of factorization. For distinguishing spectral densities, this leads to the existence of rational densities in §6, something of indispensable value in practical computations. To arrive at the required factorization we first study the rational case.

Suppose B and C have been taken inner and co-inner, respectively. Let $\tilde{B} = [B \quad \hat{B}]$ be a completion of B to a square inner matrix in $\mathcal{H}^\infty_{m\times m}$, that is, $\hat{B} \in \mathcal{RH}^\infty_{m\times(m-p)}$ is an inner complement of B. Similarly, let C be completed to a square inner matrix

$$\tilde{C} = \begin{bmatrix} C \\ \tilde{C} \end{bmatrix} \in \mathcal{RH}^\infty_{n\times n},$$

where $\tilde{C} \in \mathcal{RH}^\infty_{(n-q)\times n}$ is a co-inner complement of C. Suppose $\mathcal{S}(\Sigma)$ has a rational distinguishing spectral density Φ of rank r. Using spectral factorization we may write $\tilde{C}\Phi\tilde{C}^* = \vartheta\vartheta^*$ for some $\vartheta \in \mathcal{RH}^2_{n\times r}$ of full column rank (in fact, ϑ belongs to $\mathcal{H}^\infty_{n\times r}$). Partition ϑ as

$$\vartheta = \begin{bmatrix} \vartheta_1 \\ \vartheta_2 \end{bmatrix},$$

where $\vartheta_1 \in \mathcal{RH}^2_{q\times r}$ and $\vartheta_2 \in \mathcal{RH}^2_{(n-q)\times r}$. Suppose $\vartheta_1 \neq 0$. We may then write $\vartheta_1 = \varphi\zeta$, where $\varphi \in \mathcal{RH}^2_{q\times k}$, $k \leq r$ is co-outer, and $\zeta \in \mathcal{RH}^2_{k\times r}$ co-inner. Complete ζ to an inner matrix

$$\tilde{\zeta} = \begin{bmatrix} \zeta \\ \hat{\zeta} \end{bmatrix} \in \mathcal{RH}^\infty_{r\times r}.$$

Then

$$\vartheta\tilde{\zeta}^* = \begin{bmatrix} \varphi & 0 \\ \vartheta_2\zeta^* & \vartheta_2\hat{\zeta}^* \end{bmatrix}.$$

Define

$$\phi_1 = \tilde{C}^* \begin{bmatrix} \varphi \\ \vartheta_2\zeta^* \end{bmatrix}, \quad \phi_2 = \tilde{C}^* \begin{bmatrix} 0 \\ \vartheta_2\hat{\zeta}^* \end{bmatrix}, \quad \Phi_1 = \phi_1\phi_1^*, \quad \Phi_2 = \phi_2\phi_2^*.$$

Note that when $k = r$, Φ_2 disappears. Suppose therefore that $k < r$. Then ϕ_1 and ϕ_2 both have full column rank, and

$$\Phi = [\phi_1 \quad \phi_2] \begin{bmatrix} \phi_1^* \\ \phi_2^* \end{bmatrix} = \phi_1\phi_1^* + \phi_2\phi_2^* = \Phi_1 + \Phi_2.$$

We claim that both Φ_1 and Φ_2 are distinguishing special densities. We first check condition (12). Since Φ is distinguishing, Lemma 2 gives $H_o^*H_o\tilde{C}^*\vartheta = \lambda^2\tilde{C}^*\vartheta$, where H_o is $\mathcal{H}^\infty$-optimal, and $\lambda = \|H_o\|_\infty$. From this it follows that $H_o^*H_o\phi_1 = \lambda^2\phi_1$, and hence also $H_o^*H_o\Phi_1 = \lambda^2\Phi_1$. Thus Φ_1 satisfies condition (12). Similarly Φ_2 satisfies this condition. To check condition (11), we observe that $C\phi_2 = C\hat{C}^*\vartheta_2\hat{\zeta}^* = 0$. It follows that $\langle\Phi_2H_o^*, BQC\rangle = 0$ for each $Q \in \mathcal{H}^\infty_{p\times q}$. By Lemma 3 this means that condition (11) holds for Φ_2. Moreover, since

$$\langle\Phi_1H_o^*, BQC\rangle = \langle\Phi H_0^*, BQC\rangle - \langle\Phi_2H_0^*, BQC\rangle = 0,$$

we see that (11) holds for Φ_1, also. This shows that both Φ_1 and Φ_2 are distinguishing spectral densities for $\mathcal{S}(\Sigma)$. By considering Φ_1 we may therefore assume, without loss of generality, that a rational spectral density has a factorization

$$\Phi = \phi\phi^*, \qquad \phi = \tilde{C}^* \begin{bmatrix} \varphi \\ \chi \end{bmatrix}, \tag{17}$$

where $\chi \in \mathcal{L}^2_{(n-q)\times k}$ and either (1): $\varphi \in \mathcal{H}^2_{q\times k}$ is nonzero and co-outer; or (2): $\varphi = 0$. Of these two alternatives, the second takes care of the case $\vartheta_1 = 0$. Note that if C is square, then (17) reduces to $\phi = C^*\varphi$, with $\varphi \neq 0$ co-outer.

In the rest of this section we only consider spectral densities having a factorization of the form (17). We do not require them to be rational, however; that is, we allow φ and χ to be nonrational functions. The following lemma and its corollary show that in particular a distinguishing spectral density of this type may be assumed to have at most rank one almost everywhere on the unit circle.

LEMMA 4. *Let $\Sigma = F + B\mathcal{H}^\infty_{p\times q}C$ be a coset of the form* (4) *in which B is taken inner and C co-inner. Suppose $\Phi \in \mathcal{L}^\infty_{m\times n}$ is a spectral density having a factorization $\Phi = \phi\phi^*$, where*

$$\phi = \tilde{C}^* \begin{bmatrix} \varphi \\ \chi \end{bmatrix}, \quad \varphi \in \mathcal{H}^2_{q\times k}, \quad \chi \in \mathcal{L}^2_{(n-q)\times k},$$

and one of the following cases apply:

1. *$\varphi \neq 0$ and co-outer; or*
2. *$\varphi = 0$.*

Let ϕ_j be the jth column of ϕ, and $\hat{\Phi} = \phi_j\phi_j^*$. Then

(i) $\mathcal{A}(\Sigma;\Phi) = \mathcal{A}(\Sigma;\hat{\Phi})$.

(ii) *In case* 1 *the following statements are equivalent:*

$$H \in \mathcal{A}(\Sigma;\Phi), \tag{18}$$

$$H\phi \perp B\mathcal{H}^2_{p\times k}, \tag{19}$$

$$H\phi_j \perp B\mathcal{H}^2_p, \quad \textit{for each } 1 \leq j \leq k. \tag{20}$$

Proof. Statement (i) is proven separately for cases 1 and 2.

Case 1. Suppose $\varphi \neq 0$ is co-outer. The equivalence of (19) and (20) is a direct consequence of the definition of the inner product. We therefore prove that (18) holds if and only if (19) holds. Note first that for every $Q \in \mathcal{H}^\infty_{p\times q}$ we have

$$\langle \Phi H^*, BQC\rangle = \langle \Phi H^*, B[Q \quad 0]\tilde{C}\rangle = \langle \phi^* H^*, BQ\varphi\rangle. \tag{21}$$

From this and Lemma 3 it follows that $H \in \mathcal{A}(\Sigma;\Phi)$ if and only if

$$(H\phi, BQ\varphi) = 0 \tag{22}$$

for each $Q \in \mathcal{H}^\infty_{p\times q}$. If (19) holds then clearly (22) holds and hence, $H \in \mathcal{A}(\Sigma;\Phi)$. Conversely, suppose $H \in \mathcal{A}(\Sigma;\Phi)$. Since φ is co-outer there exists a scalar valued outer function g in $\mathcal{H}^\infty$ such that $g\varphi \in \mathcal{H}^\infty_{q\times k}$ and $\mathcal{H}^2_{p\times q}g\varphi$ is dense in $\mathcal{H}^2_{p\times k}$. Moreover, since

$B\mathcal{H}^\infty_{p\times q}$ is dense in $B\mathcal{H}^2_{p\times q}$, the functions $BQg\varphi$ must be dense in $B\mathcal{H}^2_{p\times k}$. But since (22) holds for all $BQg\varphi$, the denseness implies (19). This proves (ii).

Note that each column ϕ_j of ϕ is itself co-outer. In exactly the same way it therefore may be shown that $H \in \mathcal{A}(\Sigma;\hat{\Phi})$ if and only if (20) holds. Since (19) is equivalent to (20) it follows that $\mathcal{A}(\Sigma;\Phi) = \mathcal{A}(\Sigma;\hat{\Phi})$. This proves (i) in case 1.

Case 2. Similarly to (21) we have

$$\langle \hat{\Phi}H^*, BQC\rangle = \langle \phi_j^* H^*, BQ\varphi_j\rangle, \tag{23}$$

where φ_j is the jth column of φ. But since $\varphi = 0$, the expressions (21) and (23) both vanish. Hence by Lemma 3 (i), $\mathcal{A}(\Sigma;\Phi) = \Sigma$ and $\mathcal{A}(\Sigma;\hat{\Phi}) = \Sigma$. This proves (i) in Case 2. □

COROLLARY 1. *Suppose Φ is a distinguishing spectral density for $\mathcal{S}(\Sigma)$. Then $\hat{\Phi}$ is also a distinguishing spectral density for $\mathcal{S}(\Sigma)$.*

Proof. Let H be $\mathcal{H}^\infty$-optimal in Σ. By the lemma it suffices to prove that $\hat{\Phi}$ satisfies Lemma 2. But since Lemma 2 holds for Φ, and ϕ has full column rank, we have $H^*H\phi = \|H\|^2_\infty\phi$. Hence also $H^*H\phi_j = \|H\|^2_\infty\phi_j$, which implies $H^*H\hat{\Phi} = \|H\|^2_\infty\hat{\Phi}$. □

We have shown that a distinguishing spectral density admitting a factorization (17) may be replaced by one having the "rank one" form

$$\Phi = \phi\phi^*, \qquad \phi = \tilde{C}^* \begin{bmatrix} \varphi \\ \chi \end{bmatrix}, \tag{24}$$

where $\chi \in \mathcal{L}^2_{n-q}$, and either:

1. $0 \neq \varphi \in \mathcal{H}^2_q$ is co-outer; or
2. $\varphi = 0$.

When C is square, that is, $q = n$, this reduces to $\phi = C^*\varphi$, with $\varphi \in \mathcal{H}^2_n$ co-outer. Note again that $\varphi = 0$ if and only if $C\phi = 0$. Since we are only interested in $\mathcal{H}^\infty$-optimization over cosets with rational "data," in the sequel we assume Φ to be of the form (24). In §6 it will be shown that whenever a distinguishing Φ of this type exists, then it may in fact be taken rational. In the following two sections we may therefore assume that Φ is taken rational.

4. Dimension reduction and diagonalization. Consider a coset $\Sigma = F + B\mathcal{H}^\infty_{p\times q}C$ in $\mathcal{L}^\infty_{m\times n}$ of the form (4). We do not require B anc C to be inner and co-inner, respectively, only that B has full column rank and C full row rank on the unit circle. The purpose of this section is to determine a coset structure similar to that of Σ for a certain subset of $\mathcal{A}(\Sigma;\Phi)$. Moreover, when Φ is distinguishing, the functions in this subset have a convenient diagonalizable form and comprise all $\mathcal{H}^\infty$-optimal functions in Σ. This will play a central role in reducing class k optimization problems to class $k-1$ optimization problems.

The following lemma and its corollary state some basic facts used in the derivation of the reduction.

LEMMA 5. *Let A be a rational $m \times n$ matrix of column rank $k < n$. Then there exists a (tall) $P \in \mathcal{RH}^\infty_{n\times(n-k)}$ of full column rank on the unit circle such that:*

1. $AP = 0$.
2. *If $AX = 0$ for some $X \in \mathcal{H}^2_{n\times p}$, $(X \in \mathcal{L}^2_{n\times p})$, then $X = PM$ for some $M \in \mathcal{H}^2_{(n-k)\times p}$, $(M \in \mathcal{L}^2_{(n-k)\times p})$. Moreover, if X is rational, then M may be taken rational.*

Proof. In the trivial case $k = 0$, that is, $A = 0$, simply take P equal to the $n \times n$ identity matrix. Suppose therefore that $k > 0$. Use polynomial coprime factorization to write $A = D^{-1}N$, where D and N are polynomial matrices of dimensions $m \times m$ and

$m \times n$, respectively. By reduction to Hermite form we find an $n \times n$ unimodular polynomial matrix U such that

$$NU = [R \quad 0], \tag{25}$$

where R is an $m \times k$ polynomial matrix of full column rank. Partition $U = [\hat{U} \quad P]$, where $\hat{U}$ is $n \times k$ and P is $n \times (n-k)$. Then (25) implies $NP = 0$, and hence also $AP = 0$. Moreover, since U is unimodular P must have full column rank everywhere on the unit circle. This proves the first part of the lemma.

Suppose for $X \in \mathcal{H}^2_{n\times p}$ we have $AX = 0$, that is, $NX = 0$. Let $Y = U^{-1}X$. Since U is unimodular, Y also belongs to $\mathcal{H}^2_{n\times p}$. Partition Y as

$$Y = \begin{bmatrix} \hat{Y} \\ M \end{bmatrix},$$

where $\hat{Y} \in \mathcal{H}^2_{k\times p}$ and $M \in \mathcal{H}^2_{(n-k)\times p}$. By multiplying (25) from the right by Y and using the assumption $NX = 0$ we see that $R\hat{Y} = 0$. Moreover, since R has full column rank this yields $\hat{Y} = 0$. Hence

$$X = UY = [\hat{U} \quad P] \begin{bmatrix} 0 \\ M \end{bmatrix} = PM.$$

The last statement of the lemma follows immediately from the construction of M. The proof for the case $X \in \mathcal{L}^2_{n\times p}$ is completely analogous. □

A corresponding result holds for equations $XA = 0$.

COROLLARY 2. *Let A be a rational $m \times n$ matrix of row rank $k < m$. Then there exists a (wide) $P \in \mathcal{RH}^\infty_{(m-k)\times m}$ of full row rank on the unit circle such that:*

1. $PA = 0$.

2. *If $XA = 0$ for some $X \in \mathcal{H}^2_{p\times m}$, $(X \in \mathcal{L}^2_{p\times m})$, then $X = MP$ for some $M \in \mathcal{H}^2_{p\times(m-k)}$, $(M \in \mathcal{L}^2_{p\times(m-k)})$. Moreover, if X is rational, then M may be taken rational.*

Proof. For the proof, apply the lemma to the transpose of A. □

The matrix $P \in \mathcal{RH}^\infty_{n\times(n-k)}$ in Lemma 5 is far from unique and does not need to be chosen polynomial as in the proof. It can in fact be chosen inner, and in this form it is essentially unique. An analogous statement holds for P in Corollary 2.

Suppose $\Phi \in \mathcal{RL}^\infty_{n\times n}$ is a rational spectral density with $\Phi = \phi\phi^*$ for some $\phi \in \mathcal{L}^2_n$. In the context of a coset of the form (4) we obtain, as an immediate consequence of the previous lemma and its corollary, the following result.

LEMMA 6. *Suppose $\Sigma = F + B\mathcal{H}^\infty_{p\times q}C$ is a coset of the form* (4), *and $\Phi \in \mathcal{RL}^\infty_{n\times n}$ a rational spectral density of rank* 1. *Let $\Phi = \phi\phi^*$ be any factorization of Φ with $\phi \in \mathcal{RL}^2_n$.*

(i) *If $C\phi \neq 0$ and $q > 1$ then there exists a $P \in \mathcal{RH}^\infty_{(q-1)\times q}$ of full row rank on the unit circle such that:*

(a) $PC\phi = 0$; *and*

(b) $QC\phi = 0$ *if and only if* $Q = Q_1P$ *for some* $Q_1 \in \mathcal{H}^\infty_{p\times(q-1)}$.

(ii) *Suppose that $H_o = F + BQ_oC \in \mathcal{A}(\Sigma;\Phi)$. If $\phi^*H_o^*B \neq 0$ and $p > 1$, then there exists an $R \in \mathcal{RH}^\infty_{p\times p-1}$ of full column rank on the unit circle, not depending on the particular choice of H_o, such that*

(a) $\phi^*H_o^*BR = 0$; *and*

(b) $\phi^*H_o^*BQ = 0$ *if and only if* $Q = RQ_2$ *for some* $Q_2 \in \mathcal{H}^\infty_{(p-1)\times q}$.

Proof. By applying Corollary 2 to $C\phi$ we obtain P such that (a) and (b) hold. This proves (i). For (ii) we obtain R by applying Lemma 5 to $\phi^*H_o^*B$. By Lemma 3 (ii),

$H_o\phi = H_o'\phi$ for each $H_o' \in \mathcal{A}(\Sigma;\Phi)$. Hence R does not depend on the particular H_o we choose. □

For an arbitrary $H_o \in \Sigma$ define a set

$$\mathcal{D}(\Sigma;\Phi,H_o) = \{H \in \Sigma | H^*H\Phi = H_o^*H_o\Phi\}.$$

For $H_o \in \mathcal{A}(\Sigma;\Phi)$ the following theorem gives a useful parametrization of $\mathcal{D}(\Sigma;\Phi,H_o)$.

THEOREM 1. *Suppose* $\Sigma = F + \mathcal{B}\mathcal{H}_{p\times q}^{\infty}C \subseteq \mathcal{L}_{m\times n}^{\infty}$ *is a coset of the form* (4), *and* Φ *a rational spectral density of rank* 1. *Let* $\Phi = \phi\phi^*$ *be any factorization of* Φ *with* $\phi \in \mathcal{R}\mathcal{L}_n^2$. *Suppose* $H_o \in \mathcal{A}(\Sigma;\Phi)$. *Then* $\mathcal{D}(\Sigma;\Phi,H_o) \subseteq \mathcal{A}(\Sigma;\Phi)$ *and has one of the following forms:*

(i) *If* $C\phi = 0$ *and* $\phi^*H_o^*B = 0$, *then* $\mathcal{D}(\Sigma;\Phi,H_o) = \Sigma$.

(ii) *If* $C\phi = 0$, $\phi^*H_o^*B \neq 0$, *and* $p > 1$, *then*

$$\mathcal{D}(\Sigma;\Phi,H_o) = H_o + BR\mathcal{H}_{(p-1)\times q}^{\infty}C,$$

where $BR \in \mathcal{R}\mathcal{H}_{m\times(p-1)}^{\infty}$ *has full column rank on the unit circle*; R *chosen as in Lemma 6.*

(iii) *If* $C\phi \neq 0, q > 1$, *and* $\phi^*H_o^*B = 0$, *then*

$$\mathcal{D}(\Sigma;\Phi,H_o) = H_o + B\mathcal{H}_{p\times(q-1)}^{\infty}PC,$$

where $PC \in \mathcal{R}\mathcal{H}_{(q-1)\times n}^{\infty}$ *has full row rank on the unit circle*; P *chosen as in Lemma 6.*

(iv) *If* $C\phi \neq 0, q > 1, \phi^*H_o^*B \neq 0$, *and* $p > 1$, *then*

$$\mathcal{D}(\Sigma;\Phi,H_o) = H_o + BR\mathcal{H}_{(p-1)\times(q-1)}^{\infty}PC,$$

where $BR \in \mathcal{R}\mathcal{H}_{m\times(p-1)}^{\infty}$ *and* $PC \in \mathcal{R}\mathcal{H}_{(q-1)\times n}^{\infty}$ *have full column rank, respectively, full row rank, on the unit circle*; R *and* P *chosen as in Lemma* 6.

(v) *If either* (a): $C\phi \neq 0$ *and* $q = 1$, *or* (b): $\phi^*H_o^*B \neq 0$, *and* $p = 1$, *then*

$$D(\Sigma;\Phi,H_o) = \{H_o\}.$$

Proof. Since $H_o \in \mathcal{A}(\Sigma;\Phi)$ it is clear that $\mathcal{D}(\Sigma;\Phi,H_o) \subseteq \mathcal{A}(\Sigma;\Phi)$. To prove (i)–(v) we first note that each $H \in \Sigma$ may be written $H = H_o + BQC$ for some $Q \in \mathcal{H}_{p\times q}^{\infty}$. Moreover, the following identity then holds.

$$H^*H = H_o^*H_o + H_o^*BQC + C^*Q^*B^*H_o + C^*Q^*QC. \tag{26}$$

(i) Clearly each $H \in \mathcal{D}(\Sigma;\Phi,H_o)$ has the form $H = H_o + BQC$ for some $Q \in \mathcal{H}_{p\times q}^{\infty}$. Conversely, suppose $H = H_o + BQC$, where Q is an arbitrary function in $\mathcal{H}_{p\times q}^{\infty}$. Since $C\phi = 0$ and $\phi^*H_o^*B = 0$, identity (26) implies $H^*H\Phi = H_o^*H_o\Phi$. Thus $H \in \mathcal{D}(\Sigma;\Phi,H_o)$.

(ii) Suppose $H = H_o + BQC \in \mathcal{D}(\Sigma;\Phi,H_o)$. Then $H^*H\Phi = H_o^*H_o\Phi$, which by the full row rank property of ϕ gives $H^*H\phi = H_o^*H_o\phi$. Assumption $C\phi = 0$ together with identity (26) then implies $\phi^*H_o^*BQC = 0$, which immediately results in $\phi^*H_o^*BQ = 0$. Lemma 6(ii) therefore gives $Q = RQ_3$ for some $Q_3 \in \mathcal{H}_{(p-1)\times q}^{\infty}$, that is, $H = H_o + BRQ_3C$. Conversely, suppose $H = H_o + BQC$, with $Q = RQ_3$ for some $Q_3 \in \mathcal{H}_{(p-1)\times q}^{\infty}$. Then by Lemma 6(ii), $\phi^*H_o^*BQ = 0$. Hence identity (26) implies $H^*H\Phi = H_o^*H_o\Phi$, that is, $H \in \mathcal{D}(\Sigma;\Phi,H_o)$.

(iii) Suppose $H = H_o + BQC \in \mathcal{D}(\Sigma;\Phi,H_o)$. By Lemma 3(ii) we then have $BQC\phi = 0$, and hence $QC\phi = 0$. By Lemma 6(i) this gives $Q = Q_1P$ for some

$Q_1 \in \mathcal{H}^\infty_{p\times(q-1)}$, that is, $H = H_o + BQ_1PC$. Conversely, suppose $H = H_o + BQC$ is of this form, that is, $Q = Q_1P$ for some $Q_1 \in \mathcal{H}^\infty_{p\times(q-1)}$. Then by Lemma 6(i), $QC\phi = 0$. This, together with the assumption $\phi^* H_o^* B = 0$ and identity (26), gives $H^*H\Phi = H_o^* H_o \Phi$. Hence $H \in \mathcal{D}(\Sigma; \Phi, H_o)$.

(iv) Suppose $H = H_o + BQC \in \mathcal{D}(\Sigma; \Phi, H_o)$. As in (iii) we then have $QC\phi = 0$, and $Q = Q_1P$ for some $Q_1 \in \mathcal{H}^\infty_{p\times(q-1)}$. Hence identity (26), together with the fact that PC has full row rank, implies $\phi^* H_o^* BQ_1 = 0$. By Lemma 6(ii), we therefore have $Q_1 = RQ_2$ for some $Q_2 \in \mathcal{H}^\infty_{(p-1)\times(q-1)}$, that is, $H = H_o + BRQ_2PC$. Conversely, suppose $H = H_o + BQC$, with $Q = RQ_2P$ for some $Q_2 \in \mathcal{H}^\infty_{(p-1)\times(q-1)}$. Then $\phi^* H_o^* BQ = 0$ and $QC\phi = 0$. Hence identity (26) implies $H^*H\Phi = H_o^* H_o \Phi$, that is, $H \in \mathcal{D}(\Sigma; \Phi, H_o)$.

(v) Case a: Suppose $H = H_o + BQC \in \mathcal{D}(\Sigma; \Phi, H_o)$. By Lemma 3(ii), $BQC\phi = 0$, and hence also $QC\phi = 0$. Since $C\phi \neq 0$ is scalar valued, this gives $Q = 0$, that is, $H = H_o$.

Case b: Suppose $H = H_o + BQC \in \mathcal{D}(\Sigma; \Phi, H_o)$. Then $H^*H\phi = H_o^* H_o \phi$. Either $C\phi = 0$ or else by Lemma 3(ii), $BQC\phi = 0$, that is, $QC\phi = 0$. In either case, identity (26) implies $\phi^* H_o^* BQ = 0$. But since $\phi^* H_o^* B$ is scalar valued, this in turn implies $Q = 0$, that is, $H = H_o$. □

By Theorem 1, $\mathcal{D}(\Sigma; \Phi, H_o)$ is for each $H_o \in \mathcal{A}(\Sigma; \Phi)$ a coset, which, except for case (v), is of type (4). Since the matrices R and P do not depend on H_o, this also holds for the subspace associated with the coset, for instance, in case (iv) the subspace $BR\mathcal{H}^\infty_{p\times q}PC$.

We will identify $\mathcal{D}(\Sigma; \Phi, H_o)$ with a subset $\hat{\Sigma}(\Phi, H_o)$ of $\mathcal{H}^\infty_{m\times(n-1)}$ having a parametrization similar to that of Theorem 1. Consider therefore the rational spectral density $\Phi = \phi\phi^*, \phi \in \mathcal{L}^2_n$ in Theorem 1. Clearly

$$\phi = fg \tag{27}$$

for some inner $f \in \mathcal{RH}^2_n$ and some scalar valued function $g \in \mathcal{RL}^2$. To see this we choose a rational scalar valued inner function θ such that $\theta^*\phi \in \mathcal{H}^2_n$. By an inner-outer factorization we have $\theta\phi = fh$, where $f \in \mathcal{RH}^2_n$ is inner and $h \in \mathcal{RH}^2$ is outer. Thus we may take $g = \theta^{-1}h$. The inner vector f clearly depends on the choice of the factor ϕ and the function θ. However, by a co-inner/co-outer factorization of f and a subsequent absorption of the co-inner part in g, we could choose f to be co-outer, also. This inner/co-outer f would then be uniquely determined by Φ. Since we make no use of this fact we will only require f to be inner. Hence f is unique only up to multiplication by some scalar valued inner function. Via f we may relate Φ to a square inner matrix

$$V := [V_a \quad V_b] \in \mathcal{RH}^\infty_{n\times n}, \tag{28}$$

where $V_a = f$, and $V_b \in \mathcal{RH}^\infty_{n\times(n-1)}$ is an inner complement of V_a. If $n = 1$, this reduces to $V = V_a$. Define a mapping $\gamma : \mathcal{H}^\infty_{m\times n} \to \mathcal{H}^\infty_{m\times(n-1)}$ by $\gamma(H) = HV_b$. For $H_o \in \Sigma$ let

$$\hat{\Sigma}(\Phi, H_o) = \gamma(\mathcal{D}(\Sigma; \Phi, H_o)). \tag{29}$$

COROLLARY 3. *Suppose $H_o \in \mathcal{A}(\Sigma; \Phi)$. Then corresponding to* (i)–(v) *of Theorem* 1, *the following cases occur:*

(i) $\hat{\Sigma}(\Phi, H_o) = H_oV_b + B\mathcal{H}^\infty_{p\times q}CV_b$, *where* CV_b *has full row rank on the unit circle.*

(ii) $\hat{\Sigma}(\Phi, H_o) = H_oV_b + B\mathcal{H}^\infty_{p\times(q-1)}PCV_b$, *where* PCV_b *has full row rank on the unit circle.*

(iii) $\hat{\Sigma}(\Phi, H_o) = H_oV_b + BR\mathcal{H}^\infty_{(p-1)\times q}CV_b$, *where* BR *has full column rank, and* CV_b *has full row rank, on the unit circle.*

(iv) $\hat{\Sigma}(\Phi, H_o) = H_oV_b + BR\mathcal{H}^\infty_{(p-1)\times(q-1)}PCV_b$, *where* BR *has full column rank, and* PCV_b *has full row rank, on the unit circle.*

(v) $\hat{\Sigma}(\Phi, H_o) = \{H_oV_b\}$.

Moreover γ *determines a one-to-one correspondence between the coset* $\mathcal{D}(\Sigma; \Phi, H_o)$ *and the coset* $\hat{\Sigma}(\Phi, H_o)$.

Proof. For each case the corresponding case of Theorem 1 shows that $\hat{\Sigma}(\Phi, H_o)$ is of the required form, except possibly for the rank conditions on CV_b and PCV_b. To check these we note in cases (i) and (ii) that $CV_a = 0$, which implies that CV_b has full row rank on the unit circle. In cases (iii) and (iv), $PCV_a = 0$, which implies that PCV_b has full row rank on the unit circle. To prove the last statement we only need to show that γ restricted to $\mathcal{D}(\Sigma; \Phi, H_o)$ is injective. But in each case (i)–(iv) this follows from the full rank properties of BR, CV_b, and PCV_b. □

If so desired we could use a co-inner factor U_a of the row vector $\phi^* H_o^*$ to construct a rational square inner matrix $U = [U_a^{\mathrm{T}} \quad U_b^{\mathrm{T}}]^{\mathrm{T}}$. By redefining γ such that $\gamma(H) = U_bHV_b$ the set $\hat{\Sigma}(\Phi, H_o)$ would then take a more symmetric form. For instance, case (iv) of Corollary 3 would read

$$\hat{\Sigma}(\Phi, H_o) = U_bH_oV_b + U_bBR\mathcal{H}^\infty_{(p-1)\times(q-1)}PCV_b,$$

where U_bBR has full column rank and PCV_b has full row rank, on the unit circle. This corresponds closely to the form used in [27].

COROLLARY 4. *For any* $H \in \mathcal{D}(\Sigma; \Phi, H_o)$ *we have*

$$\mathcal{D}(\Sigma; \Phi, H) = \mathcal{D}(\Sigma; \Phi, H_o),$$
$$\hat{\Sigma}(\Phi, H) = \hat{\Sigma}(\Phi, H_o).$$

Proof. This follows directly from the coset property of $\mathcal{D}(\Sigma; \Phi, H_o)$ in Theorem 1. □

According to Corollary 3, $\hat{\Sigma}(\Phi, H_o)$ is a coset which, except for the case (v), is of type (4). Furthermore, via the mapping $\gamma, \hat{\Sigma}$ may be identified with $\mathcal{D}(\Sigma; \Phi, H_o)$. The inverse image $\gamma^{-1}(G), G \in \hat{\Sigma}(\Phi, H_o)$ is easily computed. For instance, in case (iv) of Corollary 3 we have $\gamma^{-1}(G) = H_o + BR^{-1}B^{-1}(G - H_oV_b)(PCV_b)^{-1}C$, where R^{-1} and B^{-1} are left inverses of R and B, respectively, and $(PCV_b)^{-1}$ is a right inverse of PCV_b. This becomes even easier if we know the $Q \in \mathcal{H}^\infty_{p\times q}$ determining G. Just take $\gamma^{-1}(G) = H_o + BRQPC$. This is, for instance, the case when G is an H^∞-optimal function computed via the polynomial approach.

The following theorem and its corollary accentuate the usefulness of $\hat{\Sigma}(\Phi, H_o)$.

THEOREM 2. *Suppose* $\Sigma = F + \mathcal{B}\mathcal{H}^\infty_{p\times q}C \in \mathcal{L}^\infty_{m\times n}$, $n > 1$, *is a coset of the form* (4), *and* Φ *a rational spectral density for* $\mathcal{S}(\Sigma)$ *of rank* 1. *Let* $\Phi = \phi\phi^*$ *be any factorization of* Φ *with* $\phi \in \mathcal{RL}^2_n$. *Let* $V = [V_a \quad V_b]$ *be as in* (28), *where* V_a *is the inner factor of* ϕ *given by* (27), *and* V_b *an inner complement of* V_a. *Suppose* $H_o \in \Sigma$ *satisfies* $H_o^* H_o \Phi = \lambda^2 \Phi$. *Then* $H \in \Sigma$ *belongs to* $\mathcal{D}(\Sigma; \Phi, H_o)$ *if and only if*

$$V^*H^*HV = \begin{bmatrix} \lambda^2 & 0 \\ 0 & V_b^*H^*HV_b \end{bmatrix}. \tag{30}$$

Proof. Suppose $H = F + BQC \in \mathcal{D}(\Sigma; \Phi, H_o)$. Since $\phi = gV_a$ for some scalar valued rational function g, it follows that $H^*HV_a = H_o^*H_oV_a = \lambda^2 V_a$. Thus

$$V^*H^*HV = \begin{bmatrix} V_a^*H^*HV_a & V_a^*H^*HV_b \\ V_b^*H^*HV_a & V_b^*H^*HV_b \end{bmatrix} = \begin{bmatrix} \lambda^2 & 0 \\ 0 & V_b^*H^*HV_b \end{bmatrix}. \tag{31}$$

Conversely, suppose (30) holds for some $H \in \Sigma$. Since $\Phi = gg^*V_aV_a^*$ we have

$$V^*\Phi V = \begin{bmatrix} gg^* & 0 \\ 0 & 0 \end{bmatrix}. \tag{32}$$

Hence using (30)

$$V^*H^*H\Phi V = V^*H^*HV \cdot V^*\Phi V = \lambda^2 \begin{bmatrix} gg^* & 0 \\ 0 & 0 \end{bmatrix}, \tag{33}$$

which by (32) is the same as $H^*H\Phi = \lambda^2\Phi$. □

COROLLARY 5. *Suppose H_o is $\mathcal{H}^\infty$-optimal in Σ, that is, $\|H_o\|_\infty = \lambda$, and that Φ is distinguishing for $\mathcal{S}(\Sigma)$. Then*

(i) *A function $H \in \Sigma$ is $\mathcal{H}^\infty$-optimal in Σ if and only if $H \in \mathcal{D}(\Sigma; \Phi, H_o)$ and $\|HV_b\|_\infty \le \lambda$.*

(ii) *For $k > 1$, a function $H \in \Sigma$ is class k optimal (super-optimal) in Σ if and only if $H \in \mathcal{D}(\Sigma; \Phi, H_o)$ and HV_b is class $k-1$ optimal (super-optimal) in $\hat{\Sigma}(\Phi, H_o)$.*

Proof. (i) Suppose H is $\mathcal{H}^\infty$-optimal in Σ. Then, since H_o is also $\mathcal{H}^\infty$-optimal and Φ distinguishing, it follows from Lemmas 1 and 2 that $H \in \mathcal{D}(\Sigma; \Phi, H_o)$. Moreover, by Theorem 2, H has a diagonalization (30). Since V is square and inner this means that the singular values of H are λ^2 together with those of HV_b. By the $\mathcal{H}^\infty$-optimality of H this means that $\|HV_b\|_\infty \le \lambda$. Conversely, suppose that $H \in \mathcal{D}(\Sigma; \Phi, H_o)$ and that $\|HV_b\|_\infty \le \lambda$. Then by Theorem 2, H has a diagonalization (30). Thus $\|H\|_\infty = \lambda$, that is, H is $\mathcal{H}^\infty$-optimal in Σ.

(ii) The proof is similar to the proof of (i). □

We finally consider the case where the functions in Σ are column vectors.

THEOREM 3. *Suppose $\Sigma \subseteq \mathcal{L}^\infty_{m\times n}$ is a coset of the form* (4) *with either $m = 1$ or $n = 1$. Then any $\mathcal{H}^\infty$-optimal function in Σ is necessarily super-optimal. Moreover:*

(i) *If $\mathcal{S}(\Sigma)$ admits a distingiushing spectral density, then the $\mathcal{H}^\infty$-optimal function is unique.*

(ii) *If $\mathcal{S}(\Sigma)$ does not admit a distingiushing spectral density, then an $\mathcal{H}^\infty$-optimal function may be nonunique.*

Proof. (i) Suppose $H_o \neq 0$ is $\mathcal{H}^\infty$-optimal in $\Sigma = F + B\mathcal{H}^\infty_{p\times q}C \subseteq \mathcal{H}^\infty_{m\times n}$. Consider first the case $m = 1$. Since B is tall, $m = 1$ implies $p = 1$, that is, B is scalar valued. Since $\lambda = \|H_o\|_\infty > 0$, we have, by Lemma 2, $H_o\phi \neq 0$ and hence $\phi^*H_o^*B \neq 0$. Theorem 1 (v) therefore shows that $\mathcal{D}(\Sigma; \Phi, H_o) = \{H_o\}$. Thus H_o is the only $\mathcal{H}^\infty$-optimal function in Σ. Suppose now that $n = 1$. Then $q = 1$ and C is scalar valued. Consequently, $C\phi \neq 0$. By Theorem 1(v) we therefore have $\mathcal{D}(\Sigma; \Phi, H_o) = \{H_o\}$. Hence again H_o is the only $\mathcal{H}^\infty$-optimal function in Σ.

(ii) This may be shown by constructing simple counter examples (also, cf. [10]). □

5. Super-optimization. The results of the previous section, in particular Theorem 1 and Corollaries 3 and 5, provide us with the following procedure for computing class k optimal functions in a coset

$$\Sigma_1 := F + B\mathcal{H}^\infty_{p\times q}C \subseteq \mathcal{H}^\infty_{m\times n}$$

of the type (4).

Step 1. Apply the polynomial approach, or any other approach, to obtain an $\mathcal{H}^\infty$-optimal function $H_1 = F + BQ_1C$ in Σ_1, where $Q_1 \in \mathcal{Q}_1 := \mathcal{H}^\infty_{p\times q}$. This determines a class 1 optimal function in Σ.

Step 2. If a distinguishing spectral density Φ_1 exists for $\mathcal{S}(\Sigma_1)$, then construct as in (28) the inner matrix $V_1 = [V_{1a} \quad V_{1b}]$ associated with Φ_1. In cases (i)–(vi) of Theorem 1, $\mathcal{D}(\Sigma_1; \Phi_1, H_1)$ has the form $H_1 + B\mathcal{Q}_2C$, where $\mathcal{Q}_2$ is a subspace of $\mathcal{Q}_1$. In case (iv) for instance, $\mathcal{Q}_2 = R_1\mathcal{H}^\infty_{(p-1)\times(q-1)}P_1$, where R_1 and P_1, respectively, denote the matrices R and P of Theorem 1. Define the coset

$$\Sigma_2 := H_1V_{1b} + B\mathcal{Q}_2CV_{1b}, \in \mathcal{H}^\infty_{m\times(n-1)}. \tag{34}$$

By Corollary 3 this is of the type (4).

Apply the polynomial approach, or any other approach, to Σ_2 to obtain an $\mathcal{H}^\infty$-optimal function $H_2 = H_1V_{1b} + BQ_2CV_{1b}$ in Σ_2, where $Q_2 \in \mathcal{Q}_2$. By Corollary 5,

$$H_1 + BQ_2C = F + B(Q_1 + Q_2)C$$

is then class 2 optimal in Σ_1.

Step 3. If a distinguishing spectral density Φ_2 exists for $\mathcal{S}(\Sigma_2)$, we may as before construct the inner matrix $V_2 = [V_{2a} \quad V_{2b}]$ associated with Φ_2. In cases (i)–(iv) of Theorem 1 $\mathcal{D}(\Sigma_2; \Phi_2, H_2)$ is a coset $H_2 + B\mathcal{Q}_3CV_{1b}$, where $\mathcal{Q}_3$ is a subspace of $\mathcal{Q}_2$. In case (iv) for instance, we may have $\mathcal{Q}_3 = R_1R_2\mathcal{H}^\infty_{(p-2)\times(q-2)}P_2P_1$. Define the coset

$$\Sigma_3 := H_2V_{2b} + B\mathcal{Q}_3CV_{1b}V_{2b} \in \mathcal{H}^\infty_{m\times(n-2)}. \tag{35}$$

By Corollary 3 this is of the type (4).

Apply the polynomial approach, or any other approach, to Σ_3 to obtain an $\mathcal{H}^\infty$-optimal function $H_3 = H_2V_{2b} + BQ_3CV_{1b}V_{2b}$ in Σ_3, where $Q_3 \in \mathcal{Q}_3$. By Corollary 5, $H_2 + BQ_3CV_{1b} = H_1V_{1b} + B(Q_2 + Q_3)CV_{1b}$ is then class 2 optimal in Σ_2, and

$$H_2 + B(Q_2 + Q_3)C = F + B(Q_1 + Q_2 + Q_3)C$$

is class 3 optimal in Σ_1.

Step k. In the general kth step we obtain a coset

$$\Sigma_k = H_{k-1}V_{(k-1),b} + B\mathcal{Q}_kCV_{1b}\cdots V_{(k-1),b} \in \mathcal{H}^\infty_{m\times(n-k+1)} \tag{36}$$

which, in cases (i)–(iv) of Theorem 1, is of type (4). In case (iv) for instance, we may have $\mathcal{Q}_k = R_1R_2\cdots R_{k-1}\mathcal{H}^\infty_{(p-k+1)\times(q-k+1)}P_{k-1}\cdots P_1$. Use the polynomial approach, or any other approach, to find an $\mathcal{H}^\infty$-optimal function

$$H_k = H_{k-1}V_{(k-1),b} + BQ_kCV_{1b}\cdots V_{(k-1),b}$$

in Σ_k. By Corollary 5, the function $H_{k-1} + BQ_kCV_{1b}\cdots V_{(k-2),b}$ is then class 2 optimal in Σ_{k-1}. By repeated use of this argument we find that the function

$$F + B(Q_1 + \cdots + Q_k)C$$

is class k optimal in Σ_1.

The process stops when one of the following situations occurs:

1. A distinguishing spectral density Φ_k exists for $\mathcal{S}(\Sigma_k)$ but $\mathcal{D}(\Sigma_k; \Phi_k, H_k) = \{H_k\}$; that is, we are in situation (v) of Theorem 1. The class k optimal function in Σ corresponding to H_k is then unique and hence also super-optimal.

2. No distinguishing spectral density exists, and either $m-k+1=1$ or $n-k+1=1$. The functions in Σ_k then have only one nonzero singular value, and this is clearly minimized by H_k. Consequently, the class k optimal function in Σ corresponding to H_k is super-optimal. It need not be unique, however (cf. Theorem 3).

3. No distinguishing spectral density exists for $\mathcal{S}(\Sigma_k)$ and $m-k+1>1$ and $n-k+1>1$. The class k optimal function in Σ corresponding to H_k is then, in general, neither unique nor super-optimal.

We remark that cases (2) and (3) never occur for one-block problems (cf. [18]). Moreover, as clearly demonstrated by, for instance, [10], [11], [1], [12], and [13], two- and four-block problems commonly admit a distinguishing spectral density. It is therefore believed that situation (3) occurs less frequently in practice. If the process terminates in state (3), in principle we could proceed by searching for a super-optimal function in Σ_k instead of for only an $\mathcal{H}^\infty$-optimal function. The corresponding function in Σ would then be super-optimal. This, however, seems to be a problem of totally different character, and its treatment is outside the scope of this paper.

In any case, the present method allows computation of super-optimal solutions for a large class of model-matching problems, including all one-block problems and a substantial number of genuine two- and four-block problems. For the remaining problems we only obtain a class k optimal solution, which does not need to be super-optimal.

6. Construction of the spectral density. In this section we consider the problem of constructing the distinguishing spectral density needed to carry out the dimension reduction step of the super-optimization algorithm. Since we do not make use of the spectral density itself, but rather its spectral factor, we may confine our study to spectral densities of the form $\Phi=\phi\phi^*\in\mathcal{L}^1_{n\times n}$, where ϕ is an arbitrary vector in $\mathcal{L}^2_n$, not necessarily rational. Exactly as for the rational case, Φ then has a representation (24). Unfortunately, for some four-block problems a distinguishing spectral density does not need to exist. However, we show that if one with the prescribed form exists, then there also exists a rational one. Explicit formulas for computing it are given. These in fact provide a necessary and sufficient condition for the existence of a distinguishing spectral density of the given form.

The equalizer principle binds a distinguishing spectral density Φ rather tightly to the $\mathcal{H}^\infty$-optimal functions H_o. This strongly suggests that in computing H_o some of the intermediate results may assist in finding Φ (cf., e.g., [10], [11], [12], and [13]). We take another approach, however, and consider a method of determining Φ that relies only on the knowledge of a rational $\mathcal{H}^\infty$-optimal function H_o and its norm $\|H_o\|_\infty$. With this we run the risk of repeating some of the effort taken in determining H_o. On the other hand, it has the attractive property of being independent of the $\mathcal{H}^\infty$-optimization method used to obtain H_o. In order of computational complexity we separate the one-, two-, and four-block cases. To simplify the exposition, without loss of generality, we assume that the matrices B and C of the coset $\Sigma=F+B\mathcal{H}^\infty_{p\times q}C$ have been chosen inner and co-inner, respectively.

6.1. The one-block case. We begin with the simplest case, the one-block problem. Consequently, we take $\Sigma=F+B\mathcal{H}^\infty_{p\times q}C$ and B and C square and inner. It is shown in [18] that in the one-block case a distinguishing spectral density of type (24) always exists for $\mathcal{S}(\Sigma)$. Suppose therefore that ϕ is such a distinguishing spectral density. Since C is square this means that for some co-outer $\varphi\in\mathcal{H}^2_n$ we have $\Phi=\phi\phi^*$, where $\phi=C^*\varphi$. Let H_o be an arbitrary rational $\mathcal{H}^\infty$-optimal function in Σ, and let $\lambda=\|H_o\|_\infty$. From Lemma 2

$$(H_o^*H_o-\lambda^2 I)C^*\varphi=0. \tag{37}$$

Suppose $H_o^* H_o - \lambda^2 I$ has column rank $n-k$. Apply Corollary 2 to equation (37) to get a full column rank matrix $P \in \mathcal{RH}^\infty_{n\times k}$ such that $\varphi = Pu$ for some $u \in \mathcal{H}^2_k, u \neq 0$. Moreover, from (37) it is easily seen that $H_o C^* P$ has full column rank k. Using Lemma 4(ii) we also find that $H_o C^* P u$ is perpendicular to $B\mathcal{H}^2_m$. Thus,

$$\Psi H_o C^* P u \in \mathcal{H}^2_m \ominus \Psi B\mathcal{H}^2_m, \tag{38}$$

where Ψ is any $n\times n$ rational square inner matrix such that $\Psi H_o C^* P \in \mathcal{H}^\infty_{m\times k}$. One way to obtain Ψ is to start with a polynomial coprime factorization $H_o C^* P = D^{-1}N$ and consider a polynomial spectral factorization $DD^* = \hat{D}\hat{D}^*$, where $\det(\hat{D})$ has no roots in the closed unit disk. Then $\Psi = \hat{D}^{-1}D$ has the required properties. Recall that $\mathcal{H}^2_m \ominus \Psi B\mathcal{H}^2_m$ is a finite-dimensional space, which (up to equivalence) contains only rational functions. Since $\Psi H_o C^* P$ has full column rank, this means that u in fact must be rational (modulo some null function). Choose a basis for $\mathcal{H}^2_m \ominus \Psi B\mathcal{H}^2_m$ consisting of rational vectors $E_1, \dots, E_p$. A suitable choice of $E_1, \dots, E_p$ is given in Appendix A. Let $E = [E_1, E_2, \dots, E_p]$ be the matrix having E_i as its ith column. Then for some $\xi \in C^p$,

$$\Psi H_o C^* P u = E\xi. \tag{39}$$

Conversely, suppose that $u \in \mathcal{H}^2_k$ satisfies (39). We may then assume u to be rational as well. Clearly (38) holds, also. Write $\varphi\theta = Pu$, where $\varphi \in \mathcal{RH}^2_n$ is co-outer and $\theta \in \mathcal{RH}^2$ is co-inner. Then taking $\phi = C^*\varphi$ it is easily seen from (38) that $H_o\phi \perp \theta^* B\mathcal{H}^2_m$. But since $B\mathcal{H}^2_m \subseteq \theta^* B\mathcal{H}^2_m$, this implies that $H_o\phi \perp B\mathcal{H}^2_m$. Let

$$\Phi = C^* P u u^* P^* C = \phi\phi^*, \tag{40}$$

where $\phi = C^*\varphi$. By Lemma 4(ii) we then have $H_o \in \mathcal{A}(\Sigma;\Phi)$, that is, the first condition (12) of Lemma 1 holds. On the other hand, the choice of P guarantees that $(H_o^* H_o - \lambda^2 I)\phi = 0$. By Lemma 2 this shows that Φ satisfies the second condition (12) of Lemma 1. Consequently, this Φ is distinguishing and rational.

To summarize, we have shown that the problem of finding a distinguishing spectral density Φ is equivalent to solving equation (39) for a nonzero $u \in \mathcal{RH}^2_k$ and a nonzero $\xi \in C^p$. Moreover, via (40) such a solution gives rise to rational Φ. A way to solve (39) will be considered in Appendix B.

It is worthwhile to note the following special case of equation (39). Suppose that the $\mathcal{H}^\infty$-optimal solution H_o satisfies $H_o^* H_o = \lambda^2 I$. This can always be arranged for when the polynomial approach is used to obtain H_o. We may then take $P = I$ and $\varphi = u$. Moreover, $\Psi H_o C^*$ is then invertible. In fact, $(\Psi H_o C^*)^{-1} = \lambda^{-2} C H_o^* \Psi^*$. Hence, equation (39) is equivalent to the somewhat simpler equation

$$u = \lambda^{-2} C H_o^* \Psi^* E\xi. \tag{41}$$

6.2. The two-block case. For the two-block case we may assume that the coset is of the form $\Sigma = F + B\mathcal{H}^\infty_{p\times n} C$, where B is tall and C is square. Otherwise consider transpositions. Suppose that $\mathcal{S}(\Sigma)$ has a distinguishing spectral density Φ of type (24). Since C is square we may, as for the one-block case, assume that

$$\Phi = C^*\varphi\varphi^* C,$$

where $\varphi \neq 0$ is a co-outer vector in $\mathcal{H}^2_n$. Suppose that $H_o \in \Sigma$ is $\mathcal{H}^\infty$-optimal, and let $\lambda = \|H_o\|_\infty$. By Lemma 2, $(H_o^* H_o - \lambda^2 I)C^*\varphi = 0$. Suppose $H_o^* H_o - \lambda^2 I$ has column rank $n-k$. Then by Corollary 2 there exists a rational $P \in \mathcal{RH}^\infty_{n\times k}$ such that

$$\varphi = Pu \tag{42}$$

for some $u \in \mathcal{H}_k^2, u \neq 0$. Moreover, by Lemma 4(ii)

$$H_o C^* P u \perp B\mathcal{H}_p^2. \tag{43}$$

Choose an inner matrix $\Psi \in \mathcal{RH}_{p\times p}^\infty$ such that $\Psi B^* H_o C^* P \in \mathcal{RH}_{p\times k}^\infty$. Then (43) implies

$$\Psi B^* H_o C^* P u \in \mathcal{H}_p^2 \ominus \Psi \mathcal{H}_p^2. \tag{44}$$

Note that this condition is similar to (39) except that $\Psi B^* H_o C^* P$ does not need to have full column rank. By applying the Hermite form to the numerator of a polynomial left coprime factorization of $\Psi B^* H_o C^* P$, we may write $\Psi B^* H_o C^* P = [R \quad 0]V$, where R is a rational matrix of full column rank, say r, and V is a unimodular polynomial matrix. Define $v \in \mathcal{H}_r^2$ and $w \in \mathcal{H}_{k-r}^2$ by

$$\begin{bmatrix} v \\ w \end{bmatrix} = Vu. \tag{45}$$

Then (44) implies

$$Rv \in \mathcal{H}_p^2 \ominus \Psi \mathcal{H}_p^2. \tag{46}$$

We have shown that the existence of Φ implies the existence of a $v \in \mathcal{H}_r^2$ such that (46) holds. But since the space $\mathcal{H}_p^2 \ominus \Psi \mathcal{H}_p^2$ consists of only rational functions (up to equivalence), and R has full column rank, this means that v must in fact be rational (up to equivalence).

Conversely, suppose that the matrix R can be formed in the way described, and that (46) holds for some $v \in \mathcal{H}_r^2$. We may as well assume v to be rational. Take an arbitrary $w \in \mathcal{H}_{k-r}^2$ and let

$$u := V^{-1} \begin{bmatrix} v \\ w \end{bmatrix}. \tag{47}$$

Then u satisfies (44). Write $Pu = \varphi\theta$, where $\varphi \in \mathcal{H}_n^2$ is co-outer and $\theta \in \mathcal{H}^2$ is co-inner. Take $\phi = C^*\varphi$, and

$$\Phi = C^* P u u^* P^* C = \phi\phi^*. \tag{48}$$

As in the one-block case we then find that $H_o \phi \perp B\mathcal{H}_p^2$. Consequently, $H_o \in \mathcal{A}(\Sigma; \Phi)$. Moreover, the choice of P guarantees that $(H_o^* H_o - \lambda^2 I)\phi = 0$. Thus by Lemma 2 and Lemma 1, Φ must be distinguishing. Note that if w also is taken rational, then Φ will be rational.

To summarize, we have shown that a distinguishing spectral density Φ exists if and only if the matrix R can be constructed and there exists a solution $v \neq 0$ of (46). Moreover, via (47) and (48), a solution of (46) gives rise to a distinguishing spectral density which may be chosen rational.

It remains to solve (46). Note that this is a relation exactly of type (38). We may therefore proceed as in the one-block case.

6.3. The four-block case. In the most general situation, the four-block case, we have B tall and C wide in the coset $\Sigma = F + B\mathcal{H}_{p\times p}^\infty C$. Suppose that $\mathcal{S}(\Sigma)$ has a distinguishing spectral density Φ of the form (24), that is,

$$\Phi = \phi\phi^*, \qquad \phi = \tilde{C}^* \begin{bmatrix} \varphi \\ \chi \end{bmatrix} \neq 0, \tag{49}$$

for some $\varphi \in \mathcal{H}_q^2$ and $\chi \in \mathcal{L}_{n-q}^2$, where one of the following two cases occurs:

1. $0 \neq \varphi \in \mathcal{H}_q^2$ is co-outer; or
2. $\varphi = 0$.

Here $\tilde{C} = [C^T \quad \tilde{C}^T]^T$ is a completion of C to a square $n \times n$ inner matrix, with $\hat{C}$ being a co-inner complement of C. By Lemma 2 we then have

$$(H_o^* H_o - \lambda^2 I)\tilde{C}^* \begin{bmatrix} \varphi \\ \chi \end{bmatrix} = 0. \tag{50}$$

Consequently, by Corollary 2 there exists a full column rank $n \times k$ polynomial matrix

$$P = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \tag{51}$$

partitioned with P_1 having q rows, such that $\varphi \in \mathcal{L}_q^2$ and $\chi \in \mathcal{L}_{n-q}^2$ satisfy (50) if and only if

$$\begin{bmatrix} \varphi \\ \chi \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} w \tag{52}$$

for some $w \in \mathcal{L}_k^2$.

Case A. Consider the situation where $P_1 \neq 0$ with column rank r. Suppose again that Φ is a distinguishing spectral density given by $\varphi \in \mathcal{H}_q^2$ and $\chi \in \mathcal{L}_{n-q}^2$ as in (49), and that w in $\mathcal{L}_k^2$ satisfies (52). Let

$$P_1 = U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \qquad V := \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

be, for instance, a Smith decomposition of P_1, where U and V are unimodular polynomial matrices, and D a polynomial diagonal matrix of say type $r \times r$. In the partition of V, V_1 is assumed to have r rows. Factorize $D = D_2 D_1$, with the diagonal elements of D_1 having all roots inside the open unit disk and the diagonal elements of D_2 having no roots in the open unit disk. Then $\varphi = P_1 w \in \mathcal{H}_q^2$ implies that $D_2 D_1 V_1 w \in \mathcal{H}_r^2$. Since $u := D_1 V_1 w \in \mathcal{L}_r^2$ and the elements of D_2 are outer, this means that u in fact belongs to $\mathcal{H}_r^2$ (cf. [17, Cor. 3, p. 12]). Clearly $v := V_2 w \in \mathcal{L}_{k-r}^2$. Note that $\varphi = 0$ implies $u = 0$. Moreover, let

$$M = PV^{-1} \begin{bmatrix} D_1^{-1} & 0 \\ 0 & I_{k-r} \end{bmatrix},$$

where I_{k-r} is the $(k-r) \times (k-r)$ identity matrix. Then M has full column rank and

$$\begin{bmatrix} \varphi \\ \chi \end{bmatrix} = M \begin{bmatrix} u \\ v \end{bmatrix}. \tag{53}$$

On the other hand, suppose $u \in \mathcal{H}_r^2$ and $v \in \mathcal{L}_{k-r}^2$. Let φ and χ be given by (53). Then clearly $\chi \in \mathcal{L}_{n-q}^2$, and from the definition of M it follows that

$$\varphi = U \begin{bmatrix} D_2 D_1 & 0 \\ 0 & 0 \end{bmatrix} V \cdot V^{-1} \begin{bmatrix} D_1^{-1} & 0 \\ 0 & I_{k-r} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = U \begin{bmatrix} D_2 u \\ 0 \end{bmatrix} \in \mathcal{H}_q^2.$$

Note that $u = 0$ implies $\varphi = 0$. We claim that $H_o \tilde{C}^* M$ has full column rank. To see this we take $u \in \mathcal{L}_r^2$ and $v \in \mathcal{L}_{k-r}^2$ such that

$$H_o \tilde{C}^* M \begin{bmatrix} u \\ v \end{bmatrix} = 0.$$

Then by (50) we have

$$\tilde{C}^* M \begin{bmatrix} u \\ v \end{bmatrix} = \lambda^{-2} H_o^* H_o \tilde{C}^* M \begin{bmatrix} u \\ v \end{bmatrix} = 0.$$

But since $\tilde{C}^* M$ has full column rank, this implies $u = 0$, and $v = 0$. Hence $H_o \tilde{C}^* M$ also has full column rank. We summarize these results as a lemma.

LEMMA 7. *Suppose that the matrix P_1 in (51) has column rank $0 < r \leq k$.*

(i) *There exists an $M \in \mathcal{RL}^{\infty}_{n \times k}$ of full column rank, such that $\varphi \in \mathcal{L}^2_q$ and $\chi \in \mathcal{L}^2_{n-q}$ satisfy equation (50), together with the additional requirement $\varphi \in \mathcal{H}^2_q$, if and only if*

$$\begin{bmatrix} \varphi \\ \chi \end{bmatrix} = M \begin{bmatrix} u \\ v \end{bmatrix} \tag{54}$$

for some $u \in \mathcal{H}^2_r$ and $v \in \mathcal{L}^2_{k-r}$.

(ii) *$\varphi = 0$ if and only if $u = 0$.*

(iii) *$H_o \tilde{C}^* M$ has full column rank k.*

We have shown that for the distinguishing spectral density Φ with $\varphi \in \mathcal{H}^2_q$ we can find $u \in \mathcal{H}^2_r$, and $v \in \mathcal{L}^2_{k-r}$, such that (54) holds. Note that the v part disappears when P_1 has full column rank $r = k$. To continue our construction we consider as separate subcases the two types of φ we introduced earlier.

Subcase A1. Suppose first that $\varphi \neq 0$ and is co-outer. Since Φ is distinguishing we have by Lemma 4 that $H_o \phi \perp \mathcal{B}\mathcal{H}^2_p$, or equivalently

$$\Psi B^* H_o \tilde{C}^* M \begin{bmatrix} u \\ v \end{bmatrix} \perp \Psi \mathcal{H}^2_p, \tag{55}$$

where $\Psi \in \mathcal{H}^{\infty}_{p \times p}$ is a rational square inner matrix such that $\Psi B^* H_o \tilde{C}^* M \in \mathcal{R}\mathcal{H}^{\infty}_{p \times k}$. In conformance with u and v we introduce a partition $\Psi B^* H_o \tilde{C}^* M = [K \quad A]$, where K has r columns and A has $k - r$ columns. By applying, for instance, the Hermite form to the numerator of a polynomial left coprime factorization of A, we may write $A = [R \quad 0]V$, where R is a rational matrix of full column rank, say s, and V is a unimodular polynomial matrix. Devine $v_1 \in \mathcal{L}^2_s$ and $v_2 \in \mathcal{L}^2_{k-r-s}$ by

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = Vv. \tag{56}$$

Consequently, (55) implies

$$Ku + Rv_1 \perp \Psi \mathcal{H}^2_p. \tag{57}$$

Let $R = LD^{-1}$ be a polynomial coprime factorization of R. Then $\mathcal{R}\mathcal{L}^2_s = L\mathcal{L}^2_s$, with L also having rank s. Thus (57) implies

$$Ku + L\hat{v}_1 \perp \Psi \mathcal{H}^2_p, \tag{58}$$

where

$$\hat{v}_1 = D^{-1} v_1. \tag{59}$$

Denote by P_{H^2} the orthogonal projection of $\mathcal{L}^2_p$ onto $\mathcal{H}^2_p$, and let $w = P_{H^2} z^d \hat{v}_1$. Suppose that the degree of L is d. By considering the Fourier series of $\hat{v}_1$, it is easy to see that

$L\hat{v}_1 - Lz^{-d}w$ has nonvanishing Fourier coefficients only for negative indices. Thus $L\hat{v}_1 - Lz^{-d}w$ is orthogonal to the space $\mathcal{H}_p^2$ and hence also to its subspace $\Psi\mathcal{H}_p^2$. Consequently, (58) implies

$$Ku + Lz^{-d}w \perp \Psi\mathcal{H}_p^2. \tag{60}$$

Written in a slightly different form this gives

$$[z^d K \quad L] \begin{bmatrix} u \\ w \end{bmatrix} \in \mathcal{H}_p^2 \ominus z^d \Psi \mathcal{H}_p^2. \tag{61}$$

Thus the existence of a distinguishing Φ with $\varphi \neq 0$ co-outer in $\mathcal{H}_q^2$ implies that we can construct the matrices K and L as just described, and find a $u \in \mathcal{H}_r^2$ not equal to zero, and a $w \in \mathcal{H}_s^2$, such that (61) holds.

Conversely, suppose that we can form the matrices K and L, and that (61) holds for some $u \in \mathcal{H}_r^2$, not equal to zero, and some $w \in \mathcal{H}_s^2$. Define $\hat{v}_1 = z^{-1}w \in \mathcal{L}_s^2$. Then it is easily seen that (58) holds, also. Furthermore, with $v_1 = D\hat{v}_1$ we obtain (57). Finally with

$$v = V^{-1} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = V^{-1} \begin{bmatrix} z^{-1}Dw \\ v_2 \end{bmatrix}, \tag{62}$$

where $v_2 \in \mathcal{L}_{k-r-s}^2$ is arbitrary, we find that (55) holds. Write

$$\psi = M \begin{bmatrix} u \\ v \end{bmatrix}. \tag{63}$$

By inner-outer factorization we have

$$\psi = \begin{bmatrix} \varphi \\ \chi \end{bmatrix} \theta,$$

where φ is co-outer and θ is co-inner. Hence, with

$$\phi = \tilde{C}^* \begin{bmatrix} \varphi \\ \chi \end{bmatrix}$$

condition (55) implies that $H_o\phi \perp B\mathcal{H}_p^2$. Consequently, with

$$\Phi = \tilde{C}^* \psi\psi^* \tilde{C} = \phi\phi^* \tag{64}$$

Lemma 4 shows that $H_o \in \mathcal{A}(\Sigma; \Phi)$. Hence, the first condition (11) of Lemma 1 is satisfied. Clearly the choice of M also guarantees that the second condition (12) of Lemma 1 holds. Thus Φ is a distinguishing spectral density.

To summarize subcase A1: We have shown that a distinguishing spectral density Φ with $\varphi \in \mathcal{H}_q^2$ co-outer exists if and only if the matrices K and L can be formed and there exist a nonzero $u \in \mathcal{H}_r^2$ and a $w \in \mathcal{H}_s^2$ such that (61) holds. Moreover, via (62), (63), and (64), a Φ may be computed from u and w. Thus it only remains to find u and w. Note, however, that (61) is a relation of the type occurring in (44). We may therefore proceed exactly as in the two-block case to solve (61). As in the two-lock case it then follows that a solution u, w may be taken rational. Consequently, the corresponding spectral density may also be chosen rational.

Subcase A2. Consider now the case $\varphi = 0$. We still assume that $P_1 \neq 0$. Since $\varphi = P_1 w = 0$ for $w \neq 0$, P_1 must have less than full column rank. We can still find $M, u,$

and v as in Lemma 7. However, we must have $u = 0$. Conversely, suppose P_1 has less than full column rank. Take $u = 0$ and an arbitrary v, and define φ and χ by equation (54), and Φ by (49). Then $\varphi = 0$, and hence $C\phi = 0$. But this implies $\langle \Phi H_o^* BQC \rangle = 0$ for all $Q \in \mathcal{H}^\infty_{p\times q}$. By Lemma 3 this means that condition (11) of Lemma 1 holds for Φ. On the other hand, condition (12) holds trivially by the choice of M. Consequently, Φ is distinguishing.

To summarize subcase A2: When $P_1 \neq 0$, a distinguishing spectral density with $\varphi = 0$, exists if and only if P_1 has less than full column rank. Moreover, a distinguishing Φ may be computed via (54) and (49) by taking $u = 0$ and v arbitrary. In particular, when v is chosen rational then Φ will also be rational.

Case B. Suppose $P_1 = 0$. Take an arbitrary $w \in \mathcal{L}^2_k$. Define φ and χ by (52), and ϕ and Φ by (49). Clearly $\varphi = 0$, and hence $C\phi = 0$. But this implies that $\langle \Phi H_o^* BQC \rangle = 0$ for all $Q \in \mathcal{H}^\infty_{p\times q}$. By Lemma 3 this means that condition (11) of Lemma 1 holds for Φ. On the other hand, condition (12) holds trivially by the choice of P. Consequently, Φ is distinguishing. It will be rational if w is chosen rational.

This completes the construction of a distinguishing spectral density for the four-block case. We remark that subcases A1 and A2 do not completely exclude each other. It is easy to construct "diagonal" examples with two distinguishing spectral densities, one having $\varphi \neq 0$ and the other have $\varphi = 0$.

7. Existence of the distinguishing spectral density in terms of a four-block operator. In [18] it is shown that a spectral density Φ of rank one is distinguishing for the one-block problem if and only if $\Phi = \phi\phi^*$ with ϕ a maximizing vector of a Sarason type operator associated with the one-block problem. In this section we give a similar condition for the four-block problem. Without loss of generality, we may consider the following special case of (4). Let the set of functions over which $\mathcal{H}^\infty$ optimization takes place be given by

$$\Sigma = \begin{bmatrix} F_{11} + M\mathcal{H}^\infty_{p\times q} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}, \tag{65}$$

where $F_{11} \in \mathcal{RH}^\infty_{p\times q}$, $F_{12} \in \mathcal{RH}^\infty_{p\times n-q}$, $F_{21} \in \mathcal{RH}^\infty_{m-p\times q}$, $F_{22} \in \mathcal{RH}^\infty_{m-p\times n-q}$, and M is an inner matrix in $\mathcal{RH}^\infty_{p\times p}$. Let $\mathcal{H}(M) = \mathcal{H}^2_p \ominus M\mathcal{H}^2_p$ and $\mathcal{L}(M) = \mathcal{L}^2_p \ominus M\mathcal{H}^2_p$. Denote by $P_{\mathcal{H}(M)}$ the orthogonal projection from $\mathcal{L}^2_p$ unto $\mathcal{H}(M)$, and by $P_{\mathcal{L}(M)}$ the orthogonal projection from $\mathcal{L}^2_p$ unto $\mathcal{L}(M)$. Moreover, for any matrix $V \in \mathcal{H}^\infty_{m\times n}$ we denote by $V(S)$ the operator $V(S) : \mathcal{H}^2_n \to \mathcal{H}^2_m$ of multiplying by V. Similarly, for any $W \in \mathcal{L}^\infty_{m\times n}$ we denote by $W(U)$ the operator $W(U) : \mathcal{L}^2_n \to \mathcal{L}^2_m$ of multiplying by W. Following [2], [4], and [20], we associate with (65) the four-block operator

$$A = \begin{bmatrix} P_{\mathcal{H}(M)}F_{11}(S) & P_{\mathcal{L}(M)}F_{12}(U) \\ F_{21}(S) & F_{22}(U) \end{bmatrix} \tag{66}$$

from $\mathcal{H}^2_q \oplus \mathcal{L}^2_{n-q}$ to $\mathcal{L}(M) \oplus \mathcal{L}^2_{m-p}$, the norm of which is given by

$$\|A\| = \lambda_o := \inf\{\|H\|_\infty : H \in \Sigma\}.$$

A vector $\phi \in \mathcal{H}^2_q \oplus \mathcal{L}^2_{n-q}$ is said to be *maximizing* for A if $\|A\phi\| = \|A\|\|\phi\|$.

The following theorem provides a necessary and sufficient condition for the existence of a distinguishing spectral density.

THEOREM 4. *Consider the four-block problem* Σ *in* (65) *and a spectral density* $\Phi = \phi\phi^*$ *with* $\phi \in \mathcal{L}^2_n$ *taken as in* (24). *Then* Φ *is distinguishing for* $\mathcal{S}(\Sigma)$ *if and only if* ϕ *is a maximizing vector of the four-block operator* A.

Proof. For a proof see [19].

A sufficient, but not necessary, condition for the existence of a maximizing vector is that the so-called *essential norm* of A is strictly less than $\|A\|$[4], [20]. Moreover, in [20] it is shown that the essential norm of A is given by

$$\max\left\{\|[F_{21} \quad F_{22}]\|_\infty, \left\|\begin{bmatrix} F_{12} \\ F_{22} \end{bmatrix}\right\|_\infty\right\}. \tag{67}$$

Hence, if (67) is strictly less than λ_o then the four-block problem admits a distinguishing spectral density.

8. Example. We illustrate the super-optimization method in §5 with a numerical four-block example given by the coset

$$\Sigma_1 = F + B\mathcal{H}^\infty_{2\times 2}C, \tag{68}$$

where

$$F(z) = \begin{bmatrix} \frac{(4(\sqrt{2}+1)z^2-12(\sqrt{2}+1)z-9}{3\sqrt{6}(z-3)} & -\frac{8(\sqrt{2}+1)z^2-24(\sqrt{2}+1)z+9}{6\sqrt{3}(z-3)} & \frac{\sqrt{3}z(3\sqrt{2}+16-8z)}{3\sqrt{6}(z-2)} \\ \frac{(2(4\sqrt{2}z^2-12\sqrt{2}z+9}{6\sqrt{3}(z-3)} & -\frac{8\sqrt{2}z^2-24\sqrt{2}z-9}{3\sqrt{6}(z-3)} & \frac{2z}{\sqrt{2}(z-2)} \\ \frac{(4(\sqrt{2}-1)z^2-12(\sqrt{2}z-1)-9}{3\sqrt{6}(z-3)} & -\frac{8(\sqrt{2}-1)z^2-24(\sqrt{2}-1)z+9}{6\sqrt{3}(z-3)} & \frac{\sqrt{3}z(3\sqrt{2}-16+8z)}{3\sqrt{6}(z-2)} \end{bmatrix},$$

$$B(z) = \begin{bmatrix} -\frac{5z^2-14z+5}{2\sqrt{2}(z^2-5z+6)} & \frac{z^2-1}{2\sqrt{2}(z^2-5z+6)} \\ \frac{z^2-1}{2(z^2-5z+6)} & -\frac{5z^2-14z+5}{2(z^2-5z+6)} \\ -\frac{5z^2-14z+5}{2\sqrt{2}(z^2-5z+6)} & \frac{z^2-1}{2(z^2-5z+6)} \end{bmatrix},$$

$$C(z) = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & -\frac{z}{\sqrt{2}} \\ -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} & -\frac{z}{\sqrt{2}} \end{bmatrix}.$$

B and C are both inner matrices.

Consider first the problem of finding a $F + BQC$ of minimal $\mathcal{H}^\infty$-norm in the coset (68). Invoke the polynomial approach of §1 to determine a Q_{o1} such that $H_{o1} = F + BQ_{o1}C$ is $\mathcal{H}^\infty$-optimal in Σ_1. One solution is given by

$$Q_{o1}(z) = \begin{bmatrix} -\frac{5(38z+21)}{6(49z+69)} & -\frac{2z+57}{2(49z+69)} \\ -\frac{2z+57}{2(49z+69)} & -\frac{5(38z+21)}{6(49z+69)} \end{bmatrix}.$$

The minimal norm is $\lambda_1 = \|H_{o1}\|_\infty = 4$.

To obtain the model-matching problem for the second largest singular value we first need a ϕ_1 such that $\Phi_1 = \phi_1\phi_1^*$ is a distinguishing spectral density for $\mathcal{S}(\Sigma_1)$. Using the method of §6 we obtain

$$\phi_1(z) = \begin{bmatrix} \frac{1}{\sqrt{3}(2z-1)} \\ -\frac{2}{\sqrt{6}(2z-1)} \\ -\frac{1}{2z-1} \end{bmatrix}.$$

The inner factor V_{1a} of ϕ_1 and a corresponding inner complement V_{1b} are given by

$$V_{1a}(z) = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}, \qquad V_{1b}(z) = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

As a result, $C\phi \neq 0$ and $\phi^* H_{o1}^* B = 0$. Thus we are in situation (iii) of Theorem 1. The matrix P_1 required to reduce the parameter space from $\mathcal{H}_{2\times 2}^\infty$ to $\mathcal{H}_{2\times 1}^\infty$ is, as in Lemma 6, obtained by parametrizing the set of all solutions $Q \in \mathcal{H}_{2\times 2}^\infty$ of the equation $QC\phi = 0$. This gives $P_1 = [1 \quad -1]$. The model-matching problem for the second largest singular value then becomes

$$\Sigma_2 = F_2 + B\mathcal{H}_{2\times 1}^\infty C_2,$$

where

$$F_2(z) = H_{o1}(z)V_{1b}(z) = \begin{bmatrix} -\frac{2(\sqrt{2}-1)z}{3\sqrt{2}} & \frac{2(69z+49)}{3(49z+69)} \\ \frac{2z}{3} & -\frac{4(69z+49)}{3\sqrt{2}(49z+69)} \\ \frac{2(\sqrt{2}+1)z}{3\sqrt{2}} & \frac{2(69z+49)}{3(49z+69)} \end{bmatrix},$$

$$C_2(z) = P_1(z)C(z)V_{1b}(z) = [0 \quad 1].$$

Invoking the polynomial approach once more we obtain the $\mathcal{H}^\infty$-optimal solution $H_{o2} = F_2 + B_2 Q_{o2} C_2$ determined by

$$Q_{o2}(z) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The minimal norm is $\lambda_2 = \|H_{o2}\|_\infty = 4/3$.

To obtain the model-matching problem for the smallest singular value, we construct a ϕ_2 such that $\Phi_2 = \phi_2 \phi_2^*$ is distinguishing spectral density for $\mathcal{S}(\Sigma_2)$. We may choose

$$\phi_2(z) = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

which is inner already; that is, $V_{2a} = \phi_2$. An inner complement is given by

$$V_{2b}(z) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Since $C_2\phi = 0$, we are in situation (i) of Theorem 1. Thus, no reduction of the parameter space $\mathcal{H}_{2\times 1}^\infty$ of Σ_2 takes place. Consequently, the third problem is

$$\Sigma_3 = F_3 + B\mathcal{H}_{2\times 1}^\infty C_3,$$

where

$$F_3(z) = H_{o2}(z)V_{2b}(z) = \begin{bmatrix} \frac{2(69z+49)}{3(49z+69)} \\ -\frac{4(69z+49)}{3\sqrt{2}(49z+69)} \\ \frac{2(69z+49)}{3(49z+69)} \end{bmatrix}, \qquad C_3(z) = C_2(z)V_{2b}(z) = 1.$$

Using the polynomial approach we obtain an $\mathcal{H}^\infty$-optimal solution $H_{o3} = F_3 + BQ_{o3}C_3$ determined by

$$Q_{o3}(z) = \begin{bmatrix} \dfrac{295(z-3)}{24\sqrt{2}(49z+69)} \\ -\dfrac{295(z-3)}{24\sqrt{2}(49z+69)} \end{bmatrix}.$$

The minimal norm is $\lambda_3 = \|H_{o3}\|_\infty = 9/8$. A corresponding distinguishing spectral density $\Phi_3 = \phi_3\phi_3^*$ is given by $\phi_3(z) = 1/(3z-1)$. Since Σ_3 consists of 3×1 matrix valued functions and admits a distinguishing spectral density, the solution H_{o3} is, in fact, by Theorem 3, unique.

By recovering the corresponding solution, first in Σ_2 and then in Σ_1, we obtain a unique super-optimal solution $H_{\text{sup}} = F + BQ_{\text{sup}}C$ of the original problem Σ_1, where

$$Q_{\text{sup}}(z) = -\begin{bmatrix} 25/48 & 7/48 \\ 7/48 & 25/48 \end{bmatrix},$$

and

$$H_{\text{sup}}(z) = \begin{bmatrix} \dfrac{32(\sqrt{2}+1)z+27}{24\sqrt{6}} & -\dfrac{64(\sqrt{2}+1)z-27}{48\sqrt{3}} & -\dfrac{4\sqrt{3}(2\sqrt{2}+1)z}{6\sqrt{3}} \\ \dfrac{32\sqrt{2}z-27}{24\sqrt{3}} & -\dfrac{64\sqrt{2}z+27}{24\sqrt{6}} & -\dfrac{4\sqrt{3}z}{3\sqrt{6}} \\ \dfrac{32(\sqrt{2}-1)z+27}{24\sqrt{6}} & -\dfrac{64(\sqrt{2}-1)z-27}{48\sqrt{3}} & \dfrac{4\sqrt{3}(2\sqrt{2}-1)z}{6\sqrt{3}} \end{bmatrix},$$

with singular values

$$\begin{aligned}(s_1^\infty(H_{\text{sup}}), s_2^\infty(H_{\text{sup}}), s_3^\infty(H_{\text{sup}})) &= (s_1\{H_{\text{sup}}(e^{i\theta})\}, s_2\{H_{\text{sup}}(e^{i\theta})\}, s_3(H_{\text{sup}}(e^{i\theta})\}) \\ &= (4, 4/3, 9/8)\end{aligned}$$

for all points $e^{i\theta}$ on the unit circle. This completes the construction of the super-optimal solution.

9. Conclusions. A super-optimization algorithm for the general four-block (standard problem) in $\mathcal{H}^\infty$-optimal control has been presented. It successively reduces the original problem to smaller super-optimizations problems. Each step amounts to an ordinary $\mathcal{H}^\infty$-optimization of the largest remaining singular value, and a subsequent removal of the optimized part. The central part of the paper concerns the derivation of a removal technique based on the spectral density of the equalizer principle. The $\mathcal{H}^\infty$-optimizations may be done with any method producing strictly optimal solutions. A method that readily applies, and motivated much of this work, is the polynomial approach by Kwakernaak.

The representation of the kth $\mathcal{H}^\infty$-optimization problem Σ_k in the super-optimization algorithm is not unique. Although any solution H_k of the kth optimization problem of the super-optimization algorithm, and any inner complement V_{kb} of the inner part of a corresponding distinguishing spectral density, may be used to generate data for the following optimization problem, from a numerical point of view it is desirable to obtain data of reasonably low McMillan degree. Further investigation is needed, however, in order to find H_k and V_{kb} meeting such degree constraints.

Appendix A. Let K be a rational inner matrix in $\mathcal{H}^\infty_{m\times m}$. We wish to obtain a basis for the finite-dimensional space $\mathcal{H}^2_m \ominus K\mathcal{H}^2_m$. An easily computable basis is described in [7]. For completeness we include the construction.

Write $K = PQ^{-1}$, where P and Q are coprime polynomial matrices. Then $h \in \mathcal{H}^2_m$ implies $Q^{-1}h \in \mathcal{H}^2_m$, which in turn gives $K\mathcal{H}^2_m \subseteq P\mathcal{H}^2_m$. Conversely, we have $P\mathcal{H}^2_m \subseteq K\mathcal{H}^2_m$. Thus, $\mathcal{H}^2_m \ominus K\mathcal{H}^2_m = \mathcal{H}^2_m \ominus P\mathcal{H}^2_m$. To find a basis for $\mathcal{H}^2_m \ominus P\mathcal{H}^2_m$, we will, without loss of generality, assume that P is column reduced. This can always be achieved by multiplying P (and Q) from the right by a suitable unimodular matrix. Since K is inner, this also holds for $\det(K)$. Thus, the zeros of $\det(K)$ must be in the open unit disk $\mathbb{D}$. This, in turn, implies that the zeros of $\det(P)$ are in $\mathbb{D}$, also. Let d_j be the highest power of z that occurs in the jth column of $P(z)$. Write

$$P(z) = (P_0 + P_1 z^{-1} + \cdots + P_q z^{-q})D(z) \tag{69}$$

where

$$D(z) = \operatorname{diag}(z^{d_1}, z^{d_2}, \ldots, z^{d_m}),$$

and $P_0, \ldots, P_q$ are constant $m \times m$ matrices with q the highest power of z in $P(z)$. Since P is column reduced, P_0 must be nonsingular. Define

$$\hat{P}(z) = P_0^* + P_1^* z + \cdots + P_q^* z^q.$$

Clearly $\hat{P} \in \mathcal{H}^\infty_{m\times m}$. To see also that $\hat{P}^{-1} \in H^\infty_{m\times m}$ we note, using (69), that

$$\det(\hat{P}(z)) = z^{d_1+d_2+\cdots+d_m}\det(P(z)^*) = z^{d_1+d_2+\cdots+d_m}\overline{\det(P(z))}.$$

This means that $\det(\hat{P}(z))$ cannot have any zero in the closed unit disk with a possible exception at $z = 0$. But owing to the column reduceness $\det(\hat{P}(0)) = \det(P_0^*) \neq 0$. Consequently, $\hat{P}^{-1} \in \mathcal{H}^\infty_{m\times m}$, as required. Moreover, $\hat{P}^{-1}$ determines a linear bijection on $\mathcal{H}^\infty_{m\times m}$. From (69) we also get

$$P(e^{i\theta}) = (P_o + P_1 e^{-i\theta} + \cdots + P_q e^{-iq\theta})Q(e^{-i\theta}) = \hat{P}(e^{i\theta})^* Q(e^{-i\theta}),$$

which implies that $f \in \mathcal{H}^2_m$ is perpendicular to $P\mathcal{H}^2_m$ if and only if $\hat{P}f$ is perpendicular to $Q\mathcal{H}^2_m$. Moreover, using the fact that $\hat{P}^{-1} \in H^\infty_{m\times m}$, this yields

$$\mathcal{H}^2_m \ominus P\mathcal{H}^2_m = \hat{P}^{-1}(\mathcal{H}^2_m \ominus D\mathcal{H}^2_m).$$

Since $\hat{P}^{-1}$ is a linear bijection on $\mathcal{H}^2_m$, any basis $\{b_1, b_2, \ldots\}$ of $\mathcal{H}^2_m \ominus D\mathcal{H}^2_m$ gives rise to a basis $\{\hat{P}^{-1}b_1, \hat{P}^{-1}b_2, \ldots\}$ of $\hat{P}^{-1}(\mathcal{H}^2_m \ominus D\mathcal{H}^2_m)$. However, $\mathcal{H}^2_m \ominus D\mathcal{H}^2_m$ consists of all polynomial vectors

$$v = (v^1, v^2, \ldots, v^m)^T,$$

where v^j is a polynomial of degree strictly less than d_j. A natural basis in $\mathcal{H}^2_m \ominus D\mathcal{H}^2_m$ is therefore given by the vectors

$$e_i^j(z) = (0, \ldots, 0, z^j, 0, \ldots, 0)^T, \quad i = 1, 2, \ldots, m, \quad j = 0, 1, \ldots, \deg d_i - 1,$$

where z^j occurs in the ith component. Consequently,

$$\{\hat{P}^{-1}e_i^j | i = 1, 2, \ldots, m, j = 0, 1, \ldots, \deg d_i\}$$

is a basis for $\mathcal{H}^2_m \ominus P\mathcal{H}^2_m$. In [7] this is called the *expedient basis*.

Appendix B. We show how to solve an equation of the type (39) occurring in computation of a distinguishing spectral density. We do this in the form of a lemma.

LEMMA 8. *Let A and B be rational matrices of dimensions $m \times k$ and $m \times p$, respectively, where A has full column rank. Suppose the equation*

$$Au - B\xi \tag{70}$$

has a nontrivial solution $u \in \mathcal{H}_k^2$, $(u \in \mathcal{L}_k^2), \xi \in \mathbb{C}^p$. Then there exists a complex matrix L of, say, dimensions $p \times r$, having full column rank, and a matrix $K \in \mathcal{RH}_{k\times r}^\infty, (K \in \mathcal{RL}_{k\times r}^\infty)$ such that $u \in \mathcal{RH}_k^2 (u \in \mathcal{RL}_k^2)$ and $\xi \in \mathbb{C}^p$ solves (70) *if and only if*

$$\begin{bmatrix} u \\ \xi \end{bmatrix} = \begin{bmatrix} K \\ L \end{bmatrix} \zeta \tag{71}$$

for some $\zeta \in \mathbb{C}^r$. Moreover, the linear mapping $\zeta \to (u^T, \xi^T)^T$ determined by (71) *is injective.*

Proof. Consider first the case where $u \in \mathcal{H}_k^2$. Apply, for instance, the Hermite form to the numerator of a polynomial right coprime factorization of A to write

$$A = U \begin{bmatrix} D \\ 0 \end{bmatrix}, \tag{72}$$

where U is a unimodular polynomial matrix, and D is a square rational matrix of, say, dimensions $k \times k$ having full rank. Let

$$U^{-1} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix},$$

where W_1 has k rows and W_2 has $m - k$ rows. With this, (70) takes the equivalent form

$$u = D^{-1} W_1 B\xi, \tag{73}$$

$$0 = W_2 B\xi. \tag{74}$$

We treat equation (73) first. Use polynomial coprime factorization to write $D^{-1}W_1B = \Gamma^{-1}\Delta$. Factorize $\det(\Gamma) = d_1 d_2$, where d_1 has all its roots in closed unit disk $\overline{\mathbb{D}}$ and d_2 has all its roots outside $\overline{\mathbb{D}}$. Then $\Gamma^{-1} = d_1^{-1} d_2^{-1} T$ for some polynomial matrix T. Clearly, a necessary and sufficient condition for $D^{-1}W_1B\xi \in \mathcal{H}_t^2$ is, therefore, that each component of the polynomial vector

$$T\Delta\xi = \det(\Gamma) D^{-1} W_1 B\xi,$$

cancels all the roots of d_1 (counting multiplicities). Let $M = \det(\Gamma) D^{-1} W_1 B$, and let $z_1, \ldots, z_r$ be the roots of d_1, with multiplicities $m_1, \ldots, m_r$, respectively. Thus, solving (73) is equivalent to finding a $\xi \in \mathbb{C}^p$ satisfying the following homogeneous system of linear equations:

$$M^{(j)}(z_i)\xi = 0, \qquad 1 \le i \le r,\ 0 \le j < m_i, \tag{75}$$

where $M^{(j)}$ means the jth derivative of M (with $M^{(0)} = M$).

To include the second equation (74) we take a left polynomial coprime factorization $F^{-1}N$ of W_2B, and let d be the highest degree of any element in the numerator N. Then

$$N(z) = N_0 + N_1 z + \cdots + N_d z^d,$$

where $N_k, 0 \leq k \leq d$ are ordinary complex matrices. Consequently, (74) holds if and only if

$$N_k \xi = 0, \qquad k = 0, 1, \ldots, d. \tag{76}$$

Thus the problem of solving the original equation (70) has been reduced to solving simultaneously the set of linear numerical equations in (75) and (76). If a solution ξ exists, then $u = D^{-1} W_1 B \xi$, together with ξ, constitutes a solution of the original equation.

Suppose the null space of the system of equations given by (75) and (76) has dimension k. Let L be a $p \times r$ complex matrix whose columns form a basis of this null space. Then ξ solves (75) and (76) if and only if

$$\xi = L\zeta$$

for some $\zeta \in \mathbb{C}^r$. Consequently, with $K = D^{-1} W_1 B L$, we find that u belongs to $\mathcal{H}_k^2$ and satisfies (70) for some $\xi \in \mathbb{C}^p$ if and only if

$$u = K\zeta \quad and \quad \xi = L\zeta$$

for some $\zeta \in \mathbb{C}^r$.

In the case where $u \in \mathcal{RL}_k^2$, the proof is completely analogous, except that we only need to solve (74) and then define u by (73). The last statement follows directly from the fact that L has full column rank. In fact, even the mapping $\zeta \to \xi$ is injective.

Acknowledgments. I would like to thank Professor Huibert Kwakernaak for many valuable discussions concerning this work. I am indebted to him for the idea of using the conditions of Lemma 6 to generate the matrices R and P reducing the dimension of the parameter. (An efficient MATLAB macro for computing them is available in [14].) I am also grateful to the referees for suggestions that improved the final version of the manuscript.

REFERENCES

[1] P. Boekhoudt, *The $\mathcal{H}^\infty$ control design method: A polynomial approach*, Ph.d. thesis, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 1988.

[2] A. Feintuch and B. A. Francis, *Uniformly optimal control of linear feedback systems*, Automatica, 21 (1986), pp. 563–574.

[3] Y. K. Foo and I. Postlethwaite, *All solutions, all-pass form solutions, and the "best" solutions to an $\mathcal{H}^\infty$ optimization problem in robust control*, Systems Control Lett., 7 (1986), pp. 261–268.

[4] C. Foias and A. Tannenbaum, *On the four block problem*, II: *The singular system*, Integral Equations Operator Theory, 11 (1988), pp. 726–767.

[5] C. Foias and A. E. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Birkhäuser Verlag, Basel, 1990.

[6] B. A. Francis, *A Course in $\mathcal{H}^\infty$ Control Theory*, Lecture Notes in Control and Information Sci., 88, Springer-Verlag, Berlin, New York, 1987.

[7] K. D. Gregson and N. J. Young, *Finite representations of block Hankel operators and balanced realizations*, Oper. Theory: Adv. Appl., 35 (1988), pp. 441–480.

[8] D. W. Gu, M. C. Tsai, and I. Postlethwaite, *An algorithm for super-optimal $\mathcal{H}^\infty$ design: The two-block case*, Automatica, 25 (1989), pp. 215–221.

[9] I. Jaimoukha and D. J. N. Limebeer, *State-space algorithm for the solution of the 2-block super-optimal distance problem*, preprint, Imperial College, London, 1991.

[10] H. Kwakernaak, *Minimax frequency domain performance and robustness optimization of linear feedback Systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 994–1004.

[11] ———, *A polynomial approach to minimax frequency domain optimization of multivariable feedback systems*, Internat. J. Control, 44 (1986), pp. 117–156.

[12] ———, *Progress in the polynomial solution of the standard $\mathcal{H}^\infty$-optimal control problem*, Memorandum No. 817, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 1989.

[13] ———, *The polynomial approach to $\mathcal{H}^\infty$-optimal regulation*, in $\mathcal{H}^\infty$-Control Theory, Lecture Notes in Mathematics, 1496, E. Mosca and L. Pandolfi, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[14] ———, *MATLAB macros for polynomial $\mathcal{H}^\infty$ control system optimization*, Memorandum No. 881, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 1990.

[15] H. KWAKERNAAK AND P.-O. NYMAN, *An equalizing approach to super-optimization*, Proc. 1989 IEEE International Conference on Control and Applications, Jerusalem, 1989.

[16] D. J. N. LIMEBEER, G. D. HALIKIAS, AND K. GLOVER, *State-space algorithm for the computation of superoptimal matrix interpolation functions*, Internat. J. Control, 50 (1989), pp. 2431–2466.

[17] N. K. NIKOL'SKIĬ, *Treatise on the Shift Operator*, Springer-Verlag, Berlin, New York, 1986.

[18] P. O. NYMAN, *Equalizer principle of $\mathcal{H}^\infty$ optimal control and its application to super-optimization*, Internat. J. Control, 54 (1991), pp. 393–415.

[19] ———, *Super-optimization of the four-block problem in $\mathcal{H}^\infty$-optimal control*, Ph.d. thesis, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, June, 1992.

[20] H. ÖZBAY AND A. TANNENBAUM, *A skew Toeplitz approach to the $\mathcal{H}^\infty$ optimal control of multivariable distributed systems*, SIAM J. Control Optim., 28 (1990), pp. 653–670.

[21] I. POSTLETHWAITE, M. C. TSAI, AND D.-W. GU, *A state-space approach to discrete-time super-optimal $\mathcal{H}^\infty$ control problems*, Internat. J. Control, 49 (1989), pp. 247–268.

[22] M. ROSENBLUM AND J. ROVNYAK, *Hardy Classes and Operator Theory*, Oxford University Press, New York, 1985.

[23] D. SARASON, *Generalized interpolation in $\mathcal{H}^\infty$*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.

[24] M. C. TSAI, D.-W. GU, AND I. POSTLETHWAITE, *A state-space approach to super-optimal $\mathcal{H}^\infty$ control problems*, IEEE Trans. Automat. Control, 33 (1988), pp. 833–843.

[25] M. VIDYASAGAR, *Control Systems Synthesis*, MIT Press, Cambridge, 1985.

[26] F.-B. YEH AND T.-S. HWANG, *A computational algorithm for the super-optimal solution of the model matching problem*, Systems Control Lett., 11 (1988), pp. 203–211.

[27] N. J. YOUNG, *The Nevanlinna–Pick problem for matrix-valued functions*, J. Operator Theory, 15 (1986), pp. 239–265.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 116–139, January 1994

007

RATES OF CONVERGENCE FOR AN ADAPTIVE FILTERING ALGORITHM DRIVEN BY STATIONARY DEPENDENT DATA*

ANDREW HEUNIS†

Abstract. Eweda and Macchi [IEEE *Trans. Automat. Control*, 29 (1984), pp. 119–127] and Watanabe [IEEE *Trans. Inform. Theory*, 30 (1984), pp. 134–140] show that the sequence of random vectors generated by a stochastic gradient adaptive filtering algorithm converges almost surely and in L_p (for p an even integer) to the solution of the associated Wiener–Hopf equation when the driving data process is stationary and weakly dependent. Under strong (i.e., Rosenblatt or α) and ψ-mixing conditions, together with various moment bounds on the driving data process, an almost sure functional invariance principle is obtained that approximates the sample paths of the random process generated by the stochastic gradient algorithm with the sample paths of a particular Gauss–Markov process. Almost sure rates of convergence in the form of laws of the iterated logarithm follow from the functional invariance principle. As a byproduct a functional central limit theorem is also obtained for a sequence of processes derived by suitably scaling the sequence of iterations generated by the algorithm.

Key words. adaptive filtering, strong invariance principle, law of the iterated logarithm, Lyapunov equation, strong and ψ-mixing processes, almost sure rates of convergence

AMS subject classifications. 60F15, 60F17, 93E10

1. Introduction.

Assume that $\{X_j, j \geq 1\}$ and $\{a_j, j \geq 1\}$ are given jointly stationary random processes defined on a common probability space, X_j being R^N-valued and a_j being real-valued, realizations of which can be observed but whose statistics are unknown. A basic problem in adaptive filtering is to use the single realizations $\{X_j(\omega), j \geq 1\}$ and $\{a_j(\omega), j \geq 1\}$ to compute a vector h_* in R^N that minimizes the function

$$h \to E(\langle h, X_j \rangle - a_j)^2$$

(where $\langle x, y \rangle$ denotes the inner product of vectors x and y in R^N), or, equivalently, solves the Wiener–Hopf equation

$$(*) \qquad E(X_j X_j^T)h = E(a_j X_j),$$

where X_j is regarded as a column vector. Problems of this kind arise in many contexts such as linear classification, adaptive equalization, and ARMA modeling (see Eweda and Macchi [1, Exs. 1 and 2, p. 120], and Widrow and Stearns [2, Chap. 6]), and one of the main characteristics of such problems is that successive elements of the vector-valued process $\{X_j\}$ are usually strongly correlated.

One of the most successful and widely used algorithms for computing h_* on the basis of single realizations is a recursion of the form

$$(**) \qquad h_{j+1} = h_j + \mu_j(a_j - h_j^T X_j)X_j \qquad j \geq 1,$$

in which h_j is a column vector and $\{\mu_j\}$ is a sequence of nonrandom positive scalars satisfying the two conditions

$$\sum_{j=1}^{\infty} \mu_j = \infty \quad \text{and} \quad \sum_{j=1}^{\infty} \mu_j^2 < \infty.$$

In a very significant advance, Eweda and Macchi [1, §III] show that the sequence of random vector $\{h_j\}$ resulting from this algorithm converges almost surely to h_* when the

* Received by the editors July 30, 1991; accepted for publication (in revised form) September 29, 1992.
† Department of Electrical Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

data process $\{(a_j, X_j^T)\}$ that "drives" the algorithm satisfies rather unrestrictive moment bounds, as well as a very general condition of decaying dependence between the random vectors (a_j, X_j^T) and (a_k, X_k^T), when the indices j and k are widely separated. In a related development, Watanabe [3, Thms. 1–4] proves 2 rth mean convergence of $\{h_j\}$ to h_* for all natural numbers r under moment bounds and mixing conditions that are significantly, but seemingly unavoidably, stronger than the conditions postulated in [1]. The remarkable aspect of these results is that both establish convergence without resorting to projection onto a compact neighbourhood of h_*, as is often necessary when using the established theory of stochastic algorithms (a comprehensive account of which can be found in Benveniste, Metivier, and Priouret [4]).

In probability theory we find a most celebrated result, namely, the strong invariance principle of Strassen [5, Thm. 2], which asserts that the running sum of a given sequence of independent and identically distributed zero-mean second-order random variables (of any underlying distribution) can be approximated *almost surely* by the sample paths of a suitable Brownian motion. This has many ramifications, because it allows us to use our very detailed understanding of the sample properties of Brownian motion to obtain a precise characterization of the asymptotic behaviour of the given sum of random variables. Motivated by the invariance principle of Strassen [5], our goal in this note is to obtain an almost surely approximation of the "difference" process $\{h_j - h_*\}$ by some "standard" process whose sample-path properties are well understood. We show that the difference process can in fact be approximated almost surely by a Gauss–Markov process, which is itself a linear function of a suitable Brownian motion. To illustrate the applicability of this result, we use it to prove a law of the iterated logarithm (which in turn implies a precise and unimprovable almost sure rate of convergence of h_j to h_*) and a functional central limit theorem for $\{h_j - h_*\}$. Our proof of the invariance principle relies intrinsically on (a slight variant of) the 2 rth mean convergence results of Watanabe [3] mentioned above, and hence we work under conditions on the driving data $\{(a_j, X_j^T)\}$ that are very similar to those postulated in [3].

In §2 we state the main result, which is a strong invariance principle (Proposition 2.1), as well as the consequent law of the iterated logarithm (Proposition 2.2) and functional central limit theorem (Proposition 2.3); we also give the regularity conditions that will always be assumed. In §3 these propositions are established. Section 4 is an appendix in which various subsidiary results, needed for the proofs in §3, are stated and proved (these results can be referenced at will once the reader is familiar with the basic regularity conditions and notation specified in §2). Section 5 is another appendix in which several useful results from probability theory, perhaps not well known to readers, have been gathered for convenience.

2. Main results. Let $\{a_j, j \geq 1\}$ and $\{X_j, j \geq 1\}$ be random processes, real and R^N-valued, respectively, that are defined on a common probability space $(\Omega, \mathbf{A}, P)$. We study the asymptotic properties of R^N-valued random vectors h_j generated by the stochastic gradient algorithm

$$h_{j+1} = h_j + \frac{\mu}{j}(a_j - h_j^T X_j)X_j, \qquad j \geq 1, \tag{1}$$

where h_1 is an arbitrary fixed nonrandom R^N-vector, μ is a positive constant, and X_j and h_j are column vectors. Certain regularity conditions will be needed for this problem, but before stating these we recall the notion of a strong mixing process: Suppose that $\{\xi_n, n \geq 1\}$ is a process assuming values in R^D, and for each integer $m \geq 1$ define

$$\alpha_\xi(m) \triangleq \sup |P(AB) - P(A)P(B)|,$$

where the supremum is taken over all $n \geq 1$, all $A \in \sigma\{\xi_j, 1 \leq j \leq n\}$, and all $B \in \sigma\{\xi_j, j \geq n+m\}$. Then $\{\xi_n\}$ is *strong mixing* when $\alpha_\xi(m) \to 0$ as $m \to \infty$.

Returning to the regularity conditions, it will henceforth be assumed that the following hold:

(A) The R^{N+1}-valued process $\{(a_j, X_j^T), j \geq 1\}$ is strong mixing with $\alpha_{(a,X)}(m) = O(m^{-\beta})$ for some $\beta > 3$, and strictly stationary (i.e., the finite-dimensional distribution functions of $\{(a_j, X_j^T)\}$ are invariant under translation through the positive integers);

(B) The symmetric matrix $R \stackrel{\Delta}{=} E(X_j X_j^T)$ is positive definite and the constant μ in (1) is sufficiently large to ensure that $(2\mu R - I)$ is positive definite;

(C) There are positive constants C and ε such that $E|X_j|^{4k} = O(C^k k!)$ and $E|a_j|^{37+\varepsilon} < \infty$ for all integers $j, k \geq 1$.

Define $h_* \stackrel{\Delta}{=} R^{-1}E(a_j X_j)$ (the unique solution of the Wiener–Hopf equation (*)) and let $z_j \stackrel{\Delta}{=} a_j X_j - X_j X_j^T h_*$. Clearly, $\{z_j\}$ is strictly stationary and strong mixing with $\alpha_z(m) = O(m^{-\beta})$ for some $\beta > 3$, and it is easily seen from Holder's inequality and the moment bounds in (C) that z_j has finite third order moments. It then follows from Theorem 5.2(i) in §5 that the series in

$$\Gamma \stackrel{\Delta}{=} E(z_1 z_1^T) + \sum_{j=2}^{\infty} E(z_1 z_j^T) + \sum_{j=2}^{\infty} E(z_j z_1^T)$$

converge absolutely and, clearly, that Γ is symmetric positive semidefinite. The final regularity condition is:

(D) The matrix Γ is positive definite.

The moment conditions in (C) are essentially those required by Watanabe [3, Thm. 3] (restated as Theorem 5.4 in §5), which will play an essential role in later developments. These moment conditions can be substantially weakened if the sense in which $\{(a_j, X_j^T)\}$ is mixing is strengthened. Accordingly, we state alternatives to conditions (A) and (C), but first the notion of ψ-mixing is introduced: Suppose that $\{\xi_n, n \geq 1\}$ is some R^D-valued process, and for each integer $m \geq 1$ define

$$\psi_\xi(m) \stackrel{\Delta}{=} \sup \frac{|P(AB) - P(A)P(B)|}{P(A)P(B)},$$

where the supremum is taken over all $n \geq 1$, all $A \in \sigma\{\xi_j, 1 \leq j \leq n\}$, and all $B \in \sigma\{\xi_j, j \geq n+m\}$. Then $\{\xi_n\}$ is *ψ-mixing* when $\psi_\xi(m) \to 0$ as $m \to \infty$. Note that it is possible to have $\psi_\xi(m) = +\infty$ for a given m. ψ-mixing is obviously a much more restrictive mixing condition than strong mixing. Indeed, a Gaussian ψ-mixing process is always s-dependent for some finite integer s, thus in general a Gaussian ARMA process driven by white noise cannot be ψ-mixing (the more general ϕ-mixing condition, often used to model dependent signals in adaptive filtering, is equivalent to ψ-mixing in the Gaussian case and is thus similarly restrictive in this respect (see Bradley [6, Thm. 5.1]). Nevertheless, there are several interesting examples of (non-Gaussian) ψ-mixing processes, for which we refer the reader to Bradley [6]. ψ-mixing has been used to model dependence in the driving data in several analyses of stochastic gradient algorithms quite similar to (1) (see, e.g., Bitmead [7]). The alternatives to conditions (A) and (C) are as follows:

(A*) The R^{N+1}-valued process $\{(a_j, X_j^T), j \geq 1\}$ is ψ-mixing and strictly stationary. Moreover, there is some integer M and constants $K > 0, \beta > 3$ such that $\psi_{(a,X)}(m) \leq K\, m^{-\beta}$ for all $m \geq M$ (it is possible that $\psi_{(a,X)}(m) = \infty$ when $1 \leq m < M$);

(C*) For all $j \geq 1$ we have $E|X_j|^{96M} < \infty$ and $E|a_j|^{(24+24/(4M-1))} < \infty$, where M is the constant in condition (A*). Note the trade-off between moment bounds and mixing

rates that occurs here: large M implies slower ψ-mixing, and this in turn requires more restrictive moment bounds on X_j.

The main result of this note is the following invariance principle.

PROPOSITION 2.1. (i) *Assume conditions* (A), (B), *and* (C). *Then there exists a probability space* $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ *on which is defined a real-valued process* $\{\tilde{a}_j, j \geq 1\}$, *and* R^N*-valued processes* $\{\tilde{X}_j, j \geq 1\}$, *and* $\{\tilde{W}(t), t \geq 0\}$ *such that*

(a) $\{(\tilde{a}_j, \tilde{X}_j^T), j \geq 1\} \stackrel{D}{=} \{(a_j, X_j^T), j \geq 1\}$, *where* $\stackrel{D}{=}$ *indicates that the left- and right-hand processes have identical finite-dimensional joint distributions*;

(b) $\{\tilde{W}(t), t \geq 0\}$ *is a Brownian motion with covariance matrix* $\mu^2\Gamma$;

(c) *There exists some constant where* $\gamma, 1/2 > \gamma > 0$, *such that, for almost all* $\tilde{\omega}[\tilde{P}]$,

$$t^{1/2}(\tilde{h}_{[t]+1}(\tilde{\omega}) - h_*) = t^{-1/2}\tilde{Y}(t, \tilde{\omega}) + O(t^{-\gamma}) \quad \text{for all } t > 0, \tag{2}$$

where $[a]$ *denotes the integer part of a real number* $a > 0$, *the* $\tilde{h}_j$ *are defined by*

$$\tilde{h}_1 \stackrel{\Delta}{=} h_1 \quad \text{and} \quad \tilde{h}_{j+1} = \tilde{h}_j + \frac{\mu}{j}(\tilde{a}_j - \tilde{h}_j^T \tilde{X}_j)\tilde{X}_j, \qquad j \geq 1, \tag{3}$$

and $\tilde{Y}(t)$ *is a process defined on* $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ *by*

$$\tilde{Y}(t) \stackrel{\Delta}{=} \tilde{W}(t) - (\mu R - I)\int_0^1 \tau^{(\mu R - 2I)}\tilde{W}(\tau t)d\tau, \qquad t > 0. \tag{4}$$

(ii) *Assume conditions* (A*), (B), *and* (C*). *Then all assertions in* (i) *continue to hold.*

Remark 2.1. The constant implicit in the notation $O(t^{-\gamma})$ used in (2) will, of course, depend on $\tilde{w}$. It is easily seen that the process $\tilde{Y}(t)$ in (4) is also the Gauss–Markov process that solves the following linear SDE:

$$d\tilde{Y}(t) = (I - \mu R)t^{-1}\tilde{Y}(t)dt + d\tilde{W}(t) \quad \text{subject to } \tilde{Y}(0) = 0.$$

Remark 2.2. One can consider the algorithm in (1) to be "driven" by a data process $\{(a_j, X_j^T)\}$ that is defined on $(\Omega, \mathbf{A}, P)$ and to be "generating" $\{h_j\}$ on the same probability space in accordance with (1). Proposition 2.1 introduces, in place of $\{(a_j, X_j^T)\}$, a surrogate process $\{(\tilde{a}_j, \tilde{X}_j^T)\}$ of identical distribution but defined on a different probability space $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ along with a Brownian motion $\{\tilde{W}(t), t \geq 0\}$ such that (2) holds. In general it is *not* possible to assert the existence of a Brownian motion $\{W(t), t \geq 0\}$ on the original space $(\Omega, \mathbf{A}, P)$ such that for some constant $\gamma > 0$ and almost all $\omega[P]$, we have

$$t^{1/2}(h_{[t]+1}(\omega) - h_*) = t^{-1/2}(W(t) - (\mu R - I)\int_0^1 \tau^{(\mu R - 2I)}W(\tau t)d\tau) + O(t^{-\gamma}),$$

because $(\Omega, \mathbf{A})$ may not even have enough "measure-theoretic" structure to carry a Brownian motion. It is therefore necessary, when establishing Proposition 2.1, to construct an "enriched" space $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ supporting a Brownian motion as well as a driving data process $\{(\tilde{a}_j, \tilde{X}_j^T)\}$ that is identical in distribution to $\{(a_j, X_j^T)\}$, and then to show that (2) holds. In view of (1), (3), and Proposition 2.1(a), it follows that $\{(a_j, X_j^T, h_j^T), j \geq 1\} \stackrel{D}{=} \{(\tilde{a}_j, \tilde{X}_j^T, \tilde{h}_j^T), j \geq 1\}$, and since the physical manifestation of the stochastic gradient algorithm is solely a consequence of the joint distribution of the "input-output" process $\{(a_j, X_j^T, h_j^T), j \geq 1\}$, there is no loss of generality in switching to the enriched probability space $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ and regarding the algorithm as being implemented by (3) instead. It

should be noted that strong approximation ideas are also considered, from a significantly different point of view, in Gerencser [8].

For the law of the iterated logarithm that follows from Proposition 2.1 we need the following notation due to Kuelbs [9]: If $\{x_j, j \geq 1\}$ is a sequence in a metric space $(\mathbf{X}, \rho)$, then $C(\{x_j\})$ denotes the set of *all* accumulation points (if any) of the sequence, and, if K is a compact subset of $(\mathbf{X}, \rho)$ and $\rho(x, K) \triangleq \inf\{\rho(x, y); y \in K\}$ is the distance from a point x to K, then $\{x_j, j \geq 1\} \to\to K$ indicates (i) that $\rho(x_j, K) \to 0$ as $j \to \infty$, and (ii) that $C(\{x_j\}) = K$. Let J denote a *fixed* integer such that $\log\log(j) \geq 1$ for all $j \geq J$.

PROPOSITION 2.2 (law of the iterated logarithm). (i) *Assume conditions* (A)–(D). *Then for almost all* $\omega[P]$,

$$\left\{ \frac{j^{1/2}(h_j(\omega) - h_*)}{(2\log\log(j))^{1/2}}, j \geq J \right\} \to\to K, \tag{5}$$

where $K \triangleq \{x \in R^N | \langle x, M^{-1}x\rangle \leq \mu^2\}$, *and* M *is the unique positive definite solution of the Lyapunov equation* $(2\mu R - I)M + M(2\mu R - I) = 2\Gamma$.

(ii) *Assume conditions* (A*), (B), (C*), *and* (D). *Then* (5) *holds for almost all* $\omega[P]$.

Proposition 2.2 immediately implies the following almost sure rate of convergence of $\{h_j\}$ to h_*:

$$|h_j(\omega) - h_*| = O\left(\left(\frac{\log\log(j)}{j}\right)^{1/2}\right).$$

Observe that this rate of convergence cannot be improved. Indeed, if there is a better rate for some ω, then there exists a sequence of real numbers $\{\psi_j(\omega)\}$ such that

$$|h_j(\omega) - h_*| = O(\psi_j(\omega)) \quad \text{where } \psi_j(\omega) = o\left(\left(\frac{\log\log(j)}{j}\right)^{1/2}\right).$$

However, this implies that $\lim_{j\to\infty}(j^{1/2}(h_j(\omega) - h_*))/(2\log\log(j))^{1/2}) = 0$ which, in view of (5), can take place on a set of P-measure zero only.

PROPOSITION 2.3 (functional central limit theorem). *Assume either conditions* (A), (B), (C), *or conditions* (A*), (B), (C*). *Then the sequence* $\{\xi_n(t), 0 \leq t \leq 1\}$ *of processes defined by*

$$\xi_n(t) \triangleq tn^{1/2}(h_{[nt]+1} - h_*)$$

for all $0 \leq t \leq 1$ *and integers* $n \geq 1$ *converges weakly to the process* $\{\tilde{Y}(t), 0 \leq t \leq 1\}$, *defined by* (4) *as* $n \to \infty$.

3. Proofs of propositions 2.1–2.3. We establish Proposition 2.1 by making use of the following fairly special consequence of Berger [10, Cor. 2.24, p. 541] (which is itself a significant extension of similar approximation theorems obtained in a one-dimensional setting by Ruppert [11] and in the important dissertation of Mark [12]).

THEOREM 3.1. *Assume that a sequence* $\{\nu_j, j \geq 1\}$ *of* R^N*-vectors is generated from a given sequence of* R^N*-vectors* $\{w_j, j \geq 1\}$ *by the recursion*

$$\nu_{j+1} = (I - j^{-1}\Lambda)\nu_j + j^{-1}w_j, \qquad j \geq 1, \tag{6}$$

where Λ is a symmetric N by N matrix such that $(2\Lambda - I)$ is positive definite. If $\{w_j\}$ is a random sequence defined on a probability space $(\Xi, \boldsymbol{F}, Q)$ along with an R^N-Brownian motion $\{B(t), t \geq 0\}$ such that, for some constant $\eta > 0$,

$$B(t) - \sum_{1 \leq j \leq t} w_j = O(t^{1/2-\eta}) \quad \textit{for all } t \geq 1, \textit{ a.s. } [Q], \tag{7}$$

then there exists a constant γ, where $\frac{1}{2} > \gamma > 0$, such that almost surely [Q],

$$[t]\nu_{[t]+1} - \left(B(t) - (\Lambda - I)\int_0^1 \tau^{(\Lambda - 2I)} B(\tau t)\, d\tau\right) = O(t^{1/2-\gamma}) \quad \textit{for all } t > 0. \tag{8}$$

Note the similarity between the approximating processes in (8) and (4). In using this theorem, one is usually given the sequence $\{w_j\}$ and the main task is to show existence of some $B(t)$ for which (7) holds. We see that the main technical difficulty in carrying out this step arises from the fact that w_j depends on ν_j in our application (see (9) and (10)).

Proof of Proposition 2.1(i). Define

$$\nu_j \triangleq h_j - h_*, z_j \triangleq a_j X_j - X_j X_j^T h_* \quad \text{and} \quad w_j \triangleq \mu z_j - \mu(X_j X_j^T - R)\nu_j \tag{9}$$

for $j \geq 1$. Then $E(z_j) = 0$ and, from (1) and (9),

$$\nu_1 \triangleq h_1 - h_* \quad \text{and} \quad \nu_{j+1} = (I - j^{-1}\mu R)\nu_j + j^{-1} w_j \quad \text{for } j \geq 1. \tag{10}$$

According to Lemma 4.1(i), there exists a probability space $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ carrying real and R^N-valued processes $\{\tilde{a}_j, j \geq 1\}$ and $\{\tilde{X}_j, j \geq 1\}$, respectively, along with an R^N-valued Brownian motion $\{\tilde{W}(t), t \geq 0\}$ with covariance matrix $\mu^2\Gamma$, such that:

$$\{(\tilde{a}_j, \tilde{X}_j^T), j \geq 1\} \stackrel{D}{=} \{(a_j, X_j^T), j \geq 1\}, \tag{11}$$

and

$$\sum_{1 \leq j \leq t} \mu \tilde{z}_j - \tilde{W}(t) = O(t^{1/2-\eta}) \quad \text{for all } t \geq 1, \quad \text{a.s. } [\tilde{P}], \tag{12}$$

for some constant $\eta > 0$, where

$$\tilde{z}_j \triangleq \tilde{a}_j \tilde{X}_j - \tilde{X}_j \tilde{X}_j^T h_*. \tag{13}$$

Define $\{\tilde{\nu}_j, j \geq 1\}$ and $\{\tilde{w}_j, j \geq 1\}$ by the following recursions:

$$\tilde{\nu}_1 \triangleq \nu_1 \quad \text{and} \quad \tilde{\nu}_{j+1} = (I - j^{-1}\mu R)\tilde{\nu}_j + j^{-1}\tilde{w}_j \quad \text{for } j \geq 1, \tag{14}$$

$$\tilde{w}_j \triangleq \mu \tilde{z}_j - \mu(\tilde{X}_j \tilde{X}_j^T - R)\tilde{\nu}_j. \tag{15}$$

From (10), (11), and (14) it follows that $\{(a_j, X_j^T, \nu_j^T), j \geq 1\} \stackrel{D}{=} \{(\tilde{a}_j, \tilde{X}_j^T, \tilde{\nu}_j^T), j \geq 1\}$, and by (12), (15), and Lemma 4.2, there is some constant $\eta > 0$ such that

$$\sum_{1 \leq j \leq t} \tilde{w}_j - \tilde{W}(t) = O(t^{1/2-\eta}) \quad \text{for all } t \geq 1, \quad \text{a.s. } [\tilde{P}]. \tag{16}$$

By (14), (16), and Theorem 3.1, there exists some constant $1/2 > \gamma > 0$ such that

$$t^{1/2}\tilde{\nu}_{[t]+1} = t^{-1/2}\tilde{Y}(t) + O(t^{-\gamma}) \quad \text{for all } t > 0, \quad \text{a.s. } [P], \tag{17}$$

where $\tilde{Y}(t)$ is given by (4). Let $\tilde{h}_j \stackrel{\Delta}{=} \tilde{\nu}_j + h_*$ for $j \geq 1$. From (13), (14), and (15), it is seen that the process $\{\tilde{h}_j, j \geq 1\}$ is generated by the recursion in (3), and by (17) it also satisfies the invariance principle (2), as required. □

Proof of Proposition 2.1(ii). This is identical to the proof of Proposition 2.1(i) except that Lemma 4.1(ii) is used in place of Lemma 4.1(i), and Lemma 4.3 is used instead of Lemma 4.2. □

Before establishing Proposition 2.2 let us introduce the following notation: $C_0^N[0,1]$ denotes the set of continuous functions $\phi : [0,1] \to R^N$ such that $\phi(0) = 0$, with the usual metric of uniform convergence over the interval [0,1], and $L_2^N[0,1]$ denotes the set of functions $\phi : [0,1] \to R^N$ whose components are square integrable over [0,1]. Let S denote the space of absolutely continuous functions in $C_0^N[0,1]$ defined by

$$S \stackrel{\Delta}{=} \left\{ \phi(\cdot) \in C_0^N[0,1] | \dot{\phi}(\cdot) \in L_2^N[0,1], \int_0^1 \langle \dot{\phi}(\tau), \Gamma^{-1}\dot{\phi}(\tau) \rangle d\tau \leq \mu^2 \right\}. \tag{18}$$

Proof of Proposition 2.2. The proofs of parts (i) and (ii) are identical. By Proposition 2.1 there exists some probability space $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ carrying a Brownian motion $\tilde{W}(t)$ with covariance matrix $\mu^2\Gamma$, and processes $\{\tilde{a}_j\}, \{\tilde{X}_j\}$, and $\{\tilde{h}_j\}$ such that $\{(a_j, X_j^T, h_j^T), j \geq 1\} \stackrel{D}{=} \{(\tilde{a}_j, \tilde{X}_j^T, \tilde{h}_j^T), j \geq 1\}$. The process $\{\tilde{h}_j\}$ is generated by the recursion in (3), and, for some constant $\frac{1}{2} > \gamma > 0$,

$$j^{1/2}(\tilde{h}_{j+1} - h_*) = j^{-1/2}\tilde{Y}(j) + O(j^{-\gamma}) \quad \text{for all } j \geq 1, \quad \text{a.s. } [\tilde{P}], \tag{19}$$

where $\tilde{Y}(j)$ is defined by (4). Let Δ be the set of all functions ϕ in $C_0^N[0,1]$ such that each component of the R^N-valued function $t \to t^{(\mu R - 2I)}\phi(t)$ is integrable over $0 < t \leq 1$ (Δ will be a strict subset of $C_0^N[0,1]$ when $\mu R - 2I$ has one or more negative eigenvalues). Define the function $F : \Delta \to R^N$ by

$$F(\phi) \stackrel{\Delta}{=} \phi(1) - (\mu R - I) \int_0^1 \tau^{(\mu R - 2I)}\phi(\tau) d\tau. \tag{20}$$

By Lemma 4.4(a), $S \subset \Delta$ whence $F[S]$, the image of S under F, is well defined. Thus, by Lemma 4.5 along with (19),

$$\left\{ j^{1/2} \frac{(\tilde{h}_{j+1}(\tilde{\omega}) - h_*)}{(2\,\text{loglog}(j))^{1/2}}, j \geq J \right\} \to\to F[S] \tag{21}$$

for almost all $\tilde{\omega}[\tilde{P}]$. It remains to show that $F[S]$ is equal to the set K in the statement of Proposition 2.2. Define an inner product $[\cdot, \cdot]$ in $L_2^N[0,1]$ as follows:

$$[\eta_1, \eta_2] \stackrel{\Delta}{=} \int_0^1 \langle \eta_1(\tau), \Gamma^{-1}\eta_2(\tau) \rangle d\tau \tag{22}$$

for $\eta_1, \eta_2 \in L_2^N[0,1]$. Define linear subspaces V_1 and V_2 by

$$V_1 \triangleq \{\psi(\cdot) \in L_2^N[0,1] | \psi(t) \triangleq \Gamma t^{(\mu R - I)} Dx \text{ for some } x \in R^N\}, \tag{23a}$$

$$V_2 \triangleq \{\eta(\cdot) \in L_2^N[0,1] | [\eta, \psi] = 0 \text{ for all } \psi \in V_1\}, \tag{23b}$$

where condition (B) in §2 ensures square integrability of the functions $\psi(\cdot)$ in V_1, and D denotes a non-singular $N \times N$ matrix such that $D^T D = (2\mu R - I)$. Clearly, $L_2^N[0,1]$ is a Hilbert space when the inner product is defined by (22), and this inner product will henceforth always be assumed. Let

$$\psi(t) \triangleq \Gamma t^{\mu R - I)} Dx, \tag{24}$$

where $x \in R^N$, be a typical element in V_1. Then we see that

$$[\psi, \psi] = \langle Dx, MDx \rangle, \tag{25}$$

where

$$M \triangleq \int_0^1 \tau^{(\mu R - I)} \Gamma \tau^{(\mu R - I)} d\tau, \tag{26}$$

which is clearly positive definite. In view of (25), V_1 is complete and therefore closed in the norm generated by the inner product in (22). Furthermore, because V_2 is the orthogonal complement of V_1 (in $L_2^N[0,1]$), from the projection theorem in Hilbert space it follows that each element in $L_2^N[0,1]$ can be uniquely expressed as the sum of a vector in V_1 and a vector in V_2:

$$L_2^N[0,1] = V_1 + V_2 \tag{27}$$

(see Friedman [13, p. 209]). Using the integration by parts justified by Lemma 4.4(b), the function F in (20) can be written as

$$F(\phi) = \int_0^1 \tau^{(\mu R - I)} \dot{\phi}(\tau) d\tau \tag{28}$$

for all $\phi \in S$, and because (by (22), the nonsingularity of Γ and D, and the definition of V_2)

$$\int_0^1 \tau^{(\mu R - I)} \eta(\tau) d\tau = 0 \tag{29}$$

for all $\eta \in V_2$, it follows from (18) and (27)–(29) that

$$F[S] = \left\{ \int_0^1 \tau^{(\mu R - I)} \psi(\tau) d\tau | \psi(\cdot) \in V_1, [\psi, \psi] \leq \mu^2 \right\}. \tag{30}$$

For $\psi(\cdot) \in V_1$ defined by (24), clearly

$$\int_0^1 \tau^{(\mu R - I)} \psi(\tau) d\tau = MDx, \tag{31}$$

where M is given by (26), and therefore (25), (30), and (31) imply that

$$F[S] = \{x \in R^N | \langle x, M^{-1}x\rangle \leq \mu^2\}. \tag{32}$$

Finally, substituting $\sigma = -\log\tau$ into (26) gives

$$M = \int_0^\infty e^{-\sigma(\mu R - I/2)}\Gamma e^{-\sigma(\mu R - I/2)}d\sigma, \tag{33}$$

whence M is the positive-definite and unique solution of the Lyapunov equation in Proposition 2.2 (see, e.g., Vidyasagar [14, Thm. 55 and (56), p. 175]). Proposition 2.2 follows from (21), (32), and the equality in distribution of $\{h_j, j \geq 1\}$ and $\{\tilde{h}_j, j \geq 1\}$. □

We now establish Proposition 2.3, the functional central limit theorem. It is in fact a very easy corollary of Proposition 2.1.

Proof of Proposition 2.3. Replacing t in (2) and (4) with nt, where $0 < t \leq 1$ and $n \geq 1$ is an integer, we see that for almost all $\tilde{\omega}[\tilde{P}]$,

$$\begin{aligned} &n^{1/2}t^{1/2}(\tilde{h}_{[nt]+1}(\tilde{\omega}) - h_*) \\ &= n^{-1/2}t^{-1/2}\left[\tilde{W}(nt,\tilde{\omega}) - (\mu R - I)\int_0^1 \tau^{(\mu R - 2I)}\tilde{W}(nt\tau,\tilde{\omega})d\tau\right] + t^{-\gamma}O(n^{-\gamma}) \end{aligned} \tag{34}$$

for $0 < t \leq 1$ and $n \geq 1$. Now define $\tilde{W}_n(t) \triangleq n^{-1/2}\tilde{W}(nt)$ and

$$\tilde{Y}_n(t) \triangleq \tilde{W}_n(t) - (\mu R - I)\int_0^1 \tau^{(\mu R - 2I)}\tilde{W}_n(\tau t)d\tau \tag{35}$$

for $0 < t \leq 1$ and $n \geq 1$. Multiplying each side of (34) by $t^{1/2}$ and using the fact that we can choose γ in (34) such that $0 < \gamma < \frac{1}{2}$, for almost all $\tilde{\omega}[\tilde{P}]$ it follows that

$$\tilde{\xi}_n(t,\tilde{\omega}) = \tilde{Y}_n(t,\tilde{\omega}) + O(n^{-\gamma}) \tag{36}$$

for $0 \leq t \leq 1$, and $n \geq 1$, where $\tilde{\xi}_n(t) \triangleq tn^{1/2}(\tilde{h}_{[nt]+1} - h_*)$, and the constant of proportionality implied by $O(n^{-\gamma})$ does not depend on t. Since the Brownian motions $\tilde{W}(\cdot)$ and $\tilde{W}_n(\cdot)$ have identical distributions, so also do the processes $\tilde{Y}(t)$ and $\tilde{Y}_n(t), 0 \leq t \leq 1$, defined in (4) and (35), respectively. Moreover, in view of (36),

$$\lim_{n\to\infty} \rho(\tilde{\xi}_n(\cdot,\tilde{\omega}), \tilde{Y}_n(\cdot,\tilde{\omega})) = 0 \tag{37}$$

for almost all $\tilde{\omega}[\tilde{P}]$, where $\rho(\cdot,\cdot)$ denotes the Skorohod metric in $D[0,1;R^N]$. Thus, by [15, Thm. 4.1, p. 25] of Billingsley, the sequence $\{\tilde{\xi}_n(t), 0 \leq t \leq 1\}$ converges weakly to $\{\tilde{Y}(t), 0 \leq t \leq 1\}$. □

4. Appendix I. Various technical results needed for the proofs in §3 are established in this appendix. Lemma 4.1 is used in the proof of Proposition 2.1 (see (11) and (12)).

LEMMA 4.1. (i) *Assume conditions* (A), (B), *and* (C). *Then there exists a probability space* $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ *carrying real and* R^N*-valued processes* $\{\tilde{a}_j, j \geq 1\}$ *and* $\{\tilde{X}_j, j \geq 1\}$, *respectively, along with an* R^N*-valued Brownian motion* $\{\tilde{W}(t), t \geq 0\}$, *whose covariance matrix is* $\mu^2\Gamma$, *such that*

(a) $$\{(\tilde{a}_j, \tilde{X}_j^T), j \geq 1\} \stackrel{D}{=} \{(a_j, X_j^T), j \geq 1\}.$$

Furthermore, there exists a constant $\eta > 0$ *such that*

(b) $$\sum_{1 \le j \le t} \mu(\tilde{a}_j \tilde{X}_j - \tilde{X}_j \tilde{X}_j^T h_*) = \tilde{W}(t) + O(t^{1/2-\eta}) \quad \textit{for all } t \ge 1, \quad \textit{a.s. } [\tilde{P}].$$

(ii) *Assume conditions* (A*), (B), *and* (C*). *Then the assertions in* (i) *hold.*

Proof. (i) Define the R^{2N+1}-valued process $\{\xi_n, n \ge 1\}$ by

(38) $$\xi_n^T \triangleq (a_n - E(a_n), X_n^T - E(X_n^T), (a_n X_n - X_n X_n^T h_*)^T).$$

By condition (A), this is a zero-mean process that is strong mixing with $\alpha_{\xi_n}(m) = O(n^{-\beta})$ for some $\beta > 3$. Moreover, the moment bounds in condition (C) are ample to ensure that the ξ_n have third order moments. By Theorem 5.2 the matrix

$$T \triangleq E(\xi_1 \xi_1^T) + \sum_{j=2}^{\infty} E(\xi_1 \xi_j^T) + \sum_{j=2}^{\infty} E(\xi_j \xi_1^T)$$

exists, and there is some probability space $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$ carrying an R^{2N+1}-valued process $\{\tilde{\xi}_n, n \ge 1\}$ along with an R^{2N+1}-valued Brownian motion $\{\tilde{B}(t), t \ge 0\}$ whose covariance matrix is T, such that for some number $\eta > 0$

(39) $$\{\xi_n, n \ge 1\} \stackrel{D}{=} \{\tilde{\xi}_n, n \ge 1\}$$

and

(40) $$\sum_{1 \le n \le t} \tilde{\xi}_n = \tilde{B}(t) + O(t^{1/2-\eta}), \qquad t \ge 1, \quad \text{a.s. } [\tilde{P}].$$

Now partition the vector $\tilde{\xi}_n$ as follows:

(41) $$\tilde{\xi}_n^T \triangleq (\tilde{a}_n - E(a_n), \tilde{X}_n^T - E(X_n^T), \tilde{\zeta}_n^T),$$

where $\{\tilde{a}_n\}$ is a real-valued process and $\{\tilde{X}_n\}, \{\tilde{\zeta}_n\}$ are R^N-valued processes, all defined on $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$. It is clear from (38)–(40) that

(42) $$\{(a_n, X_n^T), n \ge 1\} \stackrel{D}{=} \{(\tilde{a}_n, \tilde{X}_n^T), n \ge 1\}$$

and

(43) $$\tilde{\zeta}_n = \tilde{a}_n \tilde{X}_n - \tilde{X}_n \tilde{X}_n^T h_* \quad \text{a.s. } [\tilde{P}].$$

Now condition (D) implies that the lower $N \times N$ submatrix of T is Γ while the process $\{\tilde{V}(t), t \ge 1\}$ is an R^N-valued Brownian motion with covariance matrix Γ when $\tilde{V}(t)$ is the vector that consists of the last N elements of $\tilde{B}(t)$. From (40), (41), and (43) we see that

$$\sum_{1 \le n \le t} (\tilde{a}_n \tilde{X}_n - \tilde{X}_n \tilde{X}_n^T h_*) = \tilde{V}(t) + O(t^{1/2-\eta}) \quad \text{for all } t \ge 1.$$

The lemma follows upon defining $\tilde{W}(t) \triangleq \mu \tilde{V}(t)$.

(ii) Condition (A*) implies (A), and clearly (C*) ensures that ξ_n in (38) has third order moments. The proof goes through as in (i) with no changes. □

Lemma 4.2 is used in the proof of Proposition 2.1 (see (16)).

LEMMA 4.2. *Assume conditions* (A), (B), *and* (C). *Then there exists some constant* $\eta > 0$ *such that, almost surely* $[P]$,

$$\sum_{1\le j\le t}(X_jX_j^T - R)\nu_j = O(t^{1/2-\eta}) \quad \textit{for all } t \ge 1, \tag{45}$$

where $\nu_j \triangleq h_j - h_*$.

Proof. Fix $0 < b < 1$ and define $s(j) \triangleq [j^b], U_j \triangleq X_jX_j^T$, and $\mu_j \triangleq \mu/j$ for integers $j \ge 1$. From (1) and (9),

$$\nu_j = \nu_{j-s(j)} + \eta_j\nu_{j-s(j)} + \xi_j \quad \text{for } j \ge 2, \tag{46}$$

where

$$\eta_j \triangleq \left\{\prod_{i=1}^{s(j)}(I - \mu_{j-i}U_{j-i})\right\} - I, \tag{47}$$

$$\xi_j \triangleq \sum_{k=1}^{s(j)}\left\{\prod_{i=1}^{k-1}(I - \mu_{j-i}U_{j-i})\right\}\mu_{j-k}z_{j-k}. \tag{48}$$

(The matrix products in (47) and (48) are from left to right with increasing i.) The proof is given in three steps.

Step I. We show that there are constants $0 < b < \frac{1}{2}$ and $\eta > 0$ such that almost surely $[P]$,

$$\sum_{1\le j\le t}(U_j - R)\nu_{j-s(j)} = O(t^{1/2-\eta}), \tag{49}$$

where, for convenience, $\nu_0 \triangleq 0$. Without loss of generality, suppose $N = 1$, since the argument used in this step generalises trivially to the vector-valued case. Fix any $\eta > 0$ and define

$$u_j \triangleq j^{-(1/2-\eta)}(U_j - R)\nu_{j-s(j)} \quad \text{for all } j \ge 1. \tag{50}$$

It remains to show that

$$\sum_{j=1}^{\infty} u_j < \infty \quad \text{a.s. } [P], \tag{51}$$

for then (49) follows by the Kronecker lemma. Clearly,

$$E\left[\sum_{j=a+1}^{a+n} u_j\right]^2 = E\left[\sum_{j=a+1}^{a+n} u_j^2\right] + 2E\left[\sum_{\substack{i,j=a+1\\ a+1\le i\le j-s(j)}}^{a+n} u_iu_j\right] + 2E\left[\sum_{\substack{i,j=a+1\\ j-s(j)<i<j}}^{a+n} u_iu_j\right]. \tag{52}$$

Each of the terms on the right of (52) is now bounded. By (50) and Cauchy–Schwarz,

$$E(u_j^2) \le j^{-(1-2\eta)}\|(U_j - R)^2\|_2\|\nu_{j-s(j)}^2\|_2. \tag{53}$$

Now, clearly, $\sup_j \|(U_j - R)^2\|_2 \leq K_1 < \infty$, and taking $r = 2$ in Theorem 5.4 gives

$$E|\nu_{j-s(j)}|^4 \leq K_2(j - s(j))^{-1} \qquad j \geq 2, \tag{54}$$

for some constant K_2. Moreover, obviously there is a constant K_3 such that $(j-s(j))^{-1} \leq K_3 j^{-1}$ for each $j \geq 2$. Hence, by (53) and (54) there is a constant K_4, where

$$E\left[\sum_{j=a+1}^{a+n} u_j^2\right] \leq K_4 \left[\sum_{j=a+1}^{a+n} j^{-(3/2-2\eta)}\right] \tag{55}$$

for all integers $a \geq 0, n \geq 1$. Next we bound the second term on the right of (52). Because $a + 1 \leq i \leq j - s(j)$, the function $(U_i - R)\nu_{i-s(i)}\nu_{j-s(j)}$ is measurable with respect to $\sigma\{(a_k, X_k^T), k \leq j - s(j)\}$ and $(U_j - R)$ is clearly $\sigma\{(a_k X_k^T), k \geq j\}$-measurable. Writing $\alpha(m)$ for $\alpha_{(a,X)}(m)$ and using Theorem 5.1(i), below, along with $E(U_j - R) = 0$ and condition (A),

$$\begin{aligned} E(u_i u_j) &\leq 10(ij)^{-(1/2-\eta)}\alpha^{1/p}(s(j))\|U_j - R\|_q \|(U_i - R)\nu_{i-s(i)}\nu_{j-s(j)}\|_p \\ &\leq 10(ij)^{-(1/2-\eta)}\alpha^{1/p}(s(j))\|U_j - R\|_q \{\|U_i - R\|_{pq}\|\nu_{i-s(i)}\|_{p^2}\|\nu_{j-s(j)}\|_{p^2}\}, \end{aligned} \tag{56}$$

where $p \triangleq (4m + 1)/2m, q \triangleq (4m + 1)$, for any positive integer m, and the second line in (56) follows by Holder's inequality (the dependence of p and q on m will not be indicated). By condition (C) there is some constant $K_5(m)$, depending only on m, such that $\|U_j - R\|_q \|U_i - R\|_{pq} \leq K_5(m)$ for all $i, j \geq 1$. Also

$$|\nu_{j-s(j)}|^{p^2} \leq |\nu_{j-s(j)}|^4 + |\nu_{j-s(j)}|^8 \tag{57}$$

for all integers $m \geq 1, j \geq 2$, since $4 \leq p^2(m) \leq 8$ for all $m \geq 1$. Taking first $r = 2$, then $r = 4$ in Theorem 5.4 and using condition (B) shows that $\|\nu_{j-s(j)}\|_{p^2} \leq K_6 j^{-(1/p^2)}$. Defining $\gamma(m) \triangleq ((1/2) + (1/p^2(m)) - \eta)$, we then obtain from (56) that

$$\begin{aligned} E\left[\sum_{\substack{i,j=a+1 \\ a+1\leq i\leq j-s(j)}}^{a+n} u_i u_j\right] &\leq K_7(m) \sum_{j=a+2}^{a+n} \alpha^{1/p}(s(j)) j^{-\gamma} \sum_{i=a+1}^{j-s(j)} i^{-\gamma} \\ &\leq K_8(m) \sum_{j=a+2}^{a+n} \alpha^{1/p}(s(j)) j^{-\gamma} j^{1-\gamma} \\ &\leq K_9(m) \sum_{j=a+1}^{a+n} j^{(1-2\gamma-b\beta/p)}. \end{aligned} \tag{58}$$

Choose $0 < b < \frac{1}{2}$ such that $b\beta > 1$. Then, since $p(m) \to 2$ as $m \to \infty$, we can fix m large enough to obtain $(1 - 2\gamma - b\beta/p) = -\theta + 2\eta$ for some $\theta > 1$; thus

$$E\left[\sum_{\substack{i,j=a+1 \\ a+1\leq i\leq j-s(j)}}^{a+n} u_i u_j\right] \leq K_{10} \sum_{j=a+1}^{a+n} j^{-(\theta-2\eta)} \tag{59}$$

for all $a \geq 0, n \geq 1$, and a constant $\theta > 1$. Now we bound the third term on the right of (52). Since $j - s(j) < i < j$ and $E(U_j - R) = 0$, from Theorem 5.1(i) and an argument similar to that which gave (56),

$$\begin{aligned} E(u_i u_j) &\leq 10(ij)^{-(1/2-\eta)}\alpha^{1/p}(j - i))\|U_j - R\|_q \|(U_i - R)\nu_{i-s(i)}\nu_{j-s(j)}\|_p \\ &\leq K_{11}(m)\alpha^{1/p}(j - i) i^{-\gamma} j^{-\gamma} \leq K_{11}(m)\alpha^{1/p}(j - i) i^{-2\gamma}, \end{aligned} \tag{60}$$

where, as before, $p \triangleq (4m+1)/2m, q \triangleq (4m+1)$, and $\gamma \triangleq ((1/2)+(1/p^2)-\eta)$. For some positive integer m and with no loss of generality we can take $\gamma > 0$. Thus, from (60),

$$E\left[\sum_{\substack{i,j=a+1\\ j-s(j)<i<j}}^{a+n} u_i u_j\right] \leq K_{11}(m)\sum_{k=1}^{n-1}\sum_{i=a+1}^{a+n-k}\alpha^{1/p}(k) i^{-2\gamma}. \tag{61}$$

Choosing m large enough to ensure $(1/p)\beta > 1$ and $2\gamma = \psi - 2\eta$ for some $\psi > 1$ gives $\sum \alpha^{1/p}(k) < \infty$, and hence from (61),

$$E\left[\sum_{\substack{i,j=a+1\\ j-s(j)<i<j}}^{a+n} u_i u_j\right] \leq K_{12}\sum_{j=a+1}^{a+n} j^{-(\psi-2\eta)} \tag{62}$$

for some constant $\psi > 1$ and all $a \geq 0$ and $n \geq 1$. Putting (52), (55), (59), and (62) together shows that for some constant $\zeta > 1$,

$$E\left[\sum_{j=a+1}^{a+n} u_j\right]^2 \leq K\sum_{j=a+1}^{a+n} j^{-(\zeta-2\eta)} \tag{63}$$

for all $a \geq 0, n \geq 1$. Now choose $\eta > 0$ such that

$$\sum_{j=1}^{\infty} \log^2(j) j^{-(\zeta-2\eta)} < \infty, \tag{64}$$

and, for integers $a \geq 0, n \geq 1$, define

$$g(a,n) \triangleq K\sum_{j=a+1}^{a+n} j^{-(\zeta-2\eta)}, \qquad h(a,n) \triangleq K\sum_{j=a+1}^{a+n} \log^2(j) j^{-(\zeta-2\eta)}. \tag{65}$$

Clearly Theorem 5.3 applies and (51) follows.

Step II. We show that for each constant $0 < b < 1$ there is some $\eta > 0$ such that almost surely $[P]$,

$$\sum_{1\leq j\leq t} (U_j - R)\eta_j \nu_{j-s(j)} = O(t^{1/2-\eta}), \tag{66}$$

where η_j is given by (47) and $s(j) \triangleq [j^b]$. It is enough to show that

$$\sum_{j=1}^{\infty} j^{-(1/2-\eta)} E|(U_j - R)\eta_j \nu_{j-s(j)}| < \infty, \tag{67}$$

for this implies

$$\sum_{j=1}^{\infty} j^{-(1/2-\eta)} |(U_j - R)\eta_j \nu_{j-s(j)}| < \infty \quad \text{a.s.}, \tag{68}$$

from which (66) follows by Kronecker's lemma. Thus fix any $0 < b < 1$ and define

$$M_j \triangleq (U_j - R)\eta_j. \tag{69}$$

If $|A|$ denotes a matrix norm on the set of N by N matrices such that $|AB| \leq |A||B|$ for matrices A and B, then by the Cauchy–Schwarz inequality and an application of Theorem 4.4 with $r = 1$, there is a constant K_0 such that

$$E|(U_j - R)\eta_j \nu_{j-s(j)}| \leq E^{1/2}(|M_j|^2)E^{1/2}(|\nu_{j-s(j)}|^2) \leq K_0 E^{1/2}(|M_j|^2) j^{-1/2} \tag{70}$$

for all $j \geq 2$. To bound $E^{1/2}|M_j|^2$ note from (47) that

$$\eta_j = \sum_{k=1}^{s(j)} (-1)^k \left\{ \sum_{1 \leq i_1 < \cdots < i_k \leq s(j)} (\mu_{j-i_1} \cdots \mu_{j-i_k})(U_{j-i_1} \cdots U_{j-i_k}) \right\}. \tag{71}$$

Also, for all $1 \leq k \leq s(j)$,

$$\begin{aligned} \left\{ \sum_{i=1}^{s(j)} \mu_{j-i}|U_{j-i}| \right\}^k &= \prod_{r=1}^{k} \left\{ \sum_{i_r=1}^{s(j)} \mu_{j-i_r}|U_{j-i_r}| \right\} \\ &\geq k! \left\{ \sum_{1 \leq i_1 < \cdots < i_k \leq s(j)} (\mu_{j-i_1} \cdots \mu_{j-i_k})(|U_{j-i_1}| \cdots |U_{j-i_k}| \right\}. \end{aligned} \tag{72}$$

Taking matrix norms in (71) and using (72) gives

$$|\eta_j| \leq \sum_{k=1}^{s(j)} (1/k!) \left\{ \sum_{i=1}^{s(j)} \mu_{j-i}|U_{j-i}| \right\}^k. \tag{73}$$

Now for any $k \geq 2$, clearly $\{\mu_{j-i}|U_{j-i}|\} \leq (\mu_{j-i})^{(k-1)/k}\{(\mu_{j-i})^{1/k}|U_{j-i}|\}$; thus, by Holder's inequality,

$$\left\{ \sum_{i=1}^{s(j)} \mu_{j-i}|U_{j-i}| \right\}^k \leq \left\{ \sum_{i=1}^{s(j)} \mu_{j-i} \right\}^{(k-1)} \left\{ \sum_{i=1}^{s(j)} \mu_{j-i}|U_{j-i}|^k \right\} \tag{74}$$

for all $k \geq 1$. Hence, from (69), (73), and (74),

$$E(|M_j|^2) \leq \sum_{k_1=1}^{s(j)} \sum_{k_2=1}^{s(j)} \left\{ \sum_{i_1=1}^{s(j)} \mu_{j-i_1} \right\}^{(k_1-1)} \left\{ \sum_{i_2=1}^{s(j)} \mu_{j-i_2} \right\}^{(k_2-1)} \left\{ \sum_{i_1=1}^{s(j)} \sum_{i_2=1}^{s(j)} \mu_{j-i_1}\mu_{j-i_2} B \right\}, \tag{75}$$

where

$$\begin{aligned} B &\triangleq (1/k_1!)(1/k_2!)E[|U_j - R|^2 \cdot |U_{j-i_1}|^{k_1} \cdot |U_{j-i_2}|^{k_2}] \\ &\leq K_1\{(1/k_1!)E^{1/4}(|U_{j-i_1}|^{4k_1})\}\{(1/k_2!)E^{1/4}(|U_{j-i_2}|^{4k_2})\} \end{aligned} \tag{76}$$

for some constant K_1, where the last line follows from Holder's inequality. Now by condition (C) there is a constant K_2 such that

$$(1/k!)E^{1/4}(|U_j|^{4k}) \leq (C^{2k} \cdot (2k)!)^{1/4}/(k!) \leq K_2(2eC)^{k/2} \equiv K_2 C_1^k \tag{77}$$

where $(2k)!$ has been upper-bounded and $k!$ lower-bounded using the following extended version of Stirling's formula:

$$(2\pi)^{1/2}n^{n+1/2}e^{-n}e^{1/(1+12n)} < n! < (2\pi)^{1/2}n^{n+1/2}e^{-n}e^{1/12n}$$

(see Feller [16, (9.15), p. 54]). From (75) and (77)

$$E(|M_j|^2) \le K_3 \left\{ \sum_{k=1}^{s(j)} \left(C_1 \sum_{i=1}^{s(j)} \mu_{j-i} \right)^k \right\}^2. \tag{78}$$

Since $0 < b < 1$, there is some integer J_1 such that

$$C_1 \sum_{i=1}^{s(j)} \mu_{j-i} \le 2C_1 \mu j^{b-1} < \frac{1}{2} \tag{79}$$

for all $j \ge J_1$; hence by (78) there is a constant K_4 such that

$$E(|M_j|^2) \le K_3 \left\{ \sum_{k=1}^{\infty} (2C_1 j^{b-1})^k \right\}^2 \le K_4 j^{2b-2} \tag{80}$$

for all $j \ge J_1$. By (70) and (80), for each $0 < b < 1$ there is some $\eta > 0$ such that (67) holds, as required.

Step III. We show that for each constant $0 < b < \frac{1}{2}$ there is some $\eta > 0$ such that almost surely $[P]$,

$$\sum_{1 \le j \le t} (U_j - R)\xi_j = O(t^{1/2-\eta}), \tag{81}$$

where ξ_j is given by (48) and $s(j) \triangleq [j^b]$. Define

$$x_j \triangleq |(U_j - R)\xi_j| \le \sum_{k=1}^{s(j)} \mu_{j-k}\beta_{j,k}|C_{j,k}|, \tag{82}$$

where, from (48),

$$C_{j,k} \triangleq \prod_{i=1}^{k-1} (I - \mu_{j-i}U_{j-i}), \tag{83}$$

$$\beta_{j,k} \triangleq |U_j - R| \cdot |z_{j-k}|. \tag{84}$$

Here, for an $N \times N$ matrix A, $|A|$ denotes the norm $|A| \triangleq \sup_{|x|=1} |Ax|$, where $|x|$ is the Euclidean norm of any R^N-vector x. By Holder's inequality and condition (C) it is easily confirmed that $E^{1/2}(\beta_{j,k}) \le K_1 < \infty$ for all $j \ge 1$ and $1 \le k \le s(j)$. Thus, by (82) and Holder's inequality,

$$E(x_j) \le K_2 \sum_{k=1}^{s(j)} \mu_{j-k} E^{1/2}\{|C_{j,k}|^2\}. \tag{85}$$

Now for any $\mu > 0$ and symmetric positive semidefinite matrix U, it follows from standard properties of the above matrix norm that $|I - \mu U| \leq 1 + \mu^2 |U|^2$, so from (83),

$$|C_{j,k}| \leq \prod_{i=1}^{k-1} (1 + \mu_{j-i}^2 |U_{j-i}|^2). \tag{86}$$

Thus, by (86) and an argument identical to that used to obtain line (24) in Watanabe [3], we can show that there is some constant K_3 such that

$$E(|C_{j,k}|^2) \leq K_3 \tag{87}$$

for all $j \geq 1$ and $1 \leq k \leq j$. From (87) and (85) there is a constant K_4 such that

$$E(x_j) \leq K_4 j^{b-1} \tag{88}$$

for all $j \geq 1$. Exactly as in Step II it follows that for each $0 < b < \frac{1}{2}$ there is some $\eta > 0$ such that $\sum j^{-(1/2-\eta)} x_j < \infty$ almost surely; hence (81) follows from Kronecker's lemma.

Finally, putting together (81), (66), (49), and (46) we obtain (45). □

LEMMA 4.3. *Assume conditions* (A*), (B), *and* (C*). *Then there exists some constant* $\eta > 0$ *such that* (45) *holds almost surely* $[P]$.

Proof. The proof is similar to that of Lemma 4.2; hence we indicate only the changes in that proof. Fix $0 < b < 1$ and define $s(j) \triangleq [j^b], U_j \triangleq X_j X_j^T$, and $\mu_j \triangleq \mu/j$.

Step I. Here is it shown that for each $b > \frac{1}{6}$ there is some $\eta > 0$ such that (49) holds almost surely $[P]$. It is enough to prove (51) where μ_j is defined by (50). Clearly,

$$\begin{aligned} E\left[\sum_{j=a+1}^{a+n} u_j\right]^2 \leq\ & 2E\left[\sum_{\substack{i,j=a+1 \\ j-M \leq i \leq j}}^{a+n} u_i u_j\right] + 2E\left[\sum_{\substack{i,j=a+1 \\ a+1 \leq i \leq j-s(j)}}^{a+n} u_i u_j\right] \\ & + 2E\left[\sum_{\substack{i,j=a+1 \\ j-s(j) < i < j-M}}^{a+n} u_i u_j\right] \end{aligned} \tag{89}$$

for all $a \geq 1 + M^{1/b}$ and $n \geq 1$, since $j > a$ implies that $s(j) > M$. Consider the first term on the right of (89). Clearly,

$$\{r_1 M - (2r)(2M+1)\}(r_2 - 4(2r)) = 4(2r)^2(2M+1) \tag{90}$$

for $r = 2$, where $r_1 \triangleq 24$ and $r_2 \triangleq 24(1 + 1/(4M-1))$, thus Theorem 5.5 gives

$$E|\nu_{j-s(j)}|^4 \leq K_1 (j - s(j))^{-1} \leq K_2 j^{-1} \quad \text{for all } j \geq 2. \tag{91}$$

By Holder's inequality and condition (C*),

$$E(u_i u_j) \leq K_3 (ij)^{-(1/2-\eta)} \|\nu_{i-s(i)}\|_4 \|\nu_{j-s(j)}\|_4 \tag{92}$$

and hence, from (91),

$$E\left[\sum_{\substack{i,j=a+1 \\ j-M \leq i \leq j}}^{a+n} u_i u_j\right] \leq K_4 \left[\sum_{j=a+1}^{a+n} j^{-(3/2-2\eta)}\right] \tag{93}$$

for all $a \geq 1 + M^{1/b}, n \geq 1$. Next bound the second term on the right of (89). To lighten the notation we write $\psi(m)$ for $\psi_{(a,X)}(m)$. Since $a+1 \leq i \leq j - s(j)$ and $E(U_j - R) = 0$, it follows from Theorem 5.1(ii), (A*), and (91) that

$$\begin{aligned} E(u_i u_j) &\leq (ij)^{-(1/2-\eta)}\psi(s(j))E|U_j - R|E|(U_i - R)\nu_{i-s(i)}\nu_{j-s(j)}| \\ &\leq K_5(ij)^{-(1/2-\eta)}\psi(s(j))\|\nu_{i-s(i)}\|_4\|\nu_{j-s(j)}\|_4 \leq K_6\psi(s(j))(ij)^{-(3/4-\eta)}, \end{aligned} \tag{94}$$

and so

$$E\left[\sum_{\substack{i,j=a+1 \\ a+1\leq i\leq j-s(j)}}^{a+n} u_i u_j\right] \leq K_6 \sum_{j=a+1}^{a+n} j^{(2\eta - b\beta - 1/2)} \leq K_6 \sum_{j=a+1}^{a+n} j^{-(\theta-2\eta)} \tag{95}$$

for all $a \geq 1 + M^{1/b}, n \geq 1$, and some constant $\theta > 1$, where the second inequality follows since $b > \frac{1}{6}$ implies that $b\beta > \frac{1}{2}$ (to get the first inequality in (95), note that $j > a \geq 1 + M^{1/b}$ implies $s(j) > M$; hence by (A*) we obtain $\psi(s(j)) \leq K_7 j^{-b\beta}$). Now bound the third term on the right of (89). Since $j - s(j) < i < j - M$ and $E(U_j - R) = 0$ from Theorem 5.1(ii) it follows that

$$E(u_i u_j) \leq K_8(ij)^{-(1/2-\eta)}\psi(j-i)\|\nu_{i-s(i)}\|_4\|\nu_{j-s(j)}\|_4 \leq K_8 i^{-((3/2)-2\eta)}\psi(j-i), \tag{96}$$

and hence

$$E\left[\sum_{\substack{i,j=a+1 \\ j-s(j)<i<j-M}}^{a+n} u_i u_j\right] \leq K_9 \sum_{k=M+1}^{n-1} \sum_{i=a+1}^{a+n-k} \psi(k) i^{-(3/2-2\eta)} \leq K_{10} \sum_{j=a+1}^{a+n} j^{-(3/2-2\eta)} \tag{97}$$

since $j - i > M$. From (97), (95), (93), (89), and Theorem 4.3 we obtain (49).

Step II. In this step it is shown that for each $0 < b < \frac{1}{2}$ there is a constant $\eta > 0$ such that (66) holds almost surely $[P]$, where η_j is given by (47). As in the proof of Lemma 4.2, it is enough to prove (67). Defining M_j as in (69) we see that (70) continues to hold (but using Theorem 4.5 instead of Theorem 4.4), and hence the proof reduces to finding an upper bound similar to (80) but based on the use of (A*) and (C*). Fix any $0 < b < \frac{1}{2}$. By Lemma 5.7 and conditions (A*) and (C*) there is some constant $\gamma > 0$ such that

$$E\left[\prod_{l=1}^{k} |U_j - R|^2 |U_{j-i_l}|^2\right] \leq \gamma^{1+k} \tag{98}$$

for all $j \geq 2, k \geq 1$, and indices $1 \leq i_1 < i_2 < \cdots < i_k \leq s(j)$. By (98), (71), (69), the Cauchy–Schwarz inequality, and some simple manipulations,

$$E(|M_j|^2) \leq \gamma \left[\sum_{k=1}^{s(j)} \left\{ \sum_{1\leq i_1<\cdots i_k\leq s(j)} (\mu_{j-i_1}\cdots\mu_{j-i_k})\gamma^{k/2} \right\}\right]^2. \tag{99}$$

Clearly, there is a constant K_1 (depending only on b) such that $\mu_{j-i} \leq K_1 j^{-1}$ for all $1 \leq i \leq s(j)$ and $j \geq 2$, so

$$\left\{ \sum_{1\leq i_1<\cdots i_k\leq s(j)} (\mu_{j-i_1}\cdots\mu_{j-i_k})\gamma^{k/2} \right\} \leq \binom{s(j)}{k} \frac{C^k}{j^k} \tag{100}$$

for $j \geq 2$, where $C \triangleq K_1\gamma^{1/2}$. To bound $(C/j)^k$, fix a constant $1 > \delta > 2b$, define $p(j) \triangleq j^{-\delta/2}$, and put

$$\lambda_j \triangleq -\frac{1}{p(j)}\log(1-p(j)). \tag{101}$$

Clearly, $\lim_{j\to\infty}\lambda_j = 1$, and since $b < \delta/2$, there exists some J_1 such that

$$\frac{1}{s(j)}\log(C^{-1}j^{\delta/2}) \geq \lambda_j p(j) \quad \text{for all } j \geq J_1. \tag{102}$$

Thus, from (101) and (102), $(Cp(j))^k \leq (1-p(j))^{s(j)-k}$ for all $1 \leq k \leq s(j)$ and $j \geq J_1$, and, therefore,

$$\frac{C^k}{j^k} \leq j^{\delta-1}(p(j))^k(Cp(j))^k \leq j^{\delta-1}(p(j))^k(1-p(j))^{s(j)-k} \tag{103}$$

for all $1 \leq k \leq s(j)$ and $j \geq J_1$. In view of (100), (103), and the binomial theorem,

$$\sum_{k=1}^{s(j)}\left\{\sum_{1\leq i_1<\cdots i_k\leq s(j)}(\mu_{j-i_1}\cdots\mu_{j-i_k})\gamma^{k/2}\right\} \leq j^{\delta-1} \tag{104}$$

for all $j \geq J_1$. Thus, by (99),

$$E(|M_j|^2) \leq \gamma j^{2\delta-2} \tag{105}$$

for all $j \geq J_1$. Since $0 < \delta < 1$ we can choose $\eta > 0$ such that $2-\delta-\eta > 1$ and (67) follows by (105) and (70).

Step III. Here it is shown that for each $0 < b < \frac{1}{2}$ there is some $\eta > 0$ such that (81) holds almost surely $[P]$. The argument is identical to that used in Step III in the proof of Lemma 4.2 except that to get (87) under conditions (A*) and (C*) we follow an argument exactly like that used to obtain line (43) in Watanabe [3]. □

Lemmas 4.4 and 4.5 establish properties of the function F (defined in (20)) that are needed for the proof of Proposition 2.2. Because (B) in §2 does not ensure that $(\mu R - 2I)$ is positive definite (it may have some strictly negative eigenvalues) it is necessary to check that the integrals in (20) and (28) are well defined for $\phi \in S$, and to justify the integration by parts in going from (20) to (28).

LEMMA 4.4. *Assume condition* (B) *in* §2. *Then the following hold.*

(a) *For each* $\phi \in S, S$ *being defined by* (18), *the functions* $t \to t^{(\mu R-I)}\dot{\phi}(t)$ *and* $t \to t^{(\mu R-2I)}\phi(t)$ *are integrable over* $0 < t \leq 1$; *in particular,* $F(\phi)$ *is well defined, where* F *is given by* (20).

(b) *For each* $\phi \in S$, *the following integration-by-parts formula holds*:

$$\int_0^1 \tau^{(\mu R-I)}\dot{\phi}(\tau)d\tau = \phi(1) - (\mu R - I)\int_0^1 \tau^{(\mu R-2I)}\phi(\tau)d\tau. \tag{107}$$

Proof. It is sufficient to carry out the proof in the special case where $N = 1$; in the general case we simply diagonalize the symmetric matrix μR and carry out the same argument for each element along the main diagonal. For the proof we need the following simple consequence of the Cauchy–Schwarz inequality (see Natanson [17, Thm. 7, p. 257]): If $\phi \in C_0^N[0,1]$ is such that $\phi \in L_2^N[0,1]$, then

$$\sum_{i=0}^{n-1}\frac{|\phi(t_{i+1})-\phi(t_i)|^2}{|t_{i+1}-t_i|} \leq \int_0^1 |\dot{\phi}(\tau)|^2 d\tau \tag{108}$$

for each partition $0 = t_0 < t_1 < \cdots < t_n = 1$.

(a) Since $2\mu R - 1 > 0$, the function $t \to t^{(\mu R - 1)}$ is square integrable over the interval $0 < t \leq 1$, whence integrability of the first function in (a) follows from the Cauchy–Schwarz inequality. For integrability of the second function in (a), observe that

$$\int_0^1 \tau^{(\mu R-2)}\phi(\tau)d\tau = \int_0^1 \tau^{(\mu R-3/2)}\tau^{-1/2}\phi(\tau)d\tau. \tag{109}$$

Moreover, by (108) and the fact that $\phi(0) = 0$,

$$t^{-1/2}|\phi(t)| \leq \left(\int_0^1 |\dot{\phi}(\tau)|^2 d\tau\right)^{1/2} < \infty, \tag{110}$$

for all $0 < t \leq 1$, and (109) and (110) along with the integrability of $t \to t^{(\mu R-3/2)}$ over $0 < t \leq 1$ (which follows from the fact that $2\mu R - 1 > 0$) imply that the second function in (a) is integrable.

(b) It is enough to show that $\varepsilon^{(\mu R-1)}\phi(\varepsilon) \to 0$ as $\varepsilon \to 0$, for (107) follows upon integration by parts over the interval $(\varepsilon, 1], \varepsilon > 0$, and then taking $\epsilon \to 0$. However, the required limit follows from (110) and the fact that $\mu R - \frac{1}{2} > 0$. □

Lemma 4.5 is used to prove Proposition 2.2 (see (21)).

LEMMA 4.5. *Assume conditions* (B) *and* (D) *in* §2, *let* $\{\tilde{W}(t)\}$ *be a Brownian motion with covariance matrix* $\mu^2\Gamma$ *defined on a probability space* $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$, *and let* $\tilde{Y}(t)$ *be defined in terms of* $\tilde{W}(t)$ *by* (4). *Then for almost all* $\tilde{\omega}[\tilde{P}]$,

$$\{(2j\,\mathrm{loglog}(j))^{-1/2}\tilde{Y}(j,\tilde{\omega}), j \geq J\} \to\to F[S].$$

Proof. For $0 \leq t \leq 1, \tilde{\omega} \in \tilde{\Omega}$, and integer $j \geq J$, define

$$\tilde{\phi}_j(t,\tilde{\omega}) \triangleq (2j\,\mathrm{loglog}(j))^{-1/2}\tilde{W}(jt,\tilde{\omega}).$$

According to Theorem 5.6, the set S defined in (18) is a compact subset of $C_0^N[0,1]$ and, for almost all $\tilde{\omega}[\tilde{P}]$,

$$\{\tilde{\phi}_j(\cdot,\tilde{\omega}), j \geq J\} \to\to S. \tag{111}$$

Fix some $\tilde{\omega}$ in the set of $\tilde{P}$-unit measure such that (111) holds and the set of functions

$$t \to t^{(\mu R-2I)}\tilde{\phi}_j(t,\tilde{\omega}), \qquad j \geq J$$

has equi-absolutely continuous integrals over $0 < t \leq 1$ (see Lemma 4.6), and fix an arbitrary ϕ in S. If $\{j_r\}$ is a subsequence such that $\{\tilde{\phi}_{j_r}(\cdot,\tilde{\omega})\}$ converges in $C_0^N[0,1]$ to ϕ then the Vitali convergence theorem (Natanson [17, Thm. 2, p. 152]) along with the definition of F in (20) and the above equi-absolute integrability shows that $\{F(\tilde{\phi}_{j_r}(\cdot,\tilde{\omega}))\}$ converges to $F(\phi)$ as $r \to \infty$. From (111) it is thus clear that $F(\phi)$ is an accumulation point of the sequence $\{F(\tilde{\phi}_j(\cdot,\tilde{\omega})), j \geq 1\}$, whence, by the arbitrary choice of $\phi \in S$, it follows that $F[S] \subset C(\{F(\tilde{\phi}_j(\cdot,\tilde{\omega}))\})$. To establish the opposite inclusion, observe that compactness of S along with the fact (see (111)) that $\rho(\tilde{\phi}_j(\cdot,\tilde{\omega}), S) \to 0$ (where $\rho(\tilde{\phi}_j(\cdot,\tilde{\omega}), S)$ denotes the distance from $\tilde{\phi}_j(\cdot,\tilde{\omega})$ to S in the metric $\rho(\cdot,\cdot)$ of uniform convergence in $C_0^N[0,1]$) implies that $\{\tilde{\phi}_j(\cdot,\tilde{\omega})\}$ is a totally bounded subset of $C_0^N[0,1]$. If $z \in C(\{F(\tilde{\phi}_j(\cdot,\tilde{\omega}))\})$, then there is a subsequence $\{F(\tilde{\phi}_{j_r}(\cdot,\tilde{\omega}))\}$ of $\{F(\tilde{\phi}_j(\cdot,\tilde{\omega}))\}$ that converges to z, and by the total boundedness of $\{\tilde{\phi}_j(\cdot,\tilde{\omega})\}$, there is a further subsequence

$\{\tilde{\phi}_{j_l}(\cdot,\tilde{\omega})\}$ of $\{\tilde{\phi}_{j_r}(\cdot,\tilde{\omega})\}$ that converges uniformly to some $\phi \in S$. Thus the sequence of functions $t \to t^{(\mu R-2)}\tilde{\phi}_{j_l}(t,\tilde{\omega}), 0 < t \leq 1$, has equi-absolutely continuous integrals and converges pointwise in t to the limiting function $t \to t^{(\mu R-2)}\phi(t)$. By Vitali's theorem, $F(\phi_{j_l}(\cdot,\tilde{\omega})) \to F(\phi)$, whence $z = F(\phi) \in F[S]$ and therefore $C(\{F(\tilde{\phi}_j(\cdot,\tilde{\omega}))\}) \subset F[S]$. Thus $\{F(\tilde{\phi}_j(\cdot,\tilde{\omega})), j \geq J\} \to\to F[S]$; since $F(\tilde{\phi}_j(\cdot,\tilde{\omega})) = (2j\,\mathrm{loglog}(j))^{-1/2}\tilde{Y}(j,\tilde{\omega})$ (see (4) and (20)), this establishes Lemma 4.5. □

The next result has substance when matrix $\mu R - 2I$ has negative eigenvalues. It relies on the fact that $2\mu R - I$ is positive definite and is used to prove Lemma 4.5.

LEMMA 4.6. *Assume conditions* (B) and (D) *in* §2 *and let* $\{\tilde{W}(t)\}$ *be a Brownian motion with covariance matrix* $\mu^2\Gamma$ *defined on a probability space* $(\tilde{\Omega}, \tilde{\mathbf{A}}, \tilde{P})$. *Then for almost all* $\tilde{\omega}[\tilde{P}]$, *the set of functions*

$$t \to t^{(\mu R-2I)}(2j\,\mathrm{loglog}(j))^{-1/2}\tilde{W}(jt,\tilde{\omega}), \qquad j \geq J,$$

has equi-absolutely continuous integrals over $0 < t \leq 1$ (see Natanson [17, pp. 151–152]).

Proof. As in Lemma 4.4, it is sufficient to carry out the proof in the special case where $N = 1$. For ease of notation, let $\sigma^2 \triangleq E(\tilde{W}^2(1))$ and fix a constant $c > \sigma^2$. By the local law of the iterated logarithm for Brownian motion (Loeve [18, §41.3B, p. 249]) there exists, for almost all $\tilde{\omega}[\tilde{P}]$, some (small) $\delta(c,\tilde{\omega})$ such that $1 > \delta(c,\tilde{\omega}) > 0$ and

$$|(2jt\,\mathrm{loglog}(1/jt))^{-1/2}\tilde{W}(jt,\tilde{\omega})| < c \tag{112}$$

for $0 < jt < \delta(c,\tilde{\omega})$. Likewise, by the asymptotic law of the iterated logarithm for Brownian motion (Loeve [18, p. 249]) there exists, for almost all $\tilde{\omega}[\tilde{P}]$, some (large) $\Delta(c,\tilde{\omega}) > 1$ such that

$$|(2jt\,\mathrm{loglog}(jt))^{-1/2}\tilde{W}(jt,\tilde{\omega})| < c \tag{113}$$

for $jt > \Delta(c,\tilde{\omega})$. Fix some arbitrary Borel subset $E \subset (0,1]$, and define the sets

$$\begin{aligned} I_1(c,\tilde{\omega},j) &\triangleq E \cap \{\tau; 0 < j\tau < \delta(c,\tilde{\omega})\}, \\ I_2(c,\tilde{\omega},j) &\triangleq E \cap \{\tau; j\tau > \Delta(c,\tilde{\omega})\}, \\ I_3(c,\tilde{\omega},j) &\triangleq E \cap \{\tau; \delta(c,\tilde{\omega}) \leq j\tau \leq \Delta(c,\tilde{\omega})\}. \end{aligned}$$

For fixed $j \geq J$,

$$\begin{aligned} &\int_{I_1(c,\tilde{\omega},j)} \tau^{(\mu R-2)}(2j\,\mathrm{loglog}(j))^{-1/2}|\tilde{W}(j\tau,\tilde{\omega})|d\tau \\ &\quad \leq \int_{I_1(c,\tilde{\omega},j)} \tau^{(\mu R-3/2)}(j\tau)^{-1/2}|\tilde{W}(j\tau,\tilde{\omega})|d\tau \qquad (\mathrm{loglog}(j) \geq 1 \text{ for } j \geq J) \\ &\quad \leq c\int_{I_1(c,\tilde{\omega},j)} \tau^{(\mu R-3/2)}(2\,\mathrm{loglog}(1/j\tau))^{1/2}d\tau \qquad (\text{by (112)}) \\ &\quad \leq c\int_{I_1(c,\tilde{\omega},j)} \tau^{\alpha-1}\tau^{\gamma}(2\,\mathrm{loglog}(1/j\tau))^{1/2}d\tau \end{aligned} \tag{114}$$

(for constants $\alpha, \gamma > 0$, since $\mu R > \frac{1}{2}$),

$$\leq c_1 j^{-\gamma}\int_E \tau^{\alpha-1}d\tau \quad \text{a.s. } [\tilde{P}]$$

for some constant $c_1 > 0$ that depends only on c and $\tilde{\omega}$. The last inequality follows from the fact that

$$\sup_{\tau \in I_1(c,\tilde{\omega},j)} \tau^{\gamma}(2\operatorname{loglog}(1/j\tau))^{1/2} \leq \sup_{x \geq \delta^{-1}(c,\tilde{\omega})} (jx)^{-\gamma}(2\operatorname{loglog} x)^{1/2} = O(j^{-\gamma}),$$

where the constant implied by O depends only on c and $\tilde{\omega}$. Since $\tau \leq 1$ for all τ in E,

$$\begin{aligned} &\int_{I_2(c,\tilde{\omega},j)} \tau^{(\mu R-2)}(2j\operatorname{loglog}(j))^{-1/2}|\tilde{W}(j\tau,\tilde{\omega})|d\tau \\ &\quad \leq \int_{I_2(c,\tilde{\omega},j)} \tau^{(\mu R-2)}(2j\operatorname{loglog}(j\tau))^{-1/2}|\tilde{W}(j\tau,\tilde{\omega})|d\tau \\ &\quad = \int_{I_2(c,\tilde{\omega},j)} \tau^{(\mu R-3/2)}(2j\tau\operatorname{loglog}(j\tau))^{-1/2}|\tilde{W}(j\tau,\tilde{\omega})|d\tau \\ &\quad \leq c\int_E \tau^{(\mu R-3/2)}d\tau \quad \text{a.s. } [\tilde{P}] \end{aligned} \tag{115}$$

(by (113)). Finally, there exists some finite bound $B(c,\tilde{\omega})$ such that $(2jt)^{-1/2}|\tilde{W}(jt,\tilde{\omega})| < B(c,\tilde{\omega})$ for $\delta(c,\tilde{w}) \leq jt \leq \Delta(c,\tilde{\omega})$, whence, for $j \geq J$, it is easily seen that, almost surely $[P]$,

$$\int_{I_3(c,\tilde{\omega},j)} \tau^{(\mu R-2)}(2j\operatorname{loglog}(j))^{-1/2}|\tilde{W}(j\tau,\tilde{\omega})|d\tau \leq B(c,\tilde{\omega})\int_E \tau^{(\mu R-3/2)}d\tau. \tag{116}$$

By (114)–(116), for almost all $\tilde{\omega}[\tilde{P}]$,

$$\begin{aligned} &\int_E \tau^{(\mu R-2)}(2j\operatorname{loglog}(j))^{-1/2}|\tilde{W}(j\tau,\tilde{\omega})|d\tau \\ &\quad \leq c_1\int_E \tau^{\alpha-1}d\tau + (B(c,\tilde{\omega})+c)\int_E \tau^{(\mu R-3/2)}d\tau \end{aligned} \tag{117}$$

for all $j \geq J$ and all Borel sets $E \subset (0,1]$. Since $\alpha > 0$ and $\mu R > \frac{1}{2}$, Lemma 4.6 follows from (117). □

5. Appendix 2. In this appendix, we collect for convenience several results from probability theory that are used in the proofs of §§3 and 4.

A useful property of random variables needed to prove Lemmas 4.2 and 4.3 (see lines (56), (60), (94), and (96)) is given by Theorem 5.1.

THEOREM 5.1. *Suppose that* **G** *and* **H** *are two sub* σ*-algebras in a probability space* $(\Omega, \mathbf{A}, P)$, *and* X *and* Y *are random variables defined on* Ω *that are measurable with respect to* **G** *and* **H**, *respectively.*

(i) *If* $r, s, t > 1$ *are constants such that* $(1/r) + (1/s) + (1/t) = 1$, *and if* $\|X\|_s < \infty$ *and* $\|Y\|_t < \infty$, *then*

$$|E(XY) - (EX)(EY)| \leq 10\{\alpha(\mathbf{G}, \mathbf{H})\}^{1/r}\|X\|_s\|Y\|_t,$$

where $\alpha(\mathbf{G}, \mathbf{H}) \triangleq \sup|P(AB) - P(A)P(B)|$ *and the supremum is taken over all* $A \in \mathbf{G}$, $B \in \mathbf{H}$;

(ii) *If* $E|X| < \infty$, *and* $E|Y| < \infty$, *then*

$$|E(XY) - (EX)(EY)| \leq \{\psi(\mathbf{G}, \mathbf{H})\}E|X|E|Y|,$$

where $\psi(\mathbf{G}, \mathbf{H}) \triangleq \sup|P(AB) - P(A)P(B)|/P(A)P(B)$ *and the supremum is taken over all* $A \in \mathbf{G}, B \in \mathbf{H}$.

Part (i) of Theorem 5.1 is due to Davydov [19] and a full proof is given in Lemma 1 of Deo [20]. The proof of part (ii) is almost trivial (see, e.g., Watanabe [3, Lem. 2]).

The next result is used in the proof of Lemma 4.1 (see (39), (40)).

THEOREM 5.2. *Suppose that* $\{\xi_n, n \geq 1\}$ *is a strictly stationary sequence of* R^D*-valued zero-mean random vectors defined on a common probability space and having finite third-order moments. If* $\{\xi_n\}$ *is strong mixing with* $\alpha_\xi(m) = O(m^{-\beta})$ *for some* $\beta > 3$, *then* (i) *the series*

$$T \triangleq E(\xi_1 \xi_1^T) + \sum_{j=2}^{\infty} E(\xi_1 \xi_j^T) + \sum_{j=2}^{\infty} E(\xi_j \xi_1^T)$$

converge absolutely, and (ii) *there exists some constant* $\eta > 0$ *along with a probability space* $(\Omega^{(1)}, \mathbf{A}^{(1)}, P^{(1)})$ *on which is defined an* R^D*-valued Brownian motion* $\{B^{(1)}(t)\}$ *with covariance matrix T and a sequence* $\{\xi_n^{(1)}, n \geq 1\}$ *of* R^D*-valued random vectors such that*

(i) $\{\xi_n, n \geq 1\} \stackrel{D}{=} \{\xi_n^{(1)}, n \geq 1\}$ *and*

(ii) $\sum_{1 \leq n \leq t} \xi_n^{(1)} = B^{(1)}(t) + O(t^{1/2-\eta})$ *for all* $t \geq 1$ *a.s.* $[P^{(1)}]$.

Theorem 5.2 is a special case of [21, Thm. 4] in Kuelbs and Philipp, and is a generalization to dependent random vectors of the strong invariance principle for independent and identically distributed random variables first established by Strassen [5, Thm. 2].

The next theorem is a criterion for almost sure convergence of partial sums of random variables (see Stout [22, Thm. 2.4.2]). It is used to prove Lemmas 4.2 and 4.3 (see (65)).

THEOREM 5.3. *Suppose that* $\{\xi_k, k \geq 1\}$ *is a sequence of random variables defined on* $(\Omega, \mathbf{A}, P)$, *and let* $g(i,j)$ *and* $h(i,j)$ *be nonnegative functions whose arguments* i *and* j *are nonnegative integers, such that for all integers* $a \geq 0$ *and* $k, n > 1$:

(i) $g(a,k) + g(a+k,n) \leq g(a,k+n)$;

(ii) $h(a,k) + h(a+k,n) \leq h(a,k+n)$;

(iii) $h(a,n) \leq K < \infty$ *and* $Kh(a,n) \geq g(a,n)\log^2(a+1)$ *for some constant* K;

(iv) $$E\left[\sum_{k=a+1}^{a+n} \xi_k\right]^2 \leq g(a,n).$$

Then $\sum_{k=1}^{\infty} \xi_k < \infty$ *almost surely.*

The next theorem is a variant of a result of Watanabe [3] concerning $2r$th mean convergence of the recursion (1). It is used to prove Lemma 4.2 (see (54), (58), and (70)). (Recently Gerencser [23] has established L_q convergence for algorithms more general than (1). These entail use of a projection onto a neighbourhood of the limit, and [23] gives the first rigorous proofs of this commonly used mechanism.)

THEOREM 5.4. *Consider the recursion* (1), *where* (i) $\{(a_j, X_j^T), j \geq 1\}$ *is a strictly stationary strong mixing process with* $\alpha_{(a,X)}(m) = O(m^{-\beta_1})$ *for some* $\beta_1 > 0$; (ii) *there is a constant C such that, for each* $j, E|X_j|^{4k} = O(C^k k!)$ *for all integers* $k \geq 1$; (iii) *the matrix* $E(X_j X_j^T)$ *is positive definite, and* (iv) *for some* $\varepsilon > 0$ *and positive integer r we have* $E|a_j|^{8r+5+\varepsilon} \leq K_1 < \infty$ *for all j. If* λ *denotes the least eigenvalue of* $E(X_j X_j^T)$ *and* $\min\{2\mu\lambda, 2r\beta_1/(4r+1)\} > 1$, *then* $E|h_j - h_*|^{2r} = O(j^{-1})$.

Watanabe [3, Thm. 3] obtains the above result in the case where the driving process $\{(a_j, X_j^T)\}$ is uniform (or ϕ-) mixing. However, the proof given there adapts fairly easily to the case for which $\{(a_j, X_j^T)\}$ is strong mixing where, in place of the classical Ibragimov bound for uniform mixing processes (see [3, Lem. 1]), we use Theorem 5.1(i).

In the case of ψ-mixing driving data we have the following result on $2r$th mean convergence, which is Theorem 4 in Watanabe [3]. We must prove Lemma 4.3 (see (91)).

THEOREM 5.5. *Consider the recursion* (1), *where* (i) $E|X_j|^{4r_1M} < \infty$ and $E|a_j|^{r_2} < \infty$ *for positive integers* r_1, M, *and a positive constant* r_2; (ii) $\{(a_j, X_j^T)\}$ *is* ψ*-mixing and, for positive constants* $\beta_2, K, \psi_{(a,X)}(m) \leq Km^{-\beta_2}$ *for all* $m \geq M$; *and* (iii) *for some positive integer* r, *we have* $r_1M \geq 12r, r_2 > 8r$, *and*

$$\{r_1M - (2r)(2M+1)\}(r_2 - 8r) \geq 16r^2(2M+1).$$

If λ *is the least eigenvalue of* $E(X_jX_j^T)$ *and* $\min\{2\mu\lambda, \beta_2\} > 1$, *then* $E|h_j - h_*|^{2r} = O(j^{-1})$.

For the proof of Lemma 4.5, we need the functional law of the iterated logarithm for Brownian motion due to Strassen [5; Thm. 1].

THEOREM 5.6. *Suppose that* $\{W(t)\}$ *is an* R^N*-valued Brownian motion defined on* $(\Omega, \boldsymbol{A}, P)$, *with covariance matrix* $\mu^2\Gamma$ (*where* μ *is a scalar*). *Then the set* S *defined in* (18) *is a compact subset of* $C_0^N[0,1]$, *and for almost all* $\omega[P]$,

$$\{\phi_j(\cdot,\omega), j \geq J\} \longrightarrow\!\!\!\!\!\rightarrow S,$$

where $\phi_j(t,\omega) \stackrel{\Delta}{=} (2j\,\mathrm{loglog}(j))^{-1/2}W(jt,\omega)$ *for* $0 \leq t \leq 1$.

Lemma 5.7, which is implicit in the developments of Watanabe [3, p. 138], gives a useful property of ψ-mixing, and is needed to prove Lemma 4.3 (see (98)).

LEMMA 5.7. *Suppose that* $\{\xi_n\}$ *is a real-valued random process such that* $\psi_\xi(M) < \infty$ *and* $B \stackrel{\Delta}{=} \sup_n E|\xi_n|^M < \infty$ *for some positive integer* M. *Then there is a constant* $\gamma > 0$ *depending only on* $\psi_\xi(M)$ *and* B *such that*

$$E\left[\prod_{l=1}^{n}|\xi_l|\right] \leq \gamma^l \quad \textit{for all } n \geq 1.$$

Acknowledgment. The author thanks a reviewer for the proof of Lemma 4.1, which is simpler than the original.

REFERENCES

[1] E. EWEDA AND O. MACCHI, *Convergence of an adaptive linear estimation algorithm*, IEEE Trans. Automat. Control, 29 (1984), pp. 119–127.

[2] B. WIDROW AND S. D. STEARNS, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[3] M. WATANABE, *The 2rth mean convergence of adaptive filters with stationary dependent random variables*, IEEE Trans. Inform. Theory, 30 (1984), pp. 134–140.

[4] A. BENVENISTE, M. METIVIER, AND P. PRIOURET, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.

[5] V. STRASSEN, *An invariance principle for the law of the iterated logarithm*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 3 (1964), pp. 211–226.

[6] R. BRADLEY, *Basic properties of strong mixing conditions*, in Dependence in Probability and Statistics—Survey of Recent Results, Oberwolfach 1985, Ernst Eberlein and Murad S. Taqqu, eds., Birkhauser, Boston, 1986, pp. 165–192.

[7] R. R. BITMEAD, *Convergence properties of* LMS *adaptive estimators with unbounded dependent inputs*, IEEE Trans. Automat. Control, 29 (1984), pp. 477–479.

[8] L. GERENCSER, *Strong approximation results in estimation and adaptive control*, in Topics in Stochastic System Modeling, Estimation and Control, L. Gerencser and P. E. Caines, eds., Lecture Notes in Control and Information Sciences, v. 161, Springer, 1991.

[9] J. KUELBS, *Kolmogorov's law of the iterated logarithm for Banach space valued random variables*, Illinois J. Math., 21 (1977), pp. 784–800.

[10] E. BERGER, *Asymptotic behaviour of a class of stochastic approximation procedures*, Probab. Theory Related Fields, 71 (1986), pp. 517–552.

[11] D. Ruppert, *Almost sure approximations to the Robbins–Monro and Kiefer–Wolfowitz processes with dependent noise*, Ann. Probability, 10 (1982), pp. 178–187.
[12] G. Mark, *Loglog-invarianzprinzipien für Prozesse der stochastischen Approximation*, Mit. Math. Sem. Giessen, 153, Justus Liebig Universitat, Giessen, Germany, 1982.
[13] A. Friedman, *Foundations of Modern Analysis*, Dover, New York, 1982.
[14] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1978.
[15] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.
[16] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd ed., John Wiley, New York, 1968.
[17] I. P. Natanson, *Theory of Functions of a Real Variable*, Vol. 1, Ungar, New York, 1955.
[18] M. Loeve, *Probability Theory*, Vol. II, 4th ed., Springer-Verlag, New York, 1977.
[19] Yu. A. Davydov, *The invariance principle for stationary processes*, Theory Probability Appl., 15 (1970), pp. 487–498.
[20] C. Deo, *A note on empirical processes of strong mixing processes*, Ann. Probability, 1 (1973), pp. 870–875.
[21] J. Kuelbs and W. Philipp, *Almost sure invariance principles for partial sums of mixing B-valued random variables*, Ann. Probability, 8 (1980), pp. 1003–1036.
[22] W. F. Stout, *Almost Sure Convergence*, Academic Press, New York, 1974.
[23] L. Gerencser, *Rate of convergence of recursive estimators*, SIAM J. Control Optim., 30 (1992), pp. 1200–1227.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 140–153, January 1994

008

SHARP LIPSCHITZ CONSTANTS FOR BASIC OPTIMAL SOLUTIONS AND BASIC FEASIBLE SOLUTIONS OF LINEAR PROGRAMS*

WU LI†

Abstract. The main purpose of this paper is to give Lipschitz constants for basic optimal solutions (or vertices of solution sets) and basic feasible solutions (or vertices of feasible sets) of linear programs with respect to right-hand side perturbations. The Lipschitz constants are given in terms of norms of pseudoinverses of submatrices of the matrices involved, and are sharp under very general assumptions. There are two mathematical principles involved in deriving the Lipschitz constants: (1) the local upper Lipschitz constant of a Hausdorff lower semicontinuous mapping is equal to the Lipschitz constant of the mapping and (2) the Lipschitz constant of a finite-set-valued mapping can be inherited by its continuous submappings. Moreover, it is proved that any Lipschitz constant for basic feasible solutions can be used as an Lipschitz constant for basic optimal solutions, feasible solutions, and optimal solutions.

Key words. sharp Lipschitz constant, linear program, (basic) optimal solution, (basic) feasible solution, lower semicontinuous mapping, locally upper Lipschitz continuous mapping, pseudoinverse

AMS subject classifications. primary 90C31; secondary 90C05, 65F20, 54C60

1. Introduction.

Consider the following linear programming problem:

$$c_{\min}(b,d) := \min\{c^T x : Ax \le b, Cx = d\}, \tag{1.1}$$

where A is an $m \times n$ matrix, C a $k \times n$ matrix, $c \in \mathbb{R}^n, b \in \mathbb{R}^m$, and $d \in \mathbb{R}^k$. Let $F\binom{b}{d} := \{x \in \mathbb{R}^n : Ax \le b, Cx = d\}$ denote the feasible set of (1.1) and $S\binom{b}{d} := \{x \in \mathbb{R}^n : Ax \le b, Cx = d, c^T x = c_{\min}(b,d)\}$ the solution set of (1.1). If an optimal solution is primal degenerate, then $S\binom{b}{d}$ contains more than one solution and is a convex polyhedral set [8]. In general, F and S are set-valued mappings (or multifunctions). It was first shown by Hoffman [11] that F is Lipschitz continuous. That is, there exists a scalar $\gamma > 0$, such that

$$H\left(F\binom{b}{d}, F\binom{b'}{d'}\right)_\nu \le \gamma \cdot \left\|\binom{b}{d} - \binom{b'}{d'}\right\|_\mu \quad \text{for } b, b' \in \mathbb{R}^m, d, d' \in \mathbb{R}^k, \tag{1.2}$$

where $\|\cdot\|_\mu$, $\|\cdot\|_\nu$ denote two arbitrary norms and $H(\cdot,\cdot)_\nu$ denotes the Hausdorff metric induced by $\|\cdot\|_\nu$:

$$H(K,G)_\nu := \max\{\max_{x\in K}\min_{y\in G}\|x-y\|_\nu, \max_{y\in G}\min_{x\in K}\|y-x\|_\nu\} \quad \text{for } K, G \subset \mathbb{R}^n.$$

The scalar γ in (1.2) is also known as a Lipschitz constant (or a condition constant (cf. [21])) for the solutions of the linear system $Ax \le b, Cx = d$ with respect to right-hand side perturbations. Sixteen years later, in characterizing convex polyhedra, Walkup and Wets [32] proved that any convex polyhedral set-valued mapping is Lipschitz continuous. Even though this statement is an easy consequence of the Lipschitz continuity of F (cf. [14]), the proofs are completely different. Actually, the proof given by Walkup and Wets implies that ext F is also Lipschitz continuous, where ext $F\binom{b}{d}$ denotes the set of all basic feasible solutions (i.e., extreme points or vertices) of $F\binom{b}{d}$. That is, there exists a scalar $\beta > 0$ such that

$$H\left(\text{ext } F\binom{b}{d}, \text{ext } F\binom{b'}{d'}\right)_\nu \le \beta \cdot \left\|\binom{b}{d} - \binom{b'}{d'}\right\|_\mu \quad \text{for } b, b' \in \mathbb{R}^m, d, d' \in \mathbb{R}^k. \tag{1.3}$$

* Received by the editors March 9, 1992; accepted for publication (in revised form) October 20, 1992.
† Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia 23529, (wuli@math.odu.edu).

There are quite a few papers on estimation of γ in (1.2). Robinson's estimate [28] involves the minimum norm of points in certain polyhedral sets. Cook et al. [4] had an estimate of γ by the determinants of square submatrices of $\binom{A}{C}$ when $\|\cdot\|_\mu, \|\cdot\|_\nu$ are $\|\cdot\|_\infty$-norm and $\binom{A}{C}$ is an integer matrix. By using the norm of the inverse of $\binom{A}{C}$ on a polyhedral set, Mangasarian [21], and Mangasarian and Shiau [23] gave an estimate of γ in the case that $\|\cdot\|_\nu$ is the $\|\cdot\|_\infty$-norm. Bergthaller and Singer [2] used the norms of nonsingular submatrices of $\binom{A}{C}$ to estimate γ when $\|\cdot\|_\mu$ is the $\|\cdot\|_\infty$-norm. (In private correspondence, I. Singer showed the author that their approach actually yields estimates of γ for arbitrary norms $\|\cdot\|_\mu$ and $\|\cdot\|_\nu$.) Mangasarian and Shiau, and Bergthaller and Singer showed that their estimates are better than the one given by Cook, Gerards, Schrijver, and Tardos in the case that $\|\cdot\|_\mu$ and $\|\cdot\|_\nu$ are $\|\cdot\|_\infty$-norm. With the exception of Bergthaller and Singer's paper, all authors had similar estimates for S. It is worth mentioning that Mangasarian and Shiau's estimate of Lipschitz constant for S is independent of c. For applications of such Lipschitz constants in convergence analysis of descent methods for solving linearly constrained minimization problems, see [9], [12], and [19].

However, there is no paper on estimation of β in (1.3). Walkup and Wets proved the existence of β in (1.3) by using lifting projections of convex polyhedra [32].

The main purpose of this paper is to give the sharp Lipschitz constants for ext F and ext S. In §2, we first show that the local upper Lipschitz constant of a Hausdorff lower semicontinuous mapping is equal to the Lipschitz constant of the mapping. Then we prove that the Lipschitz constant of a finite-set-valued mapping can be inherited by its continuous submappings. The first result enables us to use a local estimate of β as a global one, and the second result is used to prove that any Lipschitz constant for ext F can be used as a Lipschitz constant for ext S when rank $\binom{A}{C} = n$. In general, ext $F\binom{b}{d}$ and ext $S\binom{b}{d}$ might be empty, so we consider the submappings F_0 (of F) and S_0 (of S) in §3. Lipschitz constants for ext F_0 and ext S_0 are given in various norms, which reduce to Lipschitz constants of ext F and ext S when rank $\binom{A}{C} = n$. The Lipschitz constants given in p-norms are extremely simple and most likely to be useful. The main objective of §4 is to show that Lipschitz constants for ext F_0 can be used as Lipschitz constants for F and S. Therefore, we obtain various Lipschitz constants for F and S. It becomes clear from the analysis given in §3 that the Lipschitz constants given for ext F when rank $\binom{A}{C} = n$ are sharp under very general assumptions. The detailed proof is given in §5. Also, we compare our Lipschitz constants for F and S with some known, better Lipschitz constants for F and S in §5. Comments and remarks are given in §6.

To conclude this section, we introduce some common notation used in the following sections. For any vector x (or matrix B) and an index set I, x_I (or B_I), denote the vector (or matrix) consisting of components (or rows) of x (or B) whose indices are in I. Let $B_{I,0}$ be the matrix obtained by replacing rows of B whose indices are not in I by 0. x^T (or B^T) is the transpose of x (or B). rank(B) denotes the rank of the matrix B. B^+ denotes the pseudoinverse of B [10]. $\mathcal{R}(B)$ is the range of B (i.e., $\mathcal{R}(B) := \{Bx : x \in \mathbb{R}^s\}$ if B is an $r \times s$ matrix). B is said to be row independent if row vectors of B are linearly independent. Let diag$(\alpha_1, \cdots, \alpha_n)$ be the $n \times n$ diagonal matrix with diagonal elements $\alpha_1, \cdots, \alpha_n$. For any set $K \subset \mathbb{R}^n$, conv(K) stands for the convex hull of K. For $x \in \mathbb{R}^n$, the p-norm of x is defined as $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $1 \le p < \infty$ and $\|x\|_\infty := \max_{1\le i\le n}|x_i|$. Let $\|\cdot\|_\nu$ and $\|\cdot\|_\mu$ be two arbitrary norms on $\mathbb{R}^n$ and $\mathbb{R}^{m+k}$, respectively. $\|\cdot\|_\mu$ is said to be a monotone norm if $\|x\|_\mu \le \|y\|_\mu$, whenever $|x_i| \le |y_i|$ for $1 \le i \le m+k$. For $x \in \mathbb{R}^n$ and $K \subset \mathbb{R}^n, d(x, K)_\nu := \min\{\|x - y\|_\nu : y \in K\}$. The upper Hausdorff metric $d(G, K)_\nu := \sup_{x\in G} d(x, K)_\nu$ for $G, K \subset \mathbb{R}^n$, and the Hausdorff metric $H(\cdot,\cdot)_\nu$ on subsets of $\mathbb{R}^n$ with respect to $\|\cdot\|_\nu$ is defined as

$$H(K,G)_\nu := \max\{d(K,G)_\nu, d(G,K)_\nu\} \quad \text{for } G,K \subset \mathbb{R}^n.$$

To avoid the case when $K = \emptyset$ or $G = \emptyset$, we assume that $d(K,\emptyset)_\nu = d(\emptyset,K)_\nu = -\infty$. $K + G := \{x + y : x \in K, y \in G\}$. For an index set I, $|I|$ denotes the number of indices in I. Define

$$\mathcal{M}(A,C) := \left\{ I \subset \{i\}_1^m : |I| = \operatorname{rank}\begin{pmatrix} A \\ C \end{pmatrix} - \operatorname{rank}(C),\ \operatorname{rank}\begin{pmatrix} A_I \\ C \end{pmatrix} = \operatorname{rank}\begin{pmatrix} A \\ C \end{pmatrix} \right\}. \tag{1.4}$$

For an $n \times (m+k)$ matrix B, the norm of B as a linear mapping from $(\mathbb{R}^{m+k}, \|\cdot\|_\mu)$ to $(\mathbb{R}^n, \|\cdot\|_\nu)$ is defined as

$$\|B\|_{\mu,\nu} := \max\{\|By\|_\nu : y \in \mathbb{R}^{m+k}, \|y\|_\mu = 1\}.$$

Then the Lipschitz constant for F, ext F, S, ext S is given by the following formula:

$$\beta_{\mu,\nu}(A,C) := \max_{I \in \mathcal{M}(A,C)} \left\| \begin{pmatrix} A_{I,0} \\ C \end{pmatrix}^+ \right\|_{\mu,\nu}. \tag{1.5}$$

In general, ext $F\binom{b}{d}$ and ext $S\binom{b}{d}$ might be empty sets. Therefore, we consider the following submappings of F and S:

$$\begin{aligned} F_0\begin{pmatrix} b \\ d \end{pmatrix} &:= \{x \in \mathbb{R}^n : Ax \le b, Cx = d, Qx = 0\}, \\ S_0\begin{pmatrix} b \\ d \end{pmatrix} &= \{x \in \mathbb{R}^n : c^T x = c_{\min}(b,d), Ax \le b, Cx = d, Qx = 0\}, \end{aligned} \tag{1.6}$$

where Q is any matrix such that $QA^T = 0, QC^T = 0$, and rank $\begin{pmatrix} A \\ C \\ Q \end{pmatrix} = n$. Note that $x \in \text{ext } F_0\binom{b}{d}$ (i.e., x is a vertex of $F_0\binom{b}{d}$) if and only if $x \in F_0\binom{b}{d}$ and there exists $J \in \mathcal{M}(A,C)$ such that $A_J x = b_J$. When rank $\binom{A}{C} = n, x \in \text{ext } F\binom{b}{d}$ is called a basic feasible solution and $x \in \text{ext } S\binom{b}{d}$ is called a basic optimal solution.

A set-valued mapping (or multifunction) T from $(\mathbb{R}^{m+k}, \|\cdot\|_\mu)$ to the subsets of $(\mathbb{R}^n, \|\cdot\|_\nu)$ (i.e., $T(x) \subset \mathbb{R}^n$ for $x \in \mathbb{R}^{k+m}$) is said to be λ-Lipschitz continuous, denoted by $T \in \text{Lip}(\lambda)_{\mu,\nu}$, if

$$H(T(x),T(y))_\nu \le \lambda \cdot \|x - y\|_\mu \quad \text{for } x,y \in \mathbb{R}^{k+m}.$$

T is said to be locally upper Lipschitz continuous with modulo λ, denoted by $UL(\lambda)_{\mu,\nu}$ [30], if, for any $x \in \mathbb{R}^{k+m}$, there exists a neighborhood U of x such that

$$d(T(y),T(x))_\nu \le \lambda \cdot \|x - y\|_\mu \quad \text{for } y \in U.$$

T is Hausdorff lower semicontinuous if $\lim_{z \to x} d(T(x),T(z))_\nu = 0$. T is Hausdorff upper semicontinuous if $\lim_{z \to x} d(T(z),T(x))_\nu = 0$. T is Hausdorff continuous if T is both Hausdorff upper semicontinuous and Hausdorff lower semicontinuous (i.e., $\lim_{z \to x} H(T(z), T(x))_\nu = 0$). T_0 is said to a submapping of T if $T_0(x) \subset T(x)$ for all x.

2. Lipschitz constant and local upper Lipschitz constant. In this section, we first show that the local upper Lipschitz constant of a Hausdorff lower semicontinuous mapping is equal to the Lipschitz constant of the mapping. This property of set-valued mappings

allows us to use local Lipschitz constants for ext F_0 and ext F as global Lipschitz constants (cf. the proof of Lemma 3.3). Then we prove that the Lipschitz constant of a finite-set-valued mapping can be inherited by its continuous submappings. So it becomes clear that the Lipschitz constants for ext F_0 and ext F can be used as Lipschitz constants for ext S_0 and ext S (cf. Lemmas 3.7 and 3.8).

THEOREM 2.1. *If T is Hausdorff lower semicontinuous and $T \in UL(\lambda)_{\mu,\nu}$, then $T \in Lip(\lambda)_{\mu,\nu}$.*

Proof. Let $x, y \in \mathbb{R}^{k+m}$. Let $y_\theta = \theta y + (1-\theta)x$. Define

$$\theta^* = \sup\{\theta : d(T(y_\theta), T(x))_\nu \le \lambda \cdot \|y_\theta - x\|_\mu \text{ and } 0 \le \theta \le 1\}.$$

First we claim that

$$d(T(y_{\theta^*}), T(x))_\nu \le \lambda \cdot \|y_{\theta^*} - x\|_\mu. \tag{2.1}$$

In fact, there exists a sequence of scalars $\theta_1, \theta_2, \theta_3, \ldots$, such that

$$d(T(y_{\theta_\tau}), T(x))_\nu \le \lambda \cdot \|y_{\theta_\tau} - x\|_\mu \quad \text{and} \quad \lim_\tau \theta_\tau = \theta^*.$$

For any θ_τ, we have

$$\begin{aligned} d(T(y_{\theta^*}), T(x))_\nu &\le d(T(y_{\theta^*}), T(y_{\theta_\tau}))_\nu + d(T(y_{\theta_\tau}), T(x))_\nu \\ &\le d(T(y_{\theta^*}), T(y_{\theta_\tau}))_\nu + \lambda \cdot \|y_{\theta_\tau} - x\|_\mu, \end{aligned}$$

where the first inequality follows from the triangle inequality for upper Hausdorff metric. Since $y_{\theta_\tau} \to y_{\theta^*}$ as $\tau \to \infty$, by the Hausdorff lower semicontinuity of $T, d(T(y_{\theta^*}), T(y_{\theta_\tau}))_\nu \to 0$ as $\tau \to \infty$. Therefore,

$$\begin{aligned} d(T(y_{\theta^*}), T(x))_\nu &\le \lim_\tau [d(T(y_{\theta^*}), T(y_{\theta_\tau}))_\nu + \lambda \cdot \|y_{\theta_\tau} - x\|_\mu] \\ &= \lim_\tau \lambda \cdot \|y_{\theta_\tau} - x\|_\mu = \lambda \cdot \|y_{\theta^*} - x\|_\mu. \end{aligned}$$

This proves (2.1).

Now we claim that $\theta^* = 1$. In fact, since $T \in UL(\lambda)_{\mu,\nu}$, if $\theta^* < 1$, then there exists $\theta^* < \theta < 1$ such that

$$d(T(y_\theta), T(y_{\theta^*}))_\nu \le \lambda \cdot \|y_\theta - y_{\theta^*}\|_\mu. \tag{2.2}$$

It follows from (2.1) and (2.2) that

$$\begin{aligned} d(T(y_\theta), T(x))_\nu &\le d(T(y_\theta), T(y_{\theta^*}))_\nu + d(T(y_{\theta^*}), T(x))_\nu \\ &\le \lambda \cdot \|y_\theta - y_{\theta^*}\|_\mu + \lambda \cdot \|y_{\theta^*} - x\|_\mu \\ &= \lambda \cdot \|y_\theta - x\|_\mu, \end{aligned}$$

which contradicts the definition of θ^*. Since $\theta^* = 1$, (2.1) implies that $T \in Lip(\lambda)_{\mu,\nu}$. □

Remark. It follows from the proof of Theorem 2.1 that T is λ-Lipschitz continuous on any convex subset where T is Hausdorff lower semicontinuous. An immediate consequence of this observation is that the solutions of linear complementarity problem $(M, q(t))$, with $q(t) := (1-t)q^1 + tq^2$, are Lipschitz continuous for $0 \le t \le 1$ if they are Hausdorff lower semicontinuous for $0 \le t \le 1$. This is an improvement of Theorem 3.2 in [23]. Furthermore, it is known that polyhedral set-valued mappings are in $UL(\lambda)$ [30]. Thus, the Lipschitz continuity of such mappings is equivalent to their Hausdorff lower semicontinuity. Theorem 2.1 might also be useful to study general parametric programs with polyhedral structures [14].

Similarly, the following "dual form" of Theorem 2.1 also holds. We say that T is locally lower Lipschitz continuous with modulo λ, if, for any x, there exists a neighborhood U of x such that

$$d(T(x), T(y))_\nu \le \lambda \cdot \|x - y\|_\mu \quad \text{for } y \in U.$$

THEOREM 2.2. *If T is Hausdorff upper semicontinuous and T is locally lower Lipschitz continuous with modulo λ, then $T \in Lip(\lambda)_{\mu,\nu}$.*

Note that ext $S\binom{b}{d} \subset$ ext $F\binom{b}{d}$, and ext $F\binom{b}{d}$ is a finite set. In this case, the Lipschitz constant of ext F can be inherited by ext S. This is actually true for general set-valued mappings.

THEOREM 2.3. *Suppose that $T \in \mathrm{Lip}(\lambda)_{\mu,\nu}$ and $T(x)$ is a finite set for any x. If T_0 is a Hausdorff continuous submapping of T, then $T_0 \in \mathrm{Lip}(\lambda)_{\mu,\nu}$.*

Proof. Since $T \in Lip(\lambda)_{\mu,\nu}$ and $T_0(x) \subset T(x)$, we have

$$d(T_0(y), T(x))_\nu \le d(T(y), T(x))_\nu \le \lambda \cdot \|x - y\|_\mu \quad \text{for all } x, y. \tag{2.3}$$

Since T_0 is Hausdorff upper semicontinuous and $T(x)$ is a finite set, there exists a neighborhood U of x such that

$$d(T_0(y), T(x))_\nu = d(T_0(y), T_0(x))_\nu \quad \text{for } y \in U. \tag{2.4}$$

It follows from (2.3) and (2.4) that $T_0 \in UL(\lambda)_{\mu,\nu}$. Since T_0 is also Hausdorff lower semicontinuous, by Theorem 2.1, $T_0 \in \mathrm{Lip}(\lambda)_{\mu,\nu}$. □

Remark. The above theorem says that the Lipschitz constant of a finite-set-valued mapping can be inherited by its continuous submappings.

3. Lipschitz constants for basic feasible solutions and basic optimal solutions. The main results in this section are Theorems 3.4, 3.6, 3.9, and 3.10, which give the Lipschitz constants for ext F and ext S in terms of pseudoinverses. The Lipschitz constant for ext F and ext S in p-norms is extremely simple and most likely to be useful. The key technical result is Lemma 3.3, which gives a Lipschitz constant for ext F_0. All Lipschitz constants for ext F, ext S, F, and S are based on Lemma 3.3. As an application of Theorem 2.3, we prove that any Lipschitz constant for ext F_0 (or ext F) can be used as a Lipschitz constant for ext S_0 (or ext S) (cf. Lemmas 3.7 and 3.8). Lemma 3.5 is stated for deriving the Lipschitz constant for F and S in p-norms.

Before proving the key technical lemma, we need two auxiliary results. The first is the Lipschitz continuity of ext F_0 implicitly proved by Walkup and Wets [32], which is stated here for easy reference, and the second is an identity about pseudoinverses, which enables us to handle nonmonotone norm $\|\cdot\|_\mu$.

LEMMA 3.1. *If T is a convex polyhedral set-valued mapping (i.e., $\{(x, y) : y \in T(x)\}$ is a convex polyhedral set), then* ext T *is Lipschitz continuous. Specifically,* ext F_0 *is Lipschitz continuous.*

LEMMA 3.2. *Suppose that matrices B and E satisfy $BE = E$ and $B^T = B$. Then $E^+B = E^+$.*

Proof. By the Moore–Penrose conditions [10], we have $EE^+E = E, E^+EE^+ = E^+, (EE^+)^T = EE^+$, and $(E^+E)^T = E^+E$. Therefore,

$$\begin{aligned}
E(E^+B)E &= EE^+(BE) = EE^+E = E,\\
(E^+B)E(E^+B) &= E^+(BE)E^+B = (E^+EE^+)B = E^+B,\\
(E(E^+B))^T &= (BEE^+B)^T = B^T(EE^+)^TB^T = B(EE^+)B = E(E^+B),\\
((E^+B)E)^T &= (E^+E)^T = E^+E = (E^+B)E.
\end{aligned}$$

Thus, E^+B is also a pseudoinverse of E. Since the pseudoinverse is unique,

$$E^+ = E^+B. \qquad \square$$

LEMMA 3.3. *For any* $b, b' \in \mathbb{R}^m$ *and* $d, d' \in \mathbb{R}^k$,

$$H\left(\text{ext } F_0\begin{pmatrix} b' \\ d' \end{pmatrix}, \text{ext } F_0\begin{pmatrix} b \\ d \end{pmatrix}\right)_\nu \leq \beta_{\mu,\nu}(A, C) \left\| \begin{pmatrix} b' - b \\ d' - d \end{pmatrix} \right\|_\mu .$$

Proof. Let w be an extreme point of $F_0\binom{b}{d}$. Let I be the index set such that

$$A_I w = b_I \quad \text{and} \quad A_i w < b_i \quad \text{for } i \notin I. \tag{3.1}$$

It follows from (3.1) that there exists $\delta_w > 0$ such that if $\|x - w\|_\nu \leq \delta_w$ and $\|\binom{b-b'}{d-d'}\|_\mu \leq \delta_w$, then

$$A_i x < b'_i \quad \text{for } i \notin I. \tag{3.2}$$

Now let $U_w = \{\binom{b'}{d'} : \|\binom{b-b'}{d-d'}\|_\mu \leq \delta_w\}, V_w = \{x : \|x - w\|_\nu < \delta_w\}, \binom{b'}{d'} \in U_w$, and $z \in \text{ext } F_0\binom{b'}{d'} \cap V_w$. Since z is a vertex, there exists an index set $J \in \mathcal{M}(A, C)$ such that $A_J z = b'_J$. By (3.2), we know that $J \subset I$. Let $\mathcal{I}_0 = \text{diag}(\lambda_1, \ldots, \lambda_{k+m})$, where $\lambda_i = 0$ for $1 \leq i \leq m$ and $i \notin J$, and $\lambda_i = 1$, otherwise. Then, $z - w$ satisfies the following system of linear equations:

$$\begin{pmatrix} A_{J,0} \\ C \end{pmatrix}(w - z) = \mathcal{I}_0 \begin{pmatrix} b - b' \\ d - d' \end{pmatrix}. \tag{3.3}$$

Since $\text{rank} \begin{pmatrix} A_{J,0} \\ C \\ Q \end{pmatrix} = n$ and $AQ^T = 0, CQ^T = 0$, we know that $\mathcal{R}(Q^T)$ is the orthogonal complement of $\mathcal{R}\left(\binom{A_{J,0}}{C}^T\right)$ [10]. It follows from $Q(w - z) = 0$ that $w - z \in \mathcal{R}\left(\binom{A_{J,0}}{C}^T\right)$. Since $\binom{A_{J,0}}{C}^T\binom{A_{J,0}}{C}$ is the orthogonal projection to $\mathcal{R}\left(\binom{A_{J,0}}{C}^T\right)$ [10], we obtain

$$\begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}(w - z) = (w - z). \tag{3.4}$$

By (3.3) and (3.4), we have

$$w - z = \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \mathcal{I}_0 \begin{pmatrix} b - b' \\ d - d' \end{pmatrix}.$$

By $\mathcal{I}_0\binom{A_{J,0}}{C} = \binom{A_{J,0}}{C}, \mathcal{I}_0^T = \mathcal{I}_0$, and Lemma 3.2, we obtain

$$\begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \mathcal{I}_0 = \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+. \tag{3.5}$$

Thus,

$$w - z = \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} b - b' \\ d - d' \end{pmatrix}, \tag{3.6}$$

which implies that

$$\|w - z\|_\nu \le \beta_{\mu,\nu}(A, C) \left\| \begin{pmatrix} b - b' \\ d - d' \end{pmatrix} \right\|_\mu . \tag{3.7}$$

Since ext F_0 is Hausdorff upper semicontinuous (cf. Lemma 3.1) and ext $F_0\binom{b}{d}$ is a finite set, by (3.7) there exists a neighborhood U of $\binom{b}{d}$ such that

$$d\left(\text{ext } F_0\begin{pmatrix} b' \\ d' \end{pmatrix}, \text{ext } F_0\begin{pmatrix} b \\ d \end{pmatrix}\right)_\nu \le \beta_{\mu,\nu}(A, C) \left\| \begin{pmatrix} b - b' \\ d - d' \end{pmatrix} \right\|_\mu \quad \text{for } \begin{pmatrix} b' \\ d' \end{pmatrix} \in U.$$

Thus, ext $F_0 \in UL(\beta_{\mu,\nu}(A, C))_{\mu,\nu}$. Since ext F_0 is Hausdorff lower semicontinuous (cf. Lemma 3.1), it follows from Theorem 2.1 that

$$H\left(\text{ext } F_0\begin{pmatrix} b' \\ d' \end{pmatrix}, \text{ext } F_0\begin{pmatrix} b \\ d \end{pmatrix}\right)_\nu \le \beta_{\mu,\nu}(A, C) \left\| \begin{pmatrix} b - b' \\ d - d' \end{pmatrix} \right\|_\mu . \qquad \square$$

Remark. If rank $\binom{A}{C} = n$, then ext $F_0\binom{b}{d} = \text{ext } F\binom{b}{d}$. Thus, an immediate consequence of Lemma 3.3 is the following estimate of the Lipschitz constant for ext F.

THEOREM 3.4. *If* rank $\binom{A}{C} = n$, *then*

$$H\left(\text{ext } F\begin{pmatrix} b \\ d \end{pmatrix}, \text{ext } F\begin{pmatrix} b' \\ d' \end{pmatrix}\right)_\nu \le \beta_{\mu,\nu}(A, C) \cdot \left\| \begin{pmatrix} b - b' \\ d - d' \end{pmatrix} \right\|_\mu , \tag{3.8}$$

i.e., ext $F \in Lip(\beta_{\mu,\nu}(A, C))_{\mu,\nu}$.

Remark. Note that, by ignoring the zero components, we can replace (3.6) by

$$w - z = \begin{pmatrix} A_J \\ C \end{pmatrix}^+ \begin{pmatrix} b_J - b'_J \\ d - d' \end{pmatrix}.$$

Therefore, for $1 \le p, q \le \infty$,

$$\|w - z\|_p \le \left\| \begin{pmatrix} A_J \\ C \end{pmatrix}^+ \right\|_{q,p} \left\| \begin{pmatrix} b_J - b'_J \\ d - d' \end{pmatrix} \right\|_q \le \beta^*_{q,p}(A, C) \cdot \left\| \begin{pmatrix} b - b' \\ d - d' \end{pmatrix} \right\|_q ,$$

where

$$\beta^*_{q,p}(A, C) := \max_{J \in \mathcal{M}(A,C)} \left\| \begin{pmatrix} A_J \\ C \end{pmatrix}^+ \right\|_{q,p} . \tag{3.9}$$

If C is row independent, then $\binom{A_J}{C}$ is also row independent, and

$$\begin{pmatrix} A_J \\ C \end{pmatrix}^+ = \begin{pmatrix} A_J \\ C \end{pmatrix}^T \left(\begin{pmatrix} A_J \\ C \end{pmatrix} \begin{pmatrix} A_J \\ C \end{pmatrix}^T \right)^{-1} .$$

Thus, the proof of Lemma 3.3 yields the following estimates of Lipschitz constants for ext F_0 in p-norm and q-norm.

LEMMA 3.5. *Let* $1 \le p, q \le \infty$. *If C is row independent, then*

$$H\left(\text{ext } F_0\begin{pmatrix} b' \\ d' \end{pmatrix}, \text{ext } F_0\begin{pmatrix} b \\ d \end{pmatrix}\right)_p \le \max_{J \in \mathcal{M}(A,C)} \left\| \begin{pmatrix} A_J \\ C \end{pmatrix}^T \left(\begin{pmatrix} A_J \\ C \end{pmatrix} \begin{pmatrix} A_J \\ C \end{pmatrix}^T \right)^{-1} \right\|_{q,p} \left\| \begin{pmatrix} b' - b \\ d' - d \end{pmatrix} \right\|_q .$$

Remark. We could verify that $\beta^*_{q,p}(A,C)$ defined in (3.9) is actually equal to $\beta_{q,p}(A,C)$ defined in (1.4) for $1 \le p,q \le \infty$ (cf. [18, Lem. 3.6]). When rank $\binom{A}{C} = n$ and C is row independent, then the pseudoinverse in (3.9) is the inverse.

THEOREM 3.6. *If* $1 \le p,q \le \infty$, *rank* $\binom{A}{C} = n$, *and* C *is row independent, then*

$$H\left(\text{ext } F\binom{b'}{d'}, \text{ext } F\binom{b}{d}\right)_p \le \max_{J\in\mathcal{M}(A,C)} \left\| \binom{A_J}{C}^{-1} \right\|_{q,p} \cdot \left\| \binom{b'-b}{d'-d} \right\|_q .$$

Note that, if $c = 0$, then $S = F$ and $S_0 = F_0$. Therefore, it seems that Lipschitz constants for ext S_0 and ext S with $c = 0$ yield Lipschitz constants for ext F_0 and ext F. However, it becomes apparent that Lipschitz constants for ext F_0 and ext F are also valid for ext S_0 and ext S.

LEMMA 3.7. *If* ext $F_0 \in Lip(\beta)_{\mu,\nu}$, *then* ext $S_0 \in Lip(\beta)_{\mu,\nu}$.

Proof. It is well known that ext $S_0\binom{b}{d} \subset$ ext $F_0\binom{b}{d}$ for all b,d (i.e., basic optimal solutions are basic feasible solutions). Since $c_{\min}(b,d)$ is a continuous function of b,d, by Lemma 3.1, ext S_0 is a continuous mapping. Note that ext $F_0\binom{b}{d}$ is a finite set for any b,d. Therefore, the above lemma is a consequence of Theorem 2.3. □

LEMMA 3.8. *If* rank $\binom{A}{C} = n$ *and* ext $F \in Lip(\beta)_{\mu,\nu}$, *then* ext $S \in Lip(\beta)_{\mu,\nu}$.

Remark. The above lemma shows that any Lipschitz constant for ext F works for ext S. Thus, the following two theorems are consequences of Lemma 3.8, and Theorems 3.4 and 3.6.

THEOREM 3.9. *If* rank $\binom{A}{C} = n$, *then*

$$H\left(\text{ext } S\binom{b}{d}, \text{ext } S\binom{b'}{d'}\right)_\nu \le \beta_{\mu,\nu}(A,C) \cdot \left\| \binom{b-b'}{d-d'} \right\|_\mu .$$

THEOREM 3.10. *If* $1 \le p,q \le \infty$, rank $\binom{A}{C} = n$, *and* C *is row independent, then*

$$H\left(\text{ext } S\binom{b'}{d'}, \text{ext } S\binom{b}{d}\right)_p \le \max_{J\in\mathcal{M}(A,C)} \left\| \binom{A_J}{C}^{-1} \right\|_{q,p} \cdot \left\| \binom{b'-b}{d'-d} \right\|_q .$$

4. Lipschitz constants for feasible and optimal solutions. It is well known that any convex polyhedral set can be represented as a convex hull of a finite set and a convex recession cone [15], [31]. The first two lemmas in this section give representations of $F\binom{b}{d}$ and $S\binom{b}{d}$ as a convex hull of a finite set and a convex recession cone in algebraic form. Then we show that the Hausdorff distance of two sets $\text{conv}(X) + K$ and $\text{conv}(Y) + K$ is bounded by the Hausdorff distance of X and Y. This result, together with Lemmas 4.1 and 4.2, implies that the Lipschitz constant for ext F_0 and ext S_0 are also valid for F and S. Therefore, we obtain various Lipschitz constants for F and S from the Lipschitz constants for ext F_0 given in §3.

LEMMA 4.1. *Let* $K := \{x : Ax \le 0, Cx = 0\}$. *Then*

$$F\binom{b}{d} = \text{conv}\left(\text{ext } F_0\binom{b}{d}\right) + K. \tag{4.1}$$

Proof. In fact, it is easy to see that $F\binom{b}{d} \supset \text{conv}\left(\text{ext } F_0\binom{b}{d}\right) + K$. Since $F\binom{b}{d} \subset F_0\binom{b}{d} + K$, it suffices to show that

$$F_0\binom{b}{d} \subset \text{conv}\left(\text{ext } F_0\binom{b}{d}\right) + K.$$

Since $F_0\binom{b}{d}$ contains no line, it can be generated by its extreme points and extreme rays [15]. Therefore, it suffices to show that any extreme ray $\ell := \{z + \lambda p : \lambda \geq 0\}$ is a subset of ext $F_0\binom{b}{d} + K$. In fact, $z \in$ ext $F_0\binom{b}{d}$. $A(z + \lambda p) \leq b$ and $C(z + \lambda p) = d$, for $\lambda \geq 0$, imply $Ap \leq 0$ and $Cp = 0$ (i.e., $p \in K$). Thus, $\ell \subset$ ext $F_0\binom{b}{d} + K$. □

Replacing C by $\binom{C}{c^T}$ in the above lemma, we obtain the following relation between ext S_0 and S.

LEMMA 4.2. *Let* $K := \{x : Ax \leq 0, Cx = 0, c^T x = 0\}$. *Then*

$$S\binom{b}{d} = \text{conv}\left(\text{ext } S_0\binom{b}{d}\right) + K. \tag{4.2}$$

LEMMA 4.3. *Let X,Y,K be closed subsets of* $\mathbb{R}^n$. *Then* $H(\text{conv}(X) + K, \text{conv}(Y) + K)_\nu \leq H(X, Y)_\nu$.

Proof. Let $u \in \text{conv}(X) + K$. Then there exist $\{x^i\}_1^t \subset X, z \in K$, and $\theta_i > 0$ for $1 \leq i \leq t$, such that $\sum_{i=1}^t \theta_i = 1$ and $u = z + \sum_{i=1}^t \theta_i x^i$. Let $\{y^i\}_1^t \subset Y$ be such that $d(x^i, Y)_\nu = \|x^i - y^i\|_\nu$ for $1 \leq i \leq t$. Then $v := z + \sum_{i=1}^t \theta_i x^i \in \text{conv}(Y) + K$ and

$$d(u, \text{conv}(Y) + K)_\nu \leq \|u - v\|_\nu \leq \sum_{i=1}^t \theta_i \cdot \|x^i - y^i\|_\nu \leq d(X, Y)_\nu.$$

Therefore, d(conv$(X) + K$, conv$(Y) + K)_\nu \leq d(X, Y)_\nu$. Similarly, d(conv$(Y) + K$, conv$(X) + K)_\nu \leq d(Y, X)_\nu$. □

Remark. The above three lemmas are algebraic versions of well-known results about convex sets. We include the proofs here for the convenience of readers. Particularly, Lemma 4.3 can be found in the proof of the main result in [32].

Now, by Lemmas 4.1–4.3, and 3.7, it is easy to see that the Lipschitz constant for ext F_0 can be used as a Lipschitz constant for F and S.

THEOREM 4.4. *If* ext $F_0 \in Lip(\beta)_{\mu,\nu}$, *then* $F, S \in Lip(\beta)_{\mu,\nu}$.

Therefore, we have the following error estimates for F and S, which are consequences of Theorem 4.4 and Lemmas 3.3, 3.5, and 3.6.

COROLLARY 4.5. *For any* $b, b' \in \mathbb{R}^m$ *and* $d, d' \in \mathbb{R}^k$,

$$H\left(F\binom{b}{d}, F\binom{b'}{d'}\right)_\nu \leq \beta_{\mu,\nu}(A, C) \cdot \left\| \binom{b - b'}{d - d'} \right\|_\mu.$$

COROLLARY 4.6. *If* $1 \leq p, q \leq \infty$ *and C is row independent, then*

$$H\left(F\binom{b'}{d'}, F\binom{b}{d}\right)_p \leq \max_{J \in \mathcal{M}(A,C)} \left\| \binom{A_J}{C}^T \left(\binom{A_J}{C} \binom{A_J}{C}^T \right)^{-1} \right\|_{q,p} \cdot \left\| \binom{b' - b}{d' - d} \right\|_q.$$

COROLLARY 4.7. *If* $1 \leq p, q \leq \infty$, rank $\binom{A}{C} = n$, *and C is row independent, then*

$$H\left(F\binom{b'}{d'}, F\binom{b}{d}\right)_p \leq \max_{J \in \mathcal{M}(A,C)} \left\| \binom{A_J}{C}^{-1} \right\|_{q,p} \cdot \left\| \binom{b' - b}{d' - d} \right\|_q.$$

COROLLARY 4.8. *For any* $b, b' \in \mathbb{R}^m$ *and* $d, d' \in \mathbb{R}^k$,

$$H\left(S\binom{b}{d}, S\binom{b'}{d'}\right)_\nu \leq \beta_{\mu,\nu}(A, C) \cdot \left\| \binom{b - b'}{d - d'} \right\|_\mu.$$

COROLLARY 4.9. *If* $1 \le p, q \le \infty$ *and* C *is row independent, then*

$$H\left(S\begin{pmatrix}b'\\d'\end{pmatrix}, S\begin{pmatrix}b\\d\end{pmatrix}\right)_p \le \max_{J\in\mathcal{M}(A,C)} \left\| \begin{pmatrix}A_J\\C\end{pmatrix}^T \left(\begin{pmatrix}A_J\\C\end{pmatrix}\begin{pmatrix}A_J\\C\end{pmatrix}^T\right)^{-1} \right\|_{q,p} \cdot \left\| \begin{pmatrix}b'-b\\d'-d\end{pmatrix} \right\|_q.$$

COROLLARY 4.10. *If* $1 \le p, q \le \infty$, rank $\begin{pmatrix}A\\C\end{pmatrix} = n$, *and* C *is row independent, then*

$$H\left(S\begin{pmatrix}b'\\d'\end{pmatrix}, S\begin{pmatrix}b\\d\end{pmatrix}\right)_p \le \max_{J\in\mathcal{M}(A,C)} \left\| \begin{pmatrix}A_J\\C\end{pmatrix}^{-1} \right\|_{q,p} \cdot \left\| \begin{pmatrix}b'-b\\d'-d\end{pmatrix} \right\|_q.$$

5. Sharpness of Lipschitz constants and comparison with known results. In this section, we first show that the Lipschitz constant for ext F and ext S is sharp under very general assumptions. Then we point out that the Lipschitz constants for F and S given in §4 can be improved.

THEOREM 5.1. *If* rank $\begin{pmatrix}A\\C\end{pmatrix} = n$, C *is row independent, and* $\|\cdot\|_\mu$ *is a monotone norm, then there exist* $\begin{pmatrix}b\\d\end{pmatrix}$ *and* $\begin{pmatrix}b'\\d'\end{pmatrix}$ *such that*

$$H\left(\text{ext } F\begin{pmatrix}b\\d\end{pmatrix}, \text{ext } F\begin{pmatrix}b'\\d'\end{pmatrix}\right)_\nu = \beta_{\mu,\nu}(A,C)\cdot \left\| \begin{pmatrix}b-b'\\d-d'\end{pmatrix} \right\|_\mu > 0, \tag{5.1}$$

i.e., $\beta_{\mu,\nu}(A,C)$ *is the sharp Lipschitz constant for* ext F.

Proof. Let $A_{J,0}$ be such that rank $\begin{pmatrix}A_{J,0}\\C\end{pmatrix} = n$ and

$$\beta_{\mu,\nu}(A,C) = \left\| \begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+ \right\|_{\mu,\nu}.$$

Let $\bar{b} \in \mathbb{R}^m$ and $\bar{d} \in \mathbb{R}^k$ be such that

$$\left\| \begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+ \begin{pmatrix}\bar{b}\\\bar{d}\end{pmatrix} \right\|_\nu = \left\| \begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+ \right\|_{\mu,\nu} \cdot \left\| \begin{pmatrix}\bar{b}\\\bar{d}\end{pmatrix} \right\|_\mu = \beta_{\mu,\nu}(A,C)\cdot \left\| \begin{pmatrix}\bar{b}\\\bar{d}\end{pmatrix} \right\|_\mu > 0. \tag{5.2}$$

Since $\begin{pmatrix}A_{J,0}\\C\end{pmatrix}$ has full column rank,

$$\begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+ = \left(\begin{pmatrix}A_{J,0}\\C\end{pmatrix}^T \begin{pmatrix}A_{J,0}\\C\end{pmatrix}\right)^{-1} \begin{pmatrix}A_{J,0}\\C\end{pmatrix}^T.$$

Let $\mathcal{I}_0 := \text{diag}(\lambda_1, \ldots, \lambda_{m+k})$, where $\lambda_i = 0$ for $1 \le i \le m, i \notin J$, and $\lambda_i = 1$, otherwise. Then

$$\mathcal{R}\left(\left(\begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+\right)^T\right) = \mathcal{R}\left(\begin{pmatrix}A_{J,0}\\C\end{pmatrix}\left(\begin{pmatrix}A_{J,0}\\C\end{pmatrix}^T \begin{pmatrix}A_{J,0}\\C\end{pmatrix}\right)^{-1}\right) = \mathcal{R}\begin{pmatrix}A_{J,0}\\C\end{pmatrix} = \mathcal{R}(\mathcal{I}_0),$$

where the last equality follows from the fact that the rank of $\begin{pmatrix}A_{J,0}\\C\end{pmatrix}$ is equal to the number of nonzero rows of $\begin{pmatrix}A_{J,0}\\C\end{pmatrix}$. Since $\begin{pmatrix}A_{J,0}\\C\end{pmatrix}\begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+$ is the orthogonal projection onto

$$\mathcal{R}\left(\left(\begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+\right)^T\right) = \mathcal{R}(\mathcal{I}_0)$$

[10], we have

$$\begin{pmatrix}A_{J,0}\\C\end{pmatrix}\begin{pmatrix}A_{J,0}\\C\end{pmatrix}^+ \mathcal{I}_0 = \mathcal{I}_0.$$

However, $\binom{A_{J,0}}{C}^+ \mathcal{I}_0 = \binom{A_{J,0}}{C}^+$ (cf. (3.5)). Therefore,

$$\begin{pmatrix} A_{J,0} \\ C \end{pmatrix} \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ = \mathcal{I}_0,$$

which implies that

$$\left\| \begin{pmatrix} A_{J,0} \\ C \end{pmatrix} \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix} \right\|_\mu = \left\| \mathcal{I}_0 \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix} \right\|_\mu \le \left\| \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix} \right\|_\mu, \tag{5.3}$$

where the last inequality follows from the monotonicity of $\|\cdot\|_\mu$.

Let $b_i^* := 0$ for $i \in J$ and $b_i^* := |(A\binom{A_{J,0}}{C}^+ \binom{\bar{b}}{\bar{d}})_i| + 1$ for $i \notin J$. Let

$$b_\theta := \theta \cdot A_{J,0} \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix} + b^*, \quad d_\theta := \theta \cdot C \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix},$$
$$x_\theta := \theta \cdot \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix}.$$

Consider the following system:

$$Ax \le b_\theta, \qquad Cx = d_\theta. \tag{5.4}$$

Note that x_θ is a vertex of the solution set of the above system, and $(Ax_\theta)_i < (b_\theta)_i$ for $i \notin J$ and $0 \le \theta \le 1$. If $\theta > 0$ is small enough, then x_0 is the vertex of the solution set of (5.4) for $\theta = 0$ that is closest to x_θ. Therefore, for $\theta > 0$ small enough,

$$\begin{aligned} H\left(\text{ext } F\begin{pmatrix} b_\theta \\ d_\theta \end{pmatrix}, \text{ext } F\begin{pmatrix} b_0 \\ d_0 \end{pmatrix}\right)_\nu &\ge d\left(x_\theta, \text{ext } F\begin{pmatrix} b_0 \\ d_0 \end{pmatrix}\right)_\nu = \|x_\theta - x_0\|_\nu \\ &= \theta \cdot \left\| \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix} \right\|_\nu = \theta \cdot \beta_{\mu,\nu}(A,C) \cdot \left\| \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix} \right\|_\mu \\ &\ge \beta_{\mu,\nu}(A,C) \cdot \theta \cdot \left\| \begin{pmatrix} A_{J,0} \\ C \end{pmatrix} \begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} \bar{b} \\ \bar{d} \end{pmatrix} \right\|_\mu \\ &= \beta_{\mu,\nu}(A,C) \cdot \left\| \begin{pmatrix} b_\theta - b_0 \\ d_\theta - d_0 \end{pmatrix} \right\|_\mu > 0, \end{aligned}$$

where the second inequality follows from (5.3) and the first equality follows from (5.2). Therefore, $\beta_{\mu,\nu}(A,C)$ is the sharp Lipschitz constant for ext F. □

Similarly, the Lipschitz constant for ext F given in Theorem 3.6 is also sharp. We leave the details for interested readers (cf. the remarks after Lemma 3.5).

THEOREM 5.2. *If* rank $\binom{A}{C} = n$, *C is row independent, and* $1 \le p, q \le \infty$, *then there exist* $\binom{b}{d}$ *and* $\binom{b'}{d'}$ *such that*

$$H\left(\text{ext } F\begin{pmatrix} b \\ d \end{pmatrix}, \text{ext } F\begin{pmatrix} b' \\ d' \end{pmatrix}\right)_p = \max_{J \in \mathcal{M}(A,C)} \left\| \begin{pmatrix} A_J \\ C \end{pmatrix}^{-1} \right\|_{q,p} \cdot \left\| \begin{pmatrix} b - b' \\ d - d' \end{pmatrix} \right\|_q > 0.$$

However, the Lipschitz constants for F and S given in §4 can be improved. In fact, we have the following sharp Lipschitz constants for F and S [18].

LEMMA 5.3. *For any* $b, b' \in \mathbb{R}^m, d, d' \in \mathbb{R}^k$,

$$H\left(F\begin{pmatrix} b \\ d \end{pmatrix}, F\begin{pmatrix} b' \\ d' \end{pmatrix}\right)_\nu \le \alpha_{\mu,\nu}(A,C)\cdot \left\| \begin{pmatrix} b-b' \\ d-d' \end{pmatrix} \right\|_\mu,$$

where $K_J := \{x \in \mathbb{R}^n : A_{J,0}x \le 0, Cx = 0\}$ *and*

$$\alpha_{\mu,\nu}(A,C) := \max_{J\in\mathcal{M}(A,C)} \sup\left\{ d\left(\begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} u \\ v \end{pmatrix}, K_J\right)_\nu : \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_\mu = 1 \right\}. \tag{5.5}$$

LEMMA 5.4. *For any* $b, b' \in \mathbb{R}^m, d, d' \in \mathbb{R}^k$,

$$H\left(S\begin{pmatrix} b \\ d \end{pmatrix}, S\begin{pmatrix} b' \\ d' \end{pmatrix}\right)_\nu \le \gamma_{\mu,\nu}(A,C)\cdot \left\| \begin{pmatrix} b-b' \\ d-d' \end{pmatrix} \right\|_\mu,$$

where $G := \{x \in \mathbb{R}^n : Ax = 0, Cx = 0\}$ *and*

$$\gamma_{\mu,\nu}(A,C) := \max_{J\in\mathcal{M}(A,C)} \sup\left\{ d\left(\begin{pmatrix} A_{J,0} \\ C \end{pmatrix}^+ \begin{pmatrix} u \\ v \end{pmatrix}, G\right)_\nu : \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_\mu = 1 \right\}. \tag{5.6}$$

Since $\{0\} \subset G \subset K_J$ and $G = \{0\}$ when $\dim \begin{pmatrix} A \\ C \end{pmatrix} = n$, it is easy to see that the following relation holds among $\gamma_{\mu,\nu}(A,C), \beta_{\mu,\nu}(A,C)$, and $\alpha_{\mu,\nu}(A,C)$.

LEMMA 5.5. $\alpha_{\mu,\nu}(A,C) \le \gamma_{\mu,\nu}(A,C) \le \beta_{\mu,\nu}(A,C)$. *Moreover,* $\gamma_{\mu,\nu}(A,C) = \beta_{\mu,\nu}(A,C)$, *if* $\dim \begin{pmatrix} A \\ C \end{pmatrix} = n$.

Let $\|x\|_{\nu^*} := \sup\{x^T y : y \in \mathbb{R}^n, \|y\|_\nu = 1\}$ denote the dual norm of $\|\cdot\|_\nu$. Then we have the following dual representations of $\alpha_{\mu,\nu}(A,C)$ and $\gamma_{\mu,\nu}(A,C)$ [18].

LEMMA 5.6. *If C is row independent, then*

$$\alpha_{\mu,\nu}(A,C) = \sup\left\{ \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_{\mu^*} : \begin{array}{l} \|A^T u + C^T v\|_{\nu^*} = 1, u \ge 0, \textit{ the rows of } A \\ \textit{corresponding to nonzero components of } u \\ \textit{and the rows of } C \textit{ are linearly independent} \end{array} \right\}, \tag{5.7}$$

$$\gamma_{\mu,\nu}(A,C) = \sup\left\{ \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_{\mu^*} : \begin{array}{l} \|A^T u + C^T v\|_{\nu^*} = 1, \textit{ the rows of } A \\ \textit{corresponding to nonzero components of } u \\ \textit{and the rows of } C \textit{ are linearly independent} \end{array} \right\}. \tag{5.8}$$

Remark. Similar Lipschitz constants, such as those given on the right-hand sides of (5.7) and (5.8), were first given by Mangasarian and Shiau [23]. When C is row independent and $\|\cdot\|_\nu \equiv \|\cdot\|_\infty$, they proved that $F \in Lip(\alpha^*_{\mu,\nu}(A,C))_{\mu,\nu}$, where

$$\alpha^*_{\mu,\nu}(A,C) := \sup\left\{ \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_{\mu^*} : \begin{array}{l} \|A^T u + C^T v\|_{\nu^*} = 1, u \ge 0, \text{ the rows of} \\ \begin{pmatrix} A \\ C \end{pmatrix} \text{ corresponding to nonzero components} \\ \text{of } \begin{pmatrix} u \\ v \end{pmatrix} \text{ are linearly independent} \end{array} \right\}.$$

When C is row independent, $\|\cdot\|_\nu \equiv \|\cdot\|_\infty$, and $\|\cdot\|_\mu$ is a monotone norm, they proved that $S \in Lip(\gamma^*_{\mu,\nu}(A,C))_{\mu,\nu}$, where

$$\gamma^*_{\mu,\nu}(A,C) := \sup\left\{ \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_{\mu^*} : \begin{array}{l} \|A^T u + C^T v\|_{\nu^*} = 1, \text{ the rows of} \\ \begin{pmatrix} A \\ C \end{pmatrix} \text{ corresponding to nonzero components} \\ \text{of } \begin{pmatrix} u \\ v \end{pmatrix} \text{ are linearly independent} \end{array} \right\}.$$

It is easy to see that $\gamma_{\mu,\nu}(A,C) \le \gamma^*_{\mu,\nu}(A,C)$ and $\alpha_{\mu,\nu}(A,C) \le \alpha^*_{\mu,\nu}(A,C)$. Also, it was shown in [18] that

$$\lim_{\epsilon\to 0+} \frac{\gamma^*_{\mu,\nu}(A_\epsilon,C)}{\gamma_{\mu,\nu}(A_\epsilon,C)} = +\infty \quad \text{and} \quad \lim_{\epsilon\to 0+} \frac{\alpha^*_{\mu,\nu}(A_\epsilon,C)}{\alpha_{\mu,\nu}(A_\epsilon,C)} = +\infty,$$

where $A_\epsilon = \begin{pmatrix} 1 & 0 \\ -1 & \epsilon \end{pmatrix}$ and $C = (0\ \ 1)$. Therefore, Lemmas 5.3 and 5.4 are improvements of Mangasarian and Shiau's results.

6. Conclusion and remarks. In this paper, the sharp Lipschitz constant for ext F is given in terms of norms of pseudoinverses of submatrices of $\binom{A}{C}$. In particular, the sharp Lipschitz constant for ext F in p-norms is given as the maximum norm of $n\times n$ nonsingular matrices of the form $\binom{A_J}{C}$. If rank $\binom{A}{C} = n$, then any Lipschitz constant for ext F can be used as Lipschitz constants for ext S, F, and S. In general, the local upper Lipschitz constant of a Hausdorff lower semicontinuous mapping is equal to the Lipschitz constant of the mapping, and the Lipschitz constant of a finite-set-valued mapping can be inherited by its continuous submappings.

There is an interesting application of Lipschitz constants for F. Consider $Q(t) := \{x \in \mathbb{R}^n : c^T x = t, Ax \le b, Cx = d\}$. Then $Q(c_{\min}(b,d)) = S\binom{b}{d}$ and $x \in Q(t)$, if $t = c^T x$ and $x \in F\binom{b}{d}$. Therefore, by applying Corollary 4.5 to Q, we establish that there exists a constant $\beta > 0$, depending on A, C, c, such that

$$d\left(x, S\begin{pmatrix} b \\ d \end{pmatrix}\right)_2 \le \beta \cdot (c^T x - c_{\min}(b,d)) \quad \text{for } x \in F\begin{pmatrix} b \\ d \end{pmatrix}, \tag{6.1}$$

which is Mangasarian and Meyer's result on the weak sharp minimum of linear programs [22]. Also it is easy to verify that

$$c^T x - c_{\min}(b,d) \le \|c\|_2 \cdot d\left(x, S\begin{pmatrix} b \\ d \end{pmatrix}\right)_2 \quad \text{for } x \in F\begin{pmatrix} b \\ d \end{pmatrix}.$$

Therefore,

$$d\left(x, S\begin{pmatrix} b \\ d \end{pmatrix}\right)_2 \le \beta \cdot (c^T x - c_{\min}(b,c)) \le \beta \cdot \|c\|_2 \cdot d\left(x, S\begin{pmatrix} b \\ d \end{pmatrix}\right)_2 \quad \text{for } x \in F\begin{pmatrix} b \\ d \end{pmatrix}.$$

If $\beta\|c\|_2$ is close to 1, then $c^T x - c_{\min}(b,d)$ is a fair measurement of the distance of a feasible point x to the solution set. Jittorntrum and Osborne proved in [13] and [25] that (6.1) holds if $S\binom{b}{d}$ is a singleton. Equation (6.1) is due to Mangasarian and Meyer [22], and is referred to as strong uniqueness in [13], [25], [26], and [17], as well as (weak) sharp minima in [27], [6], and [7]. Equation (6.1) is closely related to the strong uniqueness concept in approximation theory. For references on strong uniqueness in approximation theory, see [3], [5], [24], and [16].

Another related problem concerns Lipschitz constants for solutions of a system of nonlinear inequalities and equalities. See [29], [20], and [1] for relevant results and references. It would be interesting to know whether or not one could obtain Lipschitz constants for extreme points of solution sets of nonlinear systems.

REFERENCES

[1] A. Auslender and J.-P. Crouzeix, *Global regularity theorems*, Math. Oper. Res., 13 (1988), pp. 243–253.

[2] C. BERGTHALLER AND I. SINGER, *The distance to a polyhedron*, Linear Algebra Appl., 169 (1992), pp. 111–129.
[3] E. W. CHENEY, *Introduction to Approximation Theory*, 2nd ed., Chelsea, New York, 1982.
[4] W. COOK, A. M. H. GERARDS, A. SCHRIJVER, AND È. TARDOS, *Sensitivity theorems in integer linear programming*, Math. Programming, 34 (1986), pp. 251–264.
[5] L. CROMME, *Strong uniqueness: A far-reaching criterion for convergence analysis of iterative processes*, Numer. Math., 29 (1978), pp. 179–193.
[6] M. C. FERRIS, *Finite termination of the proximal point algorithm*, Math. Programming, 50 (1991), pp. 359–366.
[7] M. C. FERRIS AND O. L. MANGASARIAN, *Minimum principle sufficiency*, Math. Programming, 57 (1992), pp. 1–14.
[8] T. GAL, *Postoptimal Analyses, Parametric Programming and Related Topics*, McGraw-Hill, New York, 1979.
[9] J. L. GOFFIN, *The relaxation method for solving systems of linear inequalities*, Math. Oper. Res., 5 (1980), pp. 388–414.
[10] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The John Hopkins University Press, Baltimore, MD, 1983.
[11] A. J. HOFFMAN, *Approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.
[12] ALFREDO N. IUSEM AND ALVARO R. DE PIERRO, *On the convergence properties of Hildreth's quadratic programming algorithm*, Math. Programming, 47 (1990), pp. 37–51.
[13] K. JITTORNTRUM AND M. R. OSBORNE, *Strong uniqueness and second order convergence in nonlinear discrete approximation*, Numer. Math., 34 (1980), pp. 439–455.
[14] D. KLATTE, *Lipschitz continuity of infima and optimal solutions in parametric optimization: The polyhedral case*, in Parametric Optimization and Related Topics, Math. Res. 35, Akademie-Verlag, Berlin, 1987, pp. 229–248.
[15] V. KLEE, *Some characterizations of convex polyhedra*, Acta Math., 102 (1959), pp. 79–107.
[16] W. LI, *Strong uniqueness and Lipschitz continuity of metric projections: A generalization of the classical Haar theory*, J. Approx. Theory, 56 (1989), pp. 164–184.
[17] ———, *Best approximations in polyhedral spaces and linear programs*, in Approximation Theory: Proceedings of the Sixth Southeastern International Conference, G. Anastassiou, ed., Marcel Dekker, New York, pp. 393–400.
[18] ———, *The sharp Lipschitz constants for feasible 2nd optimal solutions of a perturbed linear program*, Linear Algebra Appl., 187 (1993), pp. 15–40.
[19] Z.-Q. LUO AND P. TSENG, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM J. Control Optim., 30 (1992), pp. 408–425.
[20] O. L. MANGASARIAN, *A condition number for differentiable convex inequalities*, Math. Oper. Res., 10 (1985), pp. 175–179.
[21] ———, *A condition number of linear inequalities and equalities*, in Methods of Operations Research 43, Proc. 6. Symposium über Operations Research, Universität Augsburg, September 7–9, 1981, G. Bamberg and O. Opitz, eds., Verlagsgruppe Athennäum/Hain/Scriptor/Hanstein, Konigstein, 1981, pp. 3–15.
[22] O. L. MANGASARIAN AND R. R. MEYER, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim., 17 (1979), pp. 745–752.
[23] O. L. MANGASARIAN AND T.-H. SHIAU, *Lipschitz continuity of solutions of linear inequalities, programs, and complementarity problems*, SIAM J. Control Optim., 25 (1987), pp. 583–595.
[24] G. NÜRNBERGER, *Strong unicity constants in Chebyshev approximation*, in Numerical Methods of Approximation Theory, Vol. 8, L. Collatz, G. Meinardus, and G. Nürnberger, eds., ISNM 81, 1987, Birkhäuser Verlag, Basel, pp. 144–168.
[25] M. R. OSBORNE, *Finite Algorithms in Optimization and Data Analysis*, Wiley Series in Probability and Mathematical Statistics, John Wiley, New York, 1985.
[26] M. R. OSBORNE AND R. S. WOMERSLEY, *Strong uniqueness in sequential linear programming*, J. Austral. Math. Soc. Ser. B, 31 (1990), pp. 379–384.
[27] B. T. POLYAK, *Introduction to Optimization*, Optimization Software, Inc., New York, 1987.
[28] S. M. ROBINSON, *Bounds for error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.
[29] ———, *An application of error bounds for convex programming in a linear space*, SIAM J. Control Optim., 13 (1975), pp. 271–273.
[30] S. M. ROBINSON, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.
[31] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[32] D. WALKUP AND R. J.-B. WETS, *A Lipschitzian characterization of convex polyhedra*, Proc. Amer. Math. Soc., 23 (1969), pp. 167–173.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 154–169, January 1994

009

A RELAXATION APPROACH APPLIED TO DOMAIN OPTIMIZATION*

R. B. GONZÁLEZ DE PAZ†

Abstract. A domain optimization problem related to potential theory is studied by means of a relaxation approach where a concave functional on a given convex set is defined. The functional has a minimizing point that is the characteristic function of an optimal domain. As a consequence of the necessary conditions of optimality, the domain is the solution of a free boundary value problem.

Key words. optimal design, optimal control, free boundary value problems, distributed parameter systems

AMS subject classifications. 49A22, 49B22, 35R35

1. Introduction. The mathematical aspects of the optimal design theory have been the subject of intensive research during the last decade. In its general setting, a functional which depends on a variable domain is given, and we consider its minimization on a certain class of feasible domains. To date, two main theoretical aspects have been studied. One problem concerns the existence of optimal domains. Here the main approach lies in the variational calculus framework, where a topology is defined so that the set of feasible domains is compact and the domain functional is lower semicontinuous. Some particular cases have been solved (cf. Chenais [11], Gonzalez de Paz [16], [17]), but others have been treated using some classes of "generalized domains" (cf. Murat and Tartar [22]).

The other main problem has been how to adapt differential calculus techniques in order to obtain some kind of gradient of the domain functional (cf. Cea [9], Pironneau [25], Simon [26], Zolesio [30]). However, to date, results achieved in both directions do not have a unified perspective. This work presents an approach that unifies the solution for both problems in a special case by means of a relaxation technique. We consider the problem in which we denote by Ω a domain in $\mathbb{R}^2$ that we assume to be doubly connected. We denote by Γ_0 and Γ the interior and exterior boundary of the domain Ω and by Ω_0 the domain bounded by Γ_0.

It is well known that for this case, the classical capacity problem is reduced to finding a potential function u_Ω such that:

$$\Delta u_\Omega = 0 \quad \text{in } \Omega, \tag{1.1}$$

$$u_\Omega = 0 \quad \text{on } \Gamma, \tag{1.2}$$

$$u_\Omega = 1 \quad \text{on } \Gamma_0. \tag{1.3}$$

The capacity of the domain Ω_0 related to Ω, noted $\mathrm{Cap}_\Omega(\Omega_0)$, is given by

$$\mathrm{Cap}_\Omega(\Omega_0) = \int_\Omega |\nabla u_\Omega|^2 \, d\omega.$$

Let us assume that the boundary Γ_0 and the following isoperimetric condition are given:

$$\text{meas } \Omega = A, \tag{1.4}$$

where A is a positive constant. We look for the shape of Ω such that $\mathrm{Cap}_\Omega(\Omega_0)$, the capacity of Ω_0 related to Ω, is minimized. Among others, this problem has been studied by Acker [1] and Aguilera, Alt, and Cafarelli [2].

* Received by the editors May 23, 1990; accepted for publication (in revised form) August 7, 1992.

† Departamento de Matemáticas, Universidad del Valle de Guatemala, Apdo. Postal 82, 01901 Guatemala, Guatemala.

We remark that by minimizing the functional

$$v \to J(v) = \int_{\Omega} |\nabla v|^2 \, d\omega$$

on a suitable function space, we obtain a solution for the problem (1.1)–(1.3). For the corresponding solution u_Ω we have

$$J(u_\Omega) = \int_{\Omega} |\nabla u_\Omega|^2 \, d\omega = \operatorname{Cap}_\Omega(\Omega_0).$$

Using this property, we define a new "relaxed" problem so that, by applying some convex analysis techniques, we prove the existence of the optimal domain Ω. In fact, the relaxed problem leads to the minimization of a concave function on a convex set of functions. The concavity structure will allow us to prove that there exists a characteristic function where the minimum is attained. This approach is similar to the one used by Gonzalez de Paz [17] for the study of the existence of a domain with minimal capacity when the interior boundary is unknown. The results presented were announced in [15]. In [16] a similar approach is applied to study the optimal design of elastic shafts.

Furthermore, we are able to calculate the derivative of the relaxed functional and, under certain regularity properties, we show that it is equivalent to the derivative given by other authors (cf. Simon [26], Pironneau [25], Zolesio [30]). It is worth remarking that the optimal domain solves a free boundary value problem treated by several authors (cf. Alt and Cafarelli [3], Acker [1]).

2. The relaxed problem. Let Ω_0 be a connected, bounded domain in $\mathbb{R}^2$, star-shaped related to the origin, with Lebesgue measure A_0 and boundary Γ_0 that is Lipschitz continuous. Let B_R be an open disc with center at some point in the interior of Ω_0. To allow for the feasible domains to be contained in the disc, we choose the radius R large enough so that for $d = \operatorname{dist}(\partial B_R, \Gamma_0)$, the annulus with outer boundary ∂B_R and width d has an area greater than the given constant A, and we put $D_R = B_R \backslash \overline{\Omega}_0$ and denote by $\|\|$ the usual L^2-norm in B_R. Furthermore, let μ be a nonegative function such that

$$0 \le \mu \le 1 \quad \text{almost everywhere in } B_R, \tag{2.1}$$

$$\int_{B_R} \mu \, d\omega = A_0 + A, \tag{2.2}$$

$$\int_{\Omega_0} \mu \, d\omega = A_0. \tag{2.3}$$

Here, $d\omega$ denotes the usual Lebesgue measure in $\mathbb{R}^2$. We remark also that the constraints (2.2) and (2.3) are equivalent to

$$\mu = 1 \quad \text{almost everywhere on } \Omega_0 \tag{2.2a}$$

and

$$\int_{D_R} \mu \, d\omega = A. \tag{2.3a}$$

Let $L^\infty(B_R)^+$ be the set of the nonnegative, bounded functions on B_R. We denote by C the convex subset of $L^\infty(B_R)^+$ defined by the constraints (2.1), (2.2), and (2.3).

Following the definitions introduced by Kinderlehrer and Stampacchia [20], we put

$$K_R = (v | v \in H_0^1(B_R), v \geq 1 \text{ on } \Omega_0);$$

here $H_0^1(B_R)$ denotes the usual Sobolev space, i.e., the completion of $C_0^\infty(B_R)$ related to the H^1-norm, which is defined by the application $u \to \|u\|^2 + \|\nabla u\|^2$ (cf. Neças [24]). Furthermore, the inequality $v \geq 1$ on Ω_0 must be understood in the sense of H^1, i.e., there exists a sequence of Lipschitz functions $(u_n) \subset \text{Lip}(B_R)$ such that

$$u_n \geq 1 \quad \text{on } \Omega_0$$

and

$$u_n \to v \quad \text{strongly in } H^1$$

(cf., for example, Kinderlehrer and Stampacchia [20]).

For a fixed $\mu \in L^\infty(B_R)^+$ and a fixed constant $\epsilon > 0$, we now define on the Sobolev space $H_0^1(B_R)$ the functional

$$v \to J_\mu(v) = \frac{1}{2} \int_{B_R} |\nabla v|^2 \, d\omega - \epsilon \int_{B_R} \mu \, v \, d\omega. \tag{2.4}$$

The minimization of $v \to J_\mu(v)$ on K_R was treated by Kinderlehrer and Stampacchia [20]. This functional is convex and weakly lower semicontinuous so that for a fixed $\mu \in L^\infty(B_R)^+$ there exists an element $u_\mu \in K_R$ such that the functional is minimized (cf. Ekeland and Temam [12], Moreau [23]).

The problem $P(\mu)$. As a consequence of classical arguments, the minimizer $u_\mu \in K_R$ is the weak solution of the following boundary value problem, which we denote herein as the problem $P(\mu)$:

$$-\Delta u_\mu = \mu\epsilon \quad \text{in } D_R = B_R \backslash \overline{\Omega}_0 \quad \text{in the weak sense} \tag{2.5}$$

$$u_\mu = 1 \quad \text{on } \Omega_0 \quad \text{in the } H^1 \text{sense} \tag{2.6}$$

$$u_\mu = 0 \quad \text{on } \partial B_R \quad \text{in the sense of traces.} \tag{2.7}$$

Remark 2.1. Recall that the functional $v \to J_\mu(v)$ is strictly convex, so that the solution $u_\mu \in K_R$ is unique.

Remark 2.2. The element u_μ is a nonnegative function. To prove this, define

$$u_\mu^+ = \max(u_\mu, 0).$$

This is an element of $H_0^1(B_R)$ (cf. Kinderlehrer and Stampacchia [20]). Moreover, because of the extremality property of u_μ we have $u_\mu^+ \in K_R$ and

$$\int_{B_R} \mu u_\mu^+ \, d\omega = \int_{B_R} \mu u_\mu \, d\omega.$$

If u_μ were strictly negative on a set of positive measure, then

$$\|\nabla u_\mu^+\| < \|\nabla u_\mu\|,$$

which implies

$$J_\mu(u_\mu^+) < J_\mu(u_\mu).$$

This is a contradiction, therefore $u_\mu^+ = u_\mu$.

Remark 2.3. The function u_μ is an element of $C^{1,1}_{\text{loc}}(\overline{D}_R)$. First we recall that $u_\mu \in C^{1,\alpha}(\overline{D}_R)$ for $0 \leq \alpha < 1$ (see Kinderlehrer and Stampacchia [20]).

It follows that u_μ is a Lipschitz function. Furthermore, $\Delta u_\mu \in L^\infty(D_R)$. From these results and the boundary conditions (2.6) and (2.7) it follows that

$$u_\mu \in W^{2,\infty}_{\text{loc}}(D_R)$$

(cf. Gebhardt [14] and Jensen [18]). This implies that ∇u_μ is a locally Lipschitz function (for the definition of $W^{2,\infty}(D_R)$, cf. Neças [24]). In the case where we have more regularity on the boundary $\partial\Omega_0$, for instance, $C^{2,\alpha}$-regularity, then $u_\mu \in W^{2,\alpha}(D_R)$ (cf. Frehse [13]).

The optimization problem related to μ. We now define the functional Φ on $L^\infty(B_R)^+$ as follows:

$$\Phi(\mu) = J_\mu(u_\mu) = \min_{u \in K} J_\mu(u).$$

We study the problem of minimization of Φ in $C \subset L^\infty(B_R)^+$ where C denotes the convex set defined by the constraints (2.1), (2.2), and (2.3). The convex set C is compact for the topology $\sigma(L^\infty, L^1)$. We prove that the functional Φ is continuous for the same topology in order to show the existence of a minimizing element.

THEOREM 2.1. *The functional Φ is $\sigma(L^\infty, L^1)$-continuous on C.*

Proof. First we establish the following assertion: There exists a ball B_ϱ in $H_0^1(B_R)$ of radius ϱ such that, for every $\mu \in C$,

$$\min_{u \in K_R} J_\mu(u) = \min_{u \in K_R \cap B_\varrho} J_\mu(u).$$

Let μ be given, and let u_μ be the corresponding minimizing element of J_μ in K_R. Then for every $v \in K_R$ we have

$$(\nabla u_\mu, \nabla v) = \epsilon(\mu, v).$$

Here the parentheses denote the usual scalar product in $L^2(B_R)$. For the special case $v = u_\mu$,

$$\|\nabla u_\mu\|^2 = \epsilon(\mu, u_\mu) \leq \epsilon \|\mu\|_{L^\infty} \|u_\mu\|_{L^1},$$

and by using the Cauchy–Schwarz and Poincaré inequalities,

$$\|u_\mu\|_{L^1} \leq \alpha \|u_\mu\| \leq \alpha' \|\nabla u_\mu\|,$$

where α and α' are constants depending on the ball B_R. Then we obtain, for every $\mu \in C$,

$$\epsilon \|\mu\|_{L^\infty} \leq \epsilon,$$

and finally,

$$\|\nabla u_\mu\| \leq \alpha' \epsilon,$$

so the expected ball has radius $\varrho = \alpha' \epsilon$.

Because of the Rellich–Kondrasov injection theorem, the set $K' = K_R \cap B_\varrho$ is compact in $L^1(B_R)$ (cf. Neças [24]).

Note that applying integration by parts we have

$$\Phi(\mu) = -\frac{\epsilon}{2}\int_{B_R} \mu u_\mu \, d\omega.$$

Let us give a sequence $(\mu_n)_n \subset C$ converging to an element $\mu\tilde{}$ in C in the $\sigma(L^\infty, L^1)$-topology. For the corresponding sequence $(u_n)_n$ and $u\tilde{}$ solutions of the problems $P(\mu_n)$ and $P(\mu\tilde{})$, we have for every test function $\varphi \in C_0^\infty(D_R)$,

$$(\nabla(u_n - u\tilde{}), \nabla\varphi) = \epsilon(\mu_n - \mu\tilde{}, \varphi).$$

This implies that $u_n \to u\tilde{}$ weakly in $H_0^1(B_R)$. Due to the Rellich–Kondrasov theorem, the sequence also converges strongly in $L^1(B_R)$. We now write,

$$\langle \mu_n, u_n\rangle - \langle \mu\tilde{}, u\tilde{}\rangle = \langle \mu_n - \mu\tilde{}, u_n\rangle + \langle \mu\tilde{}, u_n - u\tilde{}\rangle.$$

The brackets describe the (L^∞, L^1)-duality. The second term on the right-hand side of the equation converges to zero due to the strong convergence in $L^1(B_R)$. For the first term we recall that on the unit ball of L^∞ the $\sigma(L^\infty; L^1)$ convergence is equivalent to the uniform convergence on the compact subsets of $L^1(B_R)$ (cf. Bourbaki [6, Chap. 4]). As every u_n is in the L^1-compact set K', this term also converges to zero and it follows that the functional Φ is continuous on C.

We remark that the result of Theorem 2.1 may be restated the following way.

COROLLARY 2.2. *Let $B \subset L^\infty(B_R)$ be the unit ball. For the boundary value problem $P(\mu)$ described above, the corresponding Green operator $(\Delta)^{-1} : B \to H_0^1(B_R)$ is continuous related to the $\sigma(L^\infty, L^1)$-topology and the norm topology in $H_0^1(B_R)$, respectively.*

Furthermore, as $C \subset L^\infty(B_R)^+$ is $\sigma(L^\infty, L^1)$-compact, we have as a consequence the following corollary.

COROLLARY 2.3. *There exists an element $\mu_R \in C$ such that*

$$\Phi(\mu_R) = \min_{\mu \in C} \Phi(\mu).$$

THEOREM 2.4. *There exists a set $\Omega_R \subset D_R$ such that μ_R is the characteristic function of the set $\tilde{\Omega}_R = \Omega_0 \cup \Omega_R$.*

Proof. The functional Φ is the lower envelope of affine linear functions so that it is concave. This implies that among the minimizing elements there are extremal points of C, and these are characteristic functions of sets with measure $A + A_0$ (cf. Castaing and Valadier [7]). So there exists $\mu_R = \chi_{\tilde{\Omega}_R}$ with $\tilde{\Omega}_R = \Omega_0 \cup \Omega_R$.

We denote Ω_R as an optimal set, as its characteristic function describes a minimizing element for the relaxed problem. The necessary conditions of optimality are studied in order to obtain a description of the optimal domain as the solution of a free boundary value problem.

3. Necessary conditions of optimality and their consequences. In order to study the optimality conditions we must calculate the Gateaux directional derivative $\Phi'(\mu; \alpha)$. Classically, for every direction $\alpha \in L^\infty(B_R)$ it is defined as follows:

$$\Phi'(\mu; \alpha) = \lim_{t \to 0_+} 1/t(\Phi(\mu + t\alpha) - \Phi(\mu)).$$

THEOREM 3.1. *The functional Φ has a weak derivative in the sense of Gateaux for every $\mu \in L^\infty(B_R)^+$ and every direction $\alpha \in L^\infty(B_R)$. It has the following form:*

$$\Phi'(\mu; \alpha) = -\epsilon\langle u_\mu, \alpha\rangle_{L^1, L^\infty}.$$

Proof. We recall that

$$\Phi(\mu) = \inf_{u \in K} \frac{1}{2}\|\nabla u\|^2 - \epsilon\langle u, \mu\rangle = \frac{1}{2}\|\nabla u_\mu\|^2 - \epsilon\langle u_\mu, \mu\rangle.$$

With Φ being the lower envelope of a family of affine functions related to μ and K' a compact set, it follows from a theorem of Valadier [28] that

$$\Phi'(\mu; \alpha) = -\epsilon\langle u_\mu, \alpha\rangle_{L^1, L^\infty} \tag{3.1}$$

for every $\alpha = \gamma - \mu$ with $\gamma \in L^\infty(B_R)^+$.

Remark 3.1. Φ is concave and $\sigma(L^\infty, L^1)$-continuous, so it follows that its derivative is Fréchet (cf. Valadier [28]).

Remark 3.2. Herein let us denote $u_R = u_{\mu_R}$ as the corresponding solution for the boundary value problem $P(\mu_R)$ described by conditions (2.5)–(2.7).

The first-order necessary conditions of optimality give, for every $\alpha = \mu - \mu_R, \mu \in C$, (cf. Cea [8])

$$-\langle u_R, \alpha\rangle_{L^1, L^\infty} \geq 0. \tag{3.2}$$

If we restrict ourselves to characteristic functions of sets $\mu = \chi_{\tilde{\Omega}}$ such that $\tilde{\Omega} = \Omega_0 \cup \Omega$, we obtain, for every domain Ω in D_R with measure equal to A and such that Γ_0 is contained in $\partial\Omega$,

$$-\langle u_R, \chi_{\tilde{\Omega}} - \chi_{\tilde{\Omega}_R}\rangle \geq 0, \tag{3.3}$$

which is equivalent to

$$\int_{\Omega_R} u_R \, d\omega \geq \int_{\Omega} u_R \, d\omega. \tag{3.4}$$

Inequality (3.4) states that the integrand u_R must be "placed" in D_R so that the integral has a maximal value. We denote by $\Gamma = \partial\Omega_R \cap D_R$ the boundary of Ω_R related to D_R. The set Γ can be interpreted as a free boundary as a consequence of the following theorem.

THEOREM 3.2. *For the optimal set Ω_R there exists a positive number p_R such that*

$$\Omega_R = \{x \in D_R | u_R(x) > p_R\},$$

where the equality is understood to hold up to a null measure set, and

$$\Gamma = \{x \in D_R | u_R(x) = p_R\}.$$

Proof. The existence of a Lagrange multiplier related to the constraint (2.2) for the functional $\mu \to \int_{B_R} u_R \mu \, d\omega$ is a classical fact (cf. Cea and Malanowski [10]). This means that there exists a constant p_R such that, for all elements γ in $L^\infty(B_R)^+$ such that $0 \leq \gamma \leq 1$,

$$\int_{B_R} \mu u_R \, d\omega - p_R \int_{B_R} \mu \, d\omega \geq \int_{B_R} \gamma u_R \, d\omega - p_R \int_{B_R} \gamma \, d\omega.$$

Then we have, for almost every $x \in B_R$,

$$\begin{aligned} u_R(x) > p_R \quad &\text{implies } \mu(x) = 1, \\ u_R(x) < p_R \quad &\text{implies } \mu(x) = 0. \end{aligned}$$

Setting $\gamma = \chi_{\tilde{\Omega}_R}$ with $\tilde{\Omega}_R = \Omega_0 \cup \Omega_R$, and recalling that μ fulfills constraint (2.2), we have $\mu = \chi_{\tilde{\Omega}_R}$.

We define $G = B_R \backslash \tilde{\Omega}_R$ and it follows that

$$\{x \in B_R | u_R(x) > p_R\} \subset \tilde{\Omega}_R,$$
$$\{x \in B_R | u_R(x) < p_R\} \subset G.$$

Both inclusions must be understood up to a null measure set. Furthermore, we have the following inclusions up to a null measure set:

$$\tilde{\Omega}_R \subset \{x \in B_R | u_R(x) \geq p_R\},$$

which implies that

$$\Omega_R \subset \{x \in D_R | u_R(x) \geq p_R\}.$$

From the definition of Ω_R it follows that

$$\{x \in D_R | u_R(x) > p_R\} \subset \Omega_R.$$

A result due to Stampacchia [20] states that if $u_R \in H^1(D_R)$ is constant on a measurable set E, then $\nabla u_R = 0$ almost everywhere on E. Furthermore, if $u_R \in H^2(D_R)$ it follows that $\Delta u_R = 0$ almost everywhere on E. As equation (2.5) is verified in the sense almost everywhere on D_R, this implies that meas $(\{x \in D_R | u_R(x) = p_R\} \cap \Omega_R) = 0$ and the first assertion of the theorem is proved.

The characterization of Γ follows from the fact that the function u_R is continuous and superharmonic in D_R (cf. Gonzalez de Paz [17]).

COROLLARY 3.3. *The support of the measure $\mu_R d\omega$ is the compact set*

$$\tilde{\Omega}_R = \{x \in B_R | u_R(x) \geq p_R\}.$$

Remark 3.3. From the boundary condition (2.6) we have

$$u_R \geq p_R \quad \text{on } \Omega_0.$$

Remark 3.4. The function $u_R \in H_0^1(B_R) \cap C_{\text{loc}}^{1,1}(\overline{B}_R)$ is a solution of the following free boundary value problem:

$$-\Delta u = \epsilon \quad \text{in } \Omega_R \quad \text{in the weak sense,} \tag{3.5}$$
$$\Delta u = 0 \quad \text{in } D_R \backslash \overline{\Omega}_R, \tag{3.6}$$
$$u = p \quad \text{on } \Gamma, \tag{3.7}$$
$$u = 1 \quad \text{on } \Gamma_0. \tag{3.8}$$

Remark 3.5. We should point out that in the case where Ω_0 is not star shaped, $D_R \backslash \overline{\Omega}_R$ might have more than one connected component, which would indicate the existence of more "holes" in the domain. On the other hand, because of the maximum principle, the domain Ω_R is connected.

Remark 3.6. The gradient of u_R is continuous so that

$$(\nabla u_R)^+ = (\nabla u_R)^- \quad \text{on } \Gamma, \tag{3.9}$$

where the plus sign denotes the limit at the boundary taken in the inward direction to Ω_R and the minus sign denotes the limit in the outward direction.

Because of the regularity of u_R, it is known that free boundaries of this type are locally Lipschitz (cf. Kinderlehrer and Stampacchia [20]). If we recall the fact that the free boundary Γ is a level set of u_R, then we have in the neighborhoods of points where $|\nabla u_R| > 0$ on Γ,

$$\partial u_R/\partial n^+ = \partial u_R/\partial n^- \quad \text{on } \Gamma. \tag{3.10}$$

The condition (3.10) can be interpreted as a "transmission" condition on the free boundary.

Calculation of the regular domain functional derivative. Some analog-free boundary value problems such as the one considered above have been studied by Zolesio [29] using other optimal design techniques. Under suitable regularity assumptions, we show that our approach is related to this one.

Let $\varphi_t : B_R \to B_R$ be a C^k-diffeomorphism (k a fixed integer) that is continuously dependent on a positive parameter t, and let φ_0 be the identity application, so that for any given domain $\Omega \subset B_R$, defined as before, we have $\varphi_0(\Omega) = \Omega$. Furthermore, let us denote $\varphi_t(\Omega) = \Omega_t$, so that the application $t \to \varphi_t$ describes continuous deformations of Ω. Using a certain topology, and assuming that the application is differentiable, Zolesio in [30] calculates the derivative $D_t\varphi_t = \theta$; here θ is the vector field describing the "deformation speed." For the domain functionals $\Omega_t \to J(\Omega_t)$, the shape derivative of $J(\Omega_t)$ at Ω in the direction of the field θ is defined:

$$dJ(\Omega;\theta) = \limsup_{t\to 0^+} \{J(\Omega_t) - J(\Omega)\}/t.$$

A similar case was treated by Simon [26]: Given a regular vector field $\theta \in C^k(B_R)$, for small parameter $t > 0$ the diffeomorphism stated above is given by the application $\varphi_t = Id + t\theta$, and the derivative of domain functionals is calculated. In the case of the Dirichlet integral $\Omega \to \int_\Omega |\nabla u_\Omega|^2 \, d\omega$ where u_Ω is the corresponding capacitary potential function, the result is already classic and it is known as the Hadamard formula. We have, as the shape derivative,

$$dJ(\Omega;\theta) = -\frac{1}{2}\int_\Gamma |\partial u_\Omega/\partial n|^2 \langle n, \theta\rangle \, ds.$$

Here the term $\langle n, \theta\rangle$ describes the normal component of the vector field θ to Γ.

THEOREM 3.4. *Let $\varphi_t : B_R \to B_R$ be as above, and let Γ be a smooth curve. Then for $\alpha_t = \chi_{\Omega_t} - \chi_\Omega$ we have*

$$\lim_{t\to 0^+} \frac{1}{t}\Phi'(\chi_\Omega, \alpha_t) = \frac{1}{2}\int_\Gamma (|\partial u_e/\partial n|^2 - |\partial u_i/\partial n|^2)\langle \theta, n\rangle \, ds. \tag{3.11}$$

Here the restrictions of u_R on Ω_R and $D_R\backslash\overline{\Omega}_R$ are denoted by u_i and u_e, respectively.

Proof. We put $\mu_t = \chi_{\Omega_t}$ and $\mu_0 = x_\Omega$, and let u_t and u_0 elements of $H_0^1(B_R)$ be the solutions of the boundary value problems $P(\mu_t)$ and $P(\mu_0)$, respectively; then $\delta u = u_t - u_0$ is the weak solution of $-\Delta v = \epsilon\alpha_t$, where $\alpha_t = \mu_t - \mu_0$.

Recall that

$$\Phi(\mu) = \min_{u\in K} J_\mu(u) = -\frac{1}{2}\epsilon\int_{B_R} \mu u_\mu \, d\omega.$$

Moreover, integrating by parts we obtain

$$\int_{B_R} \mu_t u_0 \, d\omega = \int_{B_R} \mu_0 u_t \, d\omega$$

so that, if we calculate,

$$-\epsilon\int_{B_R} u_0\alpha_t\,d\omega - \frac{1}{2}\epsilon\int_{B_R}\alpha_t\delta u\,d\omega = -\frac{1}{2}\epsilon\int_{B_R}\mu_t u_t\,d\omega + \frac{1}{2}\epsilon\int_{B_R}\mu_0 u_0\,d\omega$$
$$= \Phi(\mu_t) - \Phi(\mu_0).$$

This implies that

$$\lim_{t\to 0^+}\frac{1}{t}[\Phi(\mu_t) - \Phi(\mu_0)] = \lim_{t\to 0^+}\frac{1}{t}[-\langle\epsilon u_0, \mu_t - \mu_0\rangle - -\frac{1}{2}\langle\epsilon(u_t - u_0), \mu_t - \mu_0\rangle].$$

If $t \to 0^+$, then $\mu_t \to \mu_0$ in the $\sigma(L^\infty, L^1)$-topology and $u_t \to u_0$ strongly in $L^2(B_R)$, so that the second term on the right-hand side vanishes. If we assume that the boundary Γ is smooth enough, we can apply Zolesio's techniques (cf. [30]). Thus we can write, in the case where $\mu = \chi_{\tilde{\Omega}_R}$,

$$\Phi(\mu) = \frac{1}{2}\int_{B_R}|\nabla u_\mu|^2\,d\omega - \epsilon\int_{B_R}\mu u_\mu\,d\omega = J_1(\Omega_R) + J_2(D_R\backslash\overline{\Omega}_R),$$

where

$$J_1(\Omega_R) = \frac{1}{2}\int_{\Omega_R}|\nabla u_\mu|^2\,d\omega - \epsilon\int_{\tilde{\Omega}_R} u_\mu\,d\omega$$
$$J_2(D_R\backslash\overline{\Omega}_R) = \frac{1}{2}\int_{D_R\backslash\Omega_R}|\nabla u_\mu|^2\,d\omega.$$

We have then, as shape derivatives,

$$dJ_1(\Omega_R;\theta) = -\frac{1}{2}\int_\Gamma |\partial u_i/\partial n|^2\,\langle\theta, n\rangle\,ds$$
$$dJ_2(D_R\backslash\overline{\Omega}_R;\theta) = \frac{1}{2}\int_\Gamma |\partial u_e/\partial n|^2\,\langle\theta, n\rangle\,ds,$$

where n describes the normal vector to the curve Γ. This gives, as a result,

$$\lim_{t\to 0^+}\frac{1}{t}[\Phi(\mu_t) - \Phi(\mu_0)] = \lim_{t\to 0^+}\frac{1}{t}\Phi'(\mu_0, \alpha_t)$$
$$= \frac{1}{2}\int_\Gamma(|\partial u_e/\partial n|^2 - |\partial u_i/\partial n|^2)\langle\theta, n\rangle\,ds.$$

4. The limit case for an unbounded domain. Our aim in this section is to prove that, when the ball radius increases to infinity, the corresponding solutions of our "relaxed problems" converge to the solution of the original problem.

For Ω_0 as defined in the Introduction, we note by W the set of all doubly connected domains Ω with a given measure A and Γ_0 as an inner boundary and a Lipschitz-continuous outer boundary Γ. Let us recall that for a given domain $\Omega \in W$, we have the corresponding weak solution $u_\Omega \in H_0^1(\Omega)$ of the following boundary value problem:

(4.1) $\quad -\Delta u = \mu\epsilon \quad$ in Ω $\quad$ in the weak sense

(4.2) $\quad u = 1 \quad$ on Ω_0 $\quad$ in the H^1 – sense

(4.3) $\quad u = 0 \quad$ on Γ $\quad$ in the sense of traces.

For the case of a ball B_R with radius R such that $\Omega \subset B_R$, if the boundary $\Gamma = \partial\Omega \cap D_R$ is smooth enough, the function u_Ω can be extended to a function $\tilde{u} \in H_0^1(B_R)$ such that this extension is zero on $B_R \backslash (\Omega \cup \Omega_0)$. It is clear that if μ is the characteristic function of $\tilde{\Omega} = \Omega_0 \cup \Omega$, for any feasible R we can calculate for $\tilde{u}$ as defined above:

$$J_\mu(\tilde{u}) = \frac{1}{2}\int_{B_R} |\nabla \tilde{u}|^2 \, d\omega - \epsilon \int_{B_R} \mu \tilde{u} \, d\omega = \frac{1}{2}\int_{\tilde{\Omega}} |\nabla u_\Omega|^2 \, d\omega - \epsilon \int_{\tilde{\Omega}} u_\Omega \, d\omega.$$

So formally, we write $J_\mu(\tilde{u}) = J_\mu(u_\Omega)$. This gives the motivation to define for each $\Omega \in W$ the energy functional $\Omega \to En(\Omega) = J_\mu(u_\Omega)$ with $\mu = \chi_{\tilde{\Omega}}$, which is independent of the radius R. For every $\chi_{\tilde{\Omega}} \in C$ as above, with $\Omega \in W$, there exists a ball B_R and a $\mu_R \in C$ such that

$$\Phi(\mu_R) \leq \Phi(\mu) \leq J_\mu(u_\Omega) = En(\Omega).$$

Let us take an increasing sequence of R_n such that its limit is infinity. By taking extensions of $H_0^1(B_R)$ into $H^1(\mathbb{R}^2)$, we have for $R_{n'} > R_n$ the inclusions $K_{R_n} \subset K_{R_{n'}}$, and for the corresponding minimizers μ_R as defined in Corollary 2.3 this implies that

$$\Phi(\mu_{R_{n'}}) \leq \Phi(\mu_{R_n}).$$

The sequence $(\Phi(\mu_{R_n}))_n$ is monotone decreasing; we prove that a limit does exist. We denote as Y the completion of the space of test functions in $\mathbb{R}^2$ with the L^2-norm of the gradient. Among others, this space has been used by Temam [27] to study partial differential equations in unbounded domains.

Remark 4.1. The following injections are continuous:

$$H_0^1(\mathbb{R}^2) = H^1(\mathbb{R}^2) \subset Y,$$

$$Y \subset L^\alpha_{\text{loc}}(\mathbb{R}^2)$$

for every $\alpha \geq 1$. Moreover, the injection

$$H^1_{\text{loc}}(\mathbb{R}^2) \subset L^\alpha_{\text{loc}}(\mathbb{R}^2)$$

is compact for every $1 \leq \alpha < \infty$. These injections remain valid for the unbounded domain $D = \mathbb{R}^2 \backslash \overline{\Omega}_0$.

LEMMA 4.1. *There exists a ball B_0 with radius R_0 such that for every element of a subsequence $(R_{nk})_k \subset (R_n)_n$ the optimal domains $\Omega_{R_{nk}}$ are included in B_0.*

Proof. Let C be the convex set of functions as defined before but extended to $\mathbb{R}^2$. As for every $R_n : \|\mu_{R_n}\|_{L^2}^2 = A + A_0$, there exists a subsequence also noted $\{\mu_{R_n}\}_n$ that converges weakly to a $\mu^* \in L^2(\mathbb{R}^2)$. We remark that, because the μ_{R_n} are characteristic functions of finite measure sets, it follows for every $R_n : \mu_{R_n} \in L^1(\mathbb{R}^2)$ with $\|\mu_{R_n}\|_{L^1} = A + A_0$. Following the proof of the concentration-compactness lemma given by Lions [21] we consider the function $Q_n(R) = \sup_{x \in \mathbb{R}^2} \int_{x+B_R} \mu_{R_n} \, d\omega$ and we have that $\lim_{R\to\infty} Q_n(R) = A + A_0$. The sequence $(Q_n)_n$ is nondecreasing, nonnegative, and uniformly bounded, so that classically there exists a subsequence $(n_k)_k$ and a nondecreasing, nonnegative function Q such that $Q_{nk} \to Q$ pointwise. Because of the $\sigma(L^\infty, L^1)$-convergence of the original sequence $(\mu_{R_n})_n$, it follows that $Q(R) = \sup_{x \in \mathbb{R}^2} \int_{x+B_R} \mu^* \, d\omega$; and because of the integral constraint (2.2), $\lim_{R\to\infty} Q(R) = A + A_0$. In this case, there exists a ball B_0 with radius R_0

such that for every k, $\int_{B_0} \mu_{nk}\, d\omega = A + A_0$ and our assertion follows. Without loss of generality, we let R_0 be large enough so that $x = 0$.

THEOREM 4.2. *Let the assumptions on the sequence $(R_n)_n$ be as above. Then there exist elements $u^* \in Y$ and $\mu^* \in C$ such that*

$$\frac{1}{2}\int_{\mathbb{R}^2} |\nabla u^*|^2\, d\omega - \epsilon \int_{\mathbb{R}^2} \mu^* u^*\, d\omega = \inf_{R_n} \Phi(\mu_{R_n}).$$

Proof. We note that $\tilde{u}_{R_n}$ is the natural extension of u_{R_n} in $H^1(\mathbb{R}^2)$. Recall that, for every $\varphi \in K_R$ and for every R,

$$(\nabla u_R, \nabla \varphi) = \epsilon(\mu_R, \varphi). \tag{4.4}$$

Let us define the set $K = \{\varphi \in Y | \varphi \geq 1 \text{ on } \Omega_0\}$. Due to the $\sigma - L^2$-convergence of the sequence $(\mu_{R_n})_{R_n}$, we may assume the existence of a distribution u^* such that, if $R_n \to \infty$, we have for every $\varphi \in K$

$$(\nabla \tilde{u}_{R_n}, \nabla \varphi) \to (\nabla u^*, \nabla \varphi) = \epsilon(\mu^*, \varphi) \tag{4.5}$$

so that, for the weak topology of $L^2(\mathbb{R}^2)$,

$$\nabla \tilde{u}_{R_n} \to \nabla u^* \in L^2(\mathbb{R}^2)$$

and $u^* \in Y$ is a weak solution for the Poisson equation (cf. related results in Bottaro and Marina [5])

$$-\Delta u^* = \epsilon \mu^* \quad \text{in } D. \tag{4.6}$$

The sequence $(\tilde{u}_{R_n})_n$ converges weakly in Y; because of the inclusions quoted in Remark 4.1, $u^* \in L^2_{\text{loc}}(\mathbb{R}^2)$. This means that the functions $\tilde{u}_{R_n}$ and u^* are elements of $H^1_{\text{loc}}(\mathbb{R}^2)$. Recalling again Remark 4.1 it follows that if $R_n \to \infty$, for $n \to \infty$ then

$$\tilde{u}_{R_n} \to u^* \quad \text{strongly in } L^2_{\text{loc}}(\mathbb{R}^2).$$

As for every R_n:

$$\int_{\mathbb{R}^2} \mu_{R_n} \tilde{u}_{R_n}\, d\omega = \int_{B_o} \mu_{R_n} \tilde{u}_{R_n}\, d\omega. \tag{4.7}$$

The strong $L^2_{\text{loc}}(\mathbb{R}^2)$-convergence implies that

$$\int_{\mathbb{R}^2} \mu_{R_n} \tilde{u}_{R_n}\, d\omega \to \int_{\mathbb{R}^2} \mu^* u^*\, d\omega. \tag{4.8}$$

We deduce from (4.5), (4.7), and (4.8) that

$$\|\nabla \tilde{u}_{R_n}\| \to \|\nabla u^*\|.$$

The sequence $(\nabla u_{R_n})_n$ converges strongly in $L^2(\mathbb{R}^2)$ and consequently,

$$\Phi(\mu_{R_n}) \to \Phi(\mu^*) = \inf_{R_n} \Phi(\mu_{R_n}).$$

Existence of the optimal domain. We now define $\tilde{\Omega}^* = (x \in \mathbb{R}^2 | \mu^* > 0)$ in the sense of measures, and we put $\Omega^* = \tilde{\Omega}^* \backslash \overline{\Omega}_0$.

Remark 4.2. $D\backslash\overline{\Omega}^*$ is an open set of D, and as a consequence of the differential equation (4.6) the function u^* is harmonic in $D\backslash\overline{\Omega}^*$.

LEMMA 4.3. *Let $(\tilde{u}_{R_n})_n \subset H^1(\mathbb{R}^2)$ be the same sequence as before. Then if $R_n \to \infty$, it follows that*

$$\int_D |\nabla \tilde{u}_{R_n}|^2 \, d\omega \to \int_{\Omega^*} |\nabla u^*|^2 \, d\omega.$$

Proof. Recall that the function u^* is harmonic in $D\backslash\overline{\Omega}^*$; secondly, for every R_n we have, for $\tilde{\Omega}_{R_n} = \Omega_0 \cup \Omega_{R_n}$,

$$\int_{D_{R_n}} |\nabla u_{R_n}|^2 \, d\omega = \int_{D_{R_n}\backslash\Omega_{R_n}} |\nabla u_{R_n}|^2 \, d\omega + \int_{\Omega_{R_n}} |\nabla u_{R_n}|^2 \, d\omega$$

and

$$\int_{D_{R_n}\backslash\Omega_{R_n}} |\nabla u_R|^2 \, d\omega \leq \gamma \mathrm{Cap}_{B_{R_n}}(\tilde{\Omega}_{R_n}), \tag{4.9}$$

where γ is a constant depending only on the boundary condition on Γ_0 and $\mathrm{Cap}_{B_R}(\tilde{\Omega}_R)$ is the capacity of the set $\tilde{\Omega}_R$ related to the ball B_R. Let us recall that for the ball B_0 as in Lemma 4.1, and for every $R_{nk} \geq R_0$,

$$\Omega_{R_{nk}} \subset B_0 \subset B_{R_{nk}}.$$

Then, it follows that $\mathrm{Cap}_{B_{R_{nk}}}(\tilde{\Omega}_{R_{nk}}) \leq \mathrm{Cap}_{B_{R_{nk}}}(B_0)$.

It is known that for the two-dimensional case, $\mathrm{Cap}_{B_R}(B_0) = 0(1/\log(1/R))$, so that if $R_{nk} \to \infty$ the capacity tends to zero. Consequently, from inequality (4.9) we have that

$$\int_{D\backslash\Omega^*} |\nabla u^*|^2 \, d\omega = 0.$$

COROLLARY 4.4. *The function u^* is equal to zero in the unbounded component of $D\backslash\overline{\Omega}^*$.*

Proof. As a consequence of Lemma 4.3, u^* is constant almost everywhere in $D\backslash\overline{\Omega}^*$. We note this constant on the unbounded component as p^*. If we define $v_{R_n} = u^* - \tilde{u}_{R_n} \in C(D)$, it is clear that each function is uniformly bounded at infinity and, for every R,

$$\lim_{|x|\to\infty} v_R(x) = p^*.$$

The sequence $(v_{R_n})_n$ has a limit v^* for the norm topology in Y and each v_{R_n} is a harmonic function in the complement of B_{R_n}. Moreover, v^* is bounded at infinity.

Let R_0 be as in Lemma 4.1, and let $B_{R'}$ be another ball such that $R' > R_0$, so that all the elements of the sequence $(\tilde{u}_{R_n})_{R_n \geq R'}$ are harmonic in the interior of the set $S = B_{R'}\backslash B_{R_o}$.

Because of Remark 4.1, we have that

$$\tilde{u}_{R_n} \to u^* \quad \text{strongly in } L^2(S).$$

It is known that, for every closed set $S' \subset S$, the sequence of harmonic functions converges uniformly (cf. Kellog [19]). Consequently $v^* = 0$ pointwise in S'. Let $\Sigma \subset S'$ be a regular continuous curve and let D_Σ be an unbounded, open domain contained in the complement

of B_{R_o} with Σ as boundary. Then v^* is a harmonic function equal to zero on Σ and bounded at infinity; it is a classical fact that v^* must be equal to zero everywhere in D_Σ. The function v^* is continuous and the limit of a sequence in $C(D_\Sigma)$, so it follows that, for every $x \in D_\Sigma$,

$$v^*(x) = \lim_{R_n \to \infty} v_{R_n}(x) = \lim_{R_n \to \infty} u^*(x) - \tilde{u}_{R_n}(x) = 0,$$

but the second term $\tilde{u}_{R_n}$ vanishes when $|x| \to \infty$, and this implies that

$$\lim_{|x| \to \infty} v^*(x) = \lim_{|x| \to \infty} \lim_{R_n \to \infty} v_{R_n}(x) = \lim_{|x| \to \infty} u^*(x) = p^*.$$

Thus p^* must be the null constant.

We further define the set $\Omega^+ = \{x \in D | u^*(x) > 0$ in the sense of $H^1\}$. We recall that $u(x) > 0$ at a point x in the sense of H^1 if there exists a ball centered in x with radius ϱ noted $B(x, \varrho)$ and $\varphi \in$ Lip $(B(x, \varrho)); \varphi(x) > 0$ such that $u - \varphi \geq 0$ in the sense of H^1. As a consequence of the definition the set Ω^+ is open. It is clear that up to a null measure set $\Omega^* \subset \Omega^+$.

Remark 4.3. The function u^* is an element of $C^{1,\alpha}(\Omega^+)$. This follows because $-\Delta u^* = \epsilon \mu^*$ almost everywhere in Ω^+. The right-hand term is a bounded function. Recalling regularity properties quoted in Remark 2.2, we obtain $u^* \in C^{1,\alpha}(\Omega^+)$.

THEOREM 4.5. *The element $\mu^* \in C$ has the property $\mu^* = \chi_{\Omega^*}$, and for all domains $\Omega \in W$,*

$$En(\Omega^*) \leq En(\Omega).$$

Proof. We put $K = \{u \in Y | u \geq 1$ on $\Omega_0\}$. For a given $\Omega \in W$ we have $\tilde{\Omega} = \Omega \cup \Omega_0$ and $\tilde{\Omega} \subset B_R$ for a certain R. Furthermore, we define the set $K_0(\Omega) = \{u \in K | u = 0$ in $\mathbb{R}^2 \backslash \tilde{\Omega}\}$, which is contained in K_R, closed and convex. The functional $\Omega \to$ En(Ω) can be written for $\mu = \chi_{\tilde{\Omega}}$ as follows:

$$\text{En } (\Omega) = \min_{K_o(\Omega)} J_\mu(u),$$

so that for every $\Omega \in W$ and every $\mu = \chi_{\tilde{\Omega}} \in C$,

$$\Phi(\mu^*) \leq \Phi(\mu) = \min_{K_R} J_\mu(u) \leq \min_{K_o(\Omega)} J_\mu(u) = \text{En}(\Omega). \tag{4.10}$$

For $\mu = \chi_{\tilde{\Omega}^*}$ there exists a $u_{\Omega^*} \in K_0(\Omega^*)$ such that $J_{\chi_{\tilde{\Omega}^*}}(u_{\Omega^*}) =$ En(Ω^*). Recall that $u^* \in K_0(\Omega^*)$, which implies that

$$\text{En } (\Omega^*) \leq J_{\chi_{\tilde{\Omega}^*}}(u^*). \tag{4.11}$$

As meas $\Omega^* \geq A$, we choose a set $E \subset \Omega^*$ with meas $E = A$ so that $E \in W$. It follows that

$$\Phi(\mu^*) \leq \text{En}(E) \leq \text{En}(\Omega^*). \tag{4.12}$$

From (4.11) and (4.12) we conclude that

$$\frac{1}{2} \int_{\Omega^*} |\nabla u^*|^2 \, d\omega - \epsilon \int_{\Omega^*} \mu^* u^* \, d\omega \leq \frac{1}{2} \int_{\Omega^*} |\nabla u^*|^2 \, d\omega - \epsilon \int_{\Omega^*} u^* \, d\omega.$$

This implies that

$$\int_{\Omega^*} \mu^* u^* \, d\omega \geq \int_{\Omega^*} u^* \, d\omega.$$

In our case this means that

$$\mu^* u^* \geq u^* \quad \text{almost everywhere on } \Omega^*,$$

but $\mu^* \leq 1$ almost everywhere on Ω^*. Moreover, the function u^* is continuous and nowhere zero in Ω^* as a result of regularity properties. We conclude that

$$\mu^* = \chi_{\Omega^*} \quad \text{almost everywhere in } D,$$

and consequently, meas $\Omega^* = A$.

It follows that $\Phi(\mu^*) = \text{En}(\Omega^*)$. This result and (4.10) prove the theorem.

Remark 4.4. We remark that, as a consequence of the fact that μ^* is a characteristic function, the sequence converges strongly in $L^2(D)$. Furthermore, we can choose a subsequence converging almost everywhere in D. Chenais [11] has shown that in this case, up to a null measure set we have $\Omega^* = \lim_n \inf \Omega_{R_n} = \cap_{l \in \mathbb{N}} \cup_{k \geq l} \Omega_{R_n k}$. It is known that in this case, as all the sets Ω_{R_n} are in a metric space, the set Ω^* is closed and connected. Consequently, if $D \backslash \overline{\Omega}^*$ is connected, it follows from Corollary 4.4 and the corresponding definitions that $\Omega^* = \Omega^+$, and this implies, for the free boundary $\Gamma^* = \partial\Omega^* \cap D$, meas $\Gamma^* = 0$. Up to now the regularity properties for Γ^* remain open, but in the case where it is regular enough so that the space $H_0^1(\Omega^*)$ is well defined, we are able to prove our main assertion.

THEOREM 4.6. *Let Γ_0 be a given closed Lipschitz continuous curve, nonintersecting itself so that the domain Ω_0 enclosed is star shaped. Let W be the set of all doubly connected domains Ω with a given measure A, Γ_0 as inner boundary, and a Lipschitz continuous outer boundary Γ. Then there exists a doubly connected domain Ω^* such that, for all $\Omega \in W$,*

$$\text{Cap}_{\Omega^*}(\Omega_0) \leq \text{Cap}_{\Omega}(\Omega_0).$$

Proof. Let us give a fixed, positive constant ϵ. For a given domain Ω with regular outer boundary Γ, the corresponding weak solution $u_\epsilon \in H^1(\Omega)$ of the boundary value problem

$$\begin{aligned} -\Delta u &= \epsilon \quad \text{in } \Omega \quad \text{in the weak sense} \\ u &= 0 \quad \text{on } \Gamma \quad \text{in the sense of traces} \\ u &= 1 \quad \text{on } \Omega_0 \quad \text{in the } H^1 \text{ sense} \end{aligned}$$

has the form $u_\epsilon = u_0 + u^\epsilon$, where u_0 is the capacity potential of the domain Ω_0 related to Ω, and u^ϵ is the corresponding solution of the Poisson equation in Ω with homogeneous Dirichlet conditions. Thus we have that

$$\text{En}_\epsilon(\Omega) = \frac{1}{2}\text{Cap}_\Omega(\Omega_0) + (\nabla u_0, \nabla u^\epsilon) + \frac{1}{2}\int_\Omega |\nabla u^\epsilon|^2 \, d\omega - \epsilon \int_\Omega u_\epsilon \, d\omega. \tag{4.13}$$

By applying Theorem 4.5 for a given ϵ, we know there exists a domain Ω^* such that, for every $\Omega \in W$,

$$\text{En}_\epsilon(\Omega^*) \leq \text{En}_\epsilon(\Omega). \tag{4.14}$$

Being u_0 harmonic in Ω, the second term of the right-hand side in (4.13) vanishes. Moreover, it is known that for fixed Ω, in the case where $\epsilon \to 0$, $u^\epsilon \to 0$ strongly in $H^1(\Omega)$. This implies that for every $\Omega \in W$,

$$\text{En}_\epsilon(\Omega) \to \tfrac{1}{2}\,\text{Cap}_\Omega(\Omega_0). \tag{4.15}$$

Applying property (4.15) to both sides of the inequality (4.14) gives the result.

Remark 4.4. We have that $\nabla u^* = 0$ in $D\backslash\overline{\Omega^*}$, which implies that, in the case where Γ^*, the boundary of Ω^* related to D, is smooth enough, it follows from formula (3.10) that

$$d\text{En}(\Omega^*, \theta) = -\frac{1}{2}\int_{\Gamma^*} |\partial u^*/\partial n|^2 \langle \theta, n\rangle \, ds.$$

Here $d\text{En}(\Omega^*, \theta)$ denotes the derivative of the energy functional related to domain deformation described by the vector field θ, (cf. Pironneau [25], Simon [26], Zolesio [30]).

It has been proved by other authors that from the necessary optimality condition $d\text{En}(\Omega^*, \theta) \geq 0$, which holds for every vector field θ preserving the measure, we obtain

$$|\partial u^*/\partial n| = |\nabla u^*| = \lambda \quad \text{on } \Gamma^*,$$

where λ is a positive constant that can be interpreted as a Lagrange multiplier for the functional $\Omega \to \int_\Omega |\nabla u_\Omega|^2 \, d\omega$ related to the measure constraint of the domain. Here u_Ω denotes the corresponding potential (cf., for example, Banichuk [4], Pironneau [25]).

Remark 4.5. It should be mentioned that Alt and Cafarelli [3] studied the following related problem: Find $v \in K$ that minimizes the functional

$$v \to J(v) = \int_\Omega |\nabla v|^2 \, d\omega + \int_\Omega Q^2 \chi_{v>0} \, d\omega,$$

where $K = \{v \in L^1_{\text{loc}}(\Omega) | \nabla v \in L^2(\Omega),\ v = u^0 \text{ on } S\}$; here $u^0 > 0, Q \geq 0$, and $S \subset \partial\Omega$ are given. For the case where Q and u^0 are constants, the solution of their problem solves ours for $A = \int_\Omega \chi_{u>0} \, d\omega$. Moreover, the stationary points of the functional J have the property

$$|\nabla u| = Q \quad \text{on } \Gamma = \Omega \cap \partial\{u > 0\}.$$

In their case, Q is given and the constant A is a result; in ours, A is given and the constant λ is a consequence of the necessary conditions of optimality (cf. another related result using different techniques in Acker [1]). This fact allows us to apply the regularity properties already known for the boundary Γ^* in our case, i.e., Γ^* is locally a $C^{1,\alpha}$ curve and even analytic.

Acknowledgments. The author thanks J. P. Zolesio and an anonymous reviewer for their advice and remarks, which were essential to the improvement of a first version of this paper. All possible errors remain the responsibility of the author.

REFERENCES

[1] A. Acker, *A free boundary optimization problem,* SIAM J. Math. Anal., 9 (1978), pp. 1179–1191.

[2] N. Aguilera, H. W. Alt, and L. A. Cafarelli, *An optimization problem with volume constraint,* SIAM J. Control Optim., 24 (1986), pp. 191–198.

[3] H. W. Alt and L. A. Cafarelli, *Existence and regularity for a minimum problem with free boundary,* J. Reine u. Angew. Math., 325 (1981), pp. 104–144.

[4] N. V. Banichuk, *Problems and Methods of Optimal Structural Design,* Plenum Press, London, 1983.

[5] G. Bottaro and M. E. Marina, *Problema di Dirichlet per equazioni ellitiche di tipo variazionale su insiemi non limitati,* Boll. Un. Mat. Ital., 8-A (1973), pp. 46–56.
[6] N. Bourbaki, *Espaces Vectoriels Topologiques,* Hermann, Paris, 1973.
[7] Ch. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions,* Lecture Notes in Math., 580, Springer-Verlag, Berlin, New York, 1977.
[8] J. Cea, *Optimisation, Théorie et Algorithmes,* Dunod, Paris, 1971.
[9] ———, *Problems of shape optimal design,* in Optimization of Distributed Parameter Structures, NATO Proceedings, E. J. Haug, and J. Cea, eds., Sijthoff and Noordhoff, Groningen, the Netherlands, 1981.
[10] J. Cea and K. Malanowski, *An example of a* max-min *problem in partial differential equation,* SIAM J. Control Optim., 8 (1970), pp. 305–316.
[11] D. Chenais, *On the existence of a solution in a domain identification problem,* J. Math. Anal. Appl., 52 (1975), pp. 189–219.
[12] I. Ekeland and R. Temam, *Analyse Convexe et Problémes Variationels,* Dunod, Paris, 1974.
[13] J. Frehse, *On the regularity of the solution of a second order variational inequality,* Boll. Un. Mat. Ital., 6-A (1972), pp. 312–315.
[14] Cl. Gebhardt, *Regularity of solutions of nonlinear variational inequalities,* Arch. Rat. Mech. Anal., 52 (1973), pp. 389–393.
[15] R. B. Gonzalez de Paz, *Relaxation methods for the study of domain optimization problems,* in Proc. of the 5th IFAC Symposium on Control of Distributed Parameter Systems, A. El Jai and M. Amouroux, eds., Pergamon Press, Oxford, UK, 1990.
[16] ———, *On the optimal design of elastic shafts,* Math. Modelling Numer. Anal. (M2AN), 23 (1989), pp. 615–625.
[17] ———, *Sur un probléme d'optimisation de domaine,* Numer. Funct. Anal. Optim., 5 (1982), pp. 173–197.
[18] R. Jensen, *Boundary regularity for variational inequalities,* Indiana Univ. Math. J., 29 (1980), pp. 495–511.
[19] O. Kellog, *Foundations of Potential Theory,* Dover, New York, 1954.
[20] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications,* Academic Press, New York, 1980.
[21] P. L. Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case, part* I, Ann. Inst. Henri Poincaré, 2 (1984), pp. 109–145.
[22] F. Murat and L. Tartar, *Calcul de variations et homogeneisation,* Cours Ecole d'Eté d'Analyse Numérique CEA-EDF-INRIA, Bréau sans Nappe, Juillet, 1983; Eyrolles, Paris, 1984.
[23] J. J. Moreau, *Fonctionelles Convexes,* Séminaire sur les Equations aux Derivées Partielles, Collége de France, 1966–1967.
[24] J. Neças, *Les Methodes Directes en Theorie des Equations Elliptiques,* Masson, Paris, 1967.
[25] O. Pironneau, *Optimal Shape Design for Elliptic Systems,* Springer-Verlag, Berlin, New York, 1984.
[26] J. Simon, *Differentiation with respect to the domain in boundary value problems,* Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.
[27] R. Temam, *Navier Stokes Equations,* North Holland, Amsterdam, 1979.
[28] M. Valadier, *Sous-differentiels d'une borne superieure et d'une somme continue de fonctions convexes,* C. R. Acad. Sci. Paris, Série A, 268 (1969), pp. 39–42.
[29] J. P. Zolesio, *Domain variational formulation for free boundary problems,* in Optimization of Distributed Parameter Structures, NATO Proceedings, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Amsterdam, the Netherlands, 1981.
[30] ———, *The material derivative (or speed) method for shape optimization,* in Optimization of Distributed Parameter Structures, NATO Proceedings, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Amsterdam, the Netherlands, 1981.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 170–175, January 1994

010

ON THE EXISTENCE OF OPTIMAL CONTROLS OF HILBERT SPACE-VALUED DIFFUSIONS*

DARIUSZ GĄTAREK† AND JAROSŁAW SOBCZYK‡

Abstract. An optimal control problem is studied for a Hilbert space-valued diffusion. Existence of an optimal control in the class of relaxed controls is proved. As a tool, the factorization method is used. Also, a simple example is given.

Key words. optimal control, diffusion processes in Hilbert spaces, existence theory, relaxed controls

1. Introduction. We consider the following control problem. Let X^u be a solution of the following stochastic equation on $H = L^2(\mathcal{O})$:

$$\frac{\partial}{\partial t}X^u(x,t) = AX^u(x,t) + f(X^u(x,t),u(x,t),x,t) + g(X^u(t),t)\dot{W}(x,t), X^u(x,0) = X_0(x), \tag{1}$$

where W is a Wiener process on H and A is a differential operator on H. Our objective is to minimize the following cost function:

$$J(u) = E\left\{\int_0^1 \int_{\mathcal{O}} b(X^u(x,t),u(x,t),x,t)dx\,dt + \int_{\mathcal{O}} h(X^u(x,1),x)dx\right\}.$$

We answer the question of existence of an optimal control for the above problem. We will use the compactification method to prove the existence by introducing a relaxed control q and a control policy that is a probability measure, P being the law of the pair (q, X^q). Such a method is classical in both deterministic and stochastic cases in finite-dimensional spaces; for instance, see [1], [2], [6], [7], [11]. In finite-dimensional case, the compactness of the set of control policies is proved by the use of the compactness criterion for the space of continuous functions proved by Stroock and Varadhan in [15] (see also [11]). In this paper we use the factorization method introduced by Da Prato, Kwapień, Zabczyk [4]. Due to this method we can essentially shorten and clarify proofs. This paper is the first, to our knowledge, to deal with an existence theorem of an optimal control for the equation with space-time noise in infinite-dimensional spaces.

The paper is organized as follows. In §2 we formulate the problem and give some preliminary results. In §3 we give an existence result, using the factorization method. We close the paper with an example in §4.

2. Preliminaries and setting of the problem. Let $\mathcal{O}$ be an open, bounded subset of $\mathbf{R}^n$ with smooth boundary. Define the Hilbert space $H = L^2(\mathcal{O})$. Denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the norm and the scalar product in the space H. Let $V \subseteq \mathbf{R}^m$ be a closed space of control parameters. Let U be the set of positive measures q on $V \times \overline{\mathcal{O}} \times [0,1]$, such that $q(V,dx,dt) = dxdt$ is the Lebesgue measure. We will use the disintegration of q from U given by $q(du,dx,dt) = q(du,dx,t)dt$ where $q(du,dx,t)$ is a measurable kernel of bounded mass. We call the measures $q \in U$ relaxed control parameters. Notice that U is a metrizable space in the stable topology v, defined as follows: $q_n \to q$ in v if and only if

$$\int_0^1 \int_{\mathcal{O}} \int_V \psi(u,x,t)q_n(du,dx,dt) \to \int_0^1 \int_{\mathcal{O}} \int_V \psi(u,x,t)q(du,dx,dt)$$

* Received by the editors February 13, 1992; accepted for publication (in revised form) June 18, 1992.
† Systems Research Institute, 01-447 Warsaw, Newelska 6, Poland.
‡ Department of Mathematics, Warsaw University of Technology, 00-661 Warsaw, Pl. Politechniki 1, Poland.

for any measurable bounded $\psi : V \times \overline{\mathcal{O}} \times [0,1] \to \mathbf{R}$, continuous with respect to u. Denote by $\mathcal{M}(B)$ the space of all probabilistic measures on the topological space B.

Let the following be given:

(i) Q, a nuclear operator on H;

(ii) $f \in C(\mathbf{R} \times V \times \overline{\mathcal{O}} \times [0,1]; \mathbf{R})$ such that $|f(y,u,x,t)| \leq K(1+|y|+|u|)$ for any $(y,u,x,t) \in \mathbf{R} \times V \times \overline{\mathcal{O}} \times [0,1]$;

(iii) $g \in C(H \times [0,1]; L(H;H))$ such that $\|g(X,t)\|_{L(H;H)} \leq K(1+\|X\|^{\beta})$ for any $(X,t) \in H \times [0,1]$ and a certain $\beta \in (0,1)$;

(iv) A, a differential operator on H, generating a compact semigroup $S(t)$ on H;

(v) X_0, a fixed element of H;

(vi) $b \in B(\mathbf{R} \times V \times \overline{\mathcal{O}} \times [0,1]; \mathbf{R}_+)$, such that $b(y,u,x,t)$ is lower-semicontinuous (l.s.c.) in (y,u) and $b(y,u,x,t) \geq \gamma |u|^2$ for certain $\gamma > 0$;

(vii) $h \in B(\mathbf{R} \times \mathcal{O}; \mathbf{R}_+)$, such that $h(x,y)$ is l.s.c. in x.

The cost function $J_0 : U \times C(0,1;H) \to \mathbf{R}$ takes the form:

$$J_0(q,X) = \int_0^1 \int_{\mathcal{O}} \int_V b(X(x,t),u,x,t) q(du,dx,dt) + \int_{\mathcal{O}} h(X(x,1),x)dx.$$

DEFINITION 1. We say that $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{X(t)\}, \{q_t\}, X_0)$ is a relaxed control if:

(i) $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ is a probability space equipped with a filtration;

(ii) $\{q_t\}$ is a U-valued, progressively measurable process;

(iii) $\{X(t)\}$ is a H-valued, progressively measurable process such that

$$C_t(\phi, X, q) = \phi(X(t),t) - L_1\phi(X,t) - L_2\phi(X,q,t)$$

is a continuous $(P, \mathcal{F}_t)$ martingale for any ϕ of the form

$$\phi(X(s),s) = \phi_0(\langle X(s), e_1\rangle, \ldots, \langle X(s), e_n\rangle, s),$$

where $\phi_0 \in C_0^{\infty}(\mathbf{R}^n \times (0,1)), e_i \in D(A^*)$ for any $i = 1,2,\ldots$ and $n = 1,2,3,\ldots$, and with the operators L_1 and L_2 of the form

$$L_1\phi(X,t) = \int_0^t \left(\langle A^* \nabla\phi(X(s),s), X(s)\rangle + \frac{\partial}{\partial s}\phi(X(s),s) + \frac{1}{2} \operatorname{Tr}[g^*(X(s),s)\nabla^2\phi(X(s),s)g(X(s),s)Q] \right) ds,$$

and

$$L_2\phi(X,q,t) = \int_0^1 \int_{\mathcal{O}} \int_V \nabla\phi(X(s),s)(x) \cdot f(X(x,s),u,x,s) q(du,dx,ds),$$

where $\nabla\phi$ and $\nabla^2\phi$ denote the first and second Frechet derivatives of the functional ϕ with respect to X. Moreover $X(0) = X_0 P$ almost surely:

(iv) The cost is $E^P(J_0(q,X))$. Define the probability space $\Omega = C(0,1;H) \times U$ with the filtration

$$\mathcal{F}_t = \sigma\left(X(s), \int_0^s \int_C \int_B q(du,dx,dr), \text{ where } B \subseteq V, C \subseteq \mathcal{O} \text{ are Borel sets and } s \leq t \right)$$

and $\mathcal{F}_t^X = \sigma(X(s), s \leq t)$. Elements of Ω will be denoted by $\omega = (X,q)$.

DEFINITION 2 (compare with [6], [11]). We say that a probabilistic measure P on Ω is an admissible control policy if and only if:

(i) $P(X(0) = X_0) = 1$;
(ii) the process $C_t(\phi, X, q)$ is a continuous $(P, \{\mathcal{F}_t\})$ martingale for any ϕ of the form as in Definition 1.
(iii) The cost is $J(P) = E^P(J_0(q, X))$.

We will work with a control policy in the sequel. The following lemma enables us to transfer a result proved in this paper concerning a control policy to a relaxed control case. Observe that if P is a control policy, then $(C(0,1;H) \times U, \mathcal{F}_1, P, \{\mathcal{F}_t\}, \{X(t)\}, \{q_t\})$ is a relaxed control.

LEMMA 1. *If $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{X(t)\}, \{q_t\}, X_0)$ is a relaxed control and $\{\mathcal{G}_t\}$ is a filtration such that $\mathcal{F}_t^X \subseteq \mathcal{G}_t \subseteq \mathcal{F}_t$, then there exists a process $\{\overline{q}_t\}$ such that $(\Omega, \mathcal{F}, P, \{\mathcal{G}_t\}, \{X(t)\}, \{\overline{q}_t\}, X_0)$ is a relaxed control and $J_0(\overline{q}, X) = J_0(q, X)$.*

Proof. We define $\overline{q}_t$ as a progressive version of $P(q_t|\mathcal{G}_t)$ such that for any $\gamma \in C_b(V \times \overline{\mathcal{O}})$,

$$\int_{\mathcal{O}} \int_V \gamma(u,x)\overline{q}(du,dx,t) = E^P\left(\int_{\mathcal{O}} \int_V \gamma(u,x)q(du,dx,t)|\mathcal{G}_t\right)$$

and $\overline{q}(du,dx,dt) = \overline{q}(du,dx,t)dt$.

The proof that such a progressive version exists is the same as in [6], [11]. Let ϕ be as in Definition 1 and let $t > r$. Denote $\psi(X(s),u,x,s) = \nabla\phi(X(s),s)(x) \cdot f(X(x,s),u,x,s)$. Since $\mathcal{F}_r \subseteq \mathcal{G}_r$ then

$$\begin{aligned}
&E^P(C_t(\phi,X,\overline{q}) - C_r(\phi,X,\overline{q})|\mathcal{G}_r) \\
&\quad = E^P\Bigg(\phi(X(t),t) - \phi(X(r),r) - L_1\phi(X,t) + L_1\phi(X,r) \\
&\qquad - \int_r^t \int_{\mathcal{O}} \int_V \psi(X(s),u,x,s)\overline{q}(du,dx,s)ds|\mathcal{G}_r\Bigg) \\
&\quad = E^P\Bigg(\phi(X(t),t) - \phi(X(r),r) - L_1\phi(X,t) + L_1\phi(X,r) \\
&\qquad - \int_r^t \left(E^P \int_{\mathcal{O}} \int_V \psi(X(s),u,x,s)\overline{q}(du,dx,s)|\mathcal{G}_s\right) ds|\mathcal{G}_r\Bigg) \\
&\quad = E^P\Bigg(\phi(X(t),t) - \phi(X(r),r) - L_1\phi(X,t) + L_1\phi(X,r) \\
&\qquad - \int_r^t \int_{\mathcal{O}} \int_V \psi(X(s),u,x,s)q(du,dx,s)ds|\mathcal{G}_r\Bigg) \\
&\quad = E^P\Bigg(E^P\Bigg(\phi(X(t),t) - \phi(X(r),r) - L_1\phi(X,t) + L_1\phi(X,r) \\
&\qquad - \int_r^t \int_{\mathcal{O}} \int_V \psi(X(s),u,x,s)q(du,dx,s)ds|\mathcal{F}_r\Bigg)|\mathcal{G}_r\Bigg) = 0,
\end{aligned}$$

because, by assumption, the internal conditional expectation is zero. Hence process $\{C_t(\phi, X, \overline{q})\}$ is $(P, \{\mathcal{G}_t\})$ martingale. The equality of the costs is obvious. □

Denote by $\mathcal{P}$ the space of all admissible control policies. The cost function associated to a measure $\mathcal{P} \in \mathcal{P}$ is given by $J(P) = E^P J_0$. Notice that $J : \mathcal{P} \to \mathbf{R}_+$ is l.s.c. with respect to the weak topology.

PROPOSITION 1 [10]. *For any admissible control policy P there exists a Wiener process W on the space H with covariance operator Q, such that*

$$X(t) = S(t)X_0 + \int_0^t S(t-s)f(X,q,s)ds + \int_0^t S(t-s)g(X(s),s)dW(s), \tag{2}$$

where the function f is extended to $f : C(0,1;H) \times U \times [0,1] \to H$ by the following: For any Borel subset B of $\overline{\mathcal{O}} \times [0,1]$,

$$\int_B f(X,q,t)(x)dxdt = \int_B \int_V f(X(x,t),u,x,t)q(du,dx,dt).$$

Here the state equation (2) takes the symbolic form:

$$dX(t) = [AX(t) + f(X,q,t)]dt + g(X(t),t)dW(t). \tag{3}$$

By Proposition 1 the set $\mathcal{P}$ is nonempty.

3. Solution of the problem. Define

$$\mathcal{P}(M) = \{P \in \mathcal{P} : J(P) \leq M\} \text{ for any } M \in \mathbf{R}_+.$$

THEOREM 1. *The set $\mathcal{P}(M)$ is compact for any $M \in \boldsymbol{R}_+$.*

The proof is based on the following facts.

We now recall the factorization method as presented in [4]. For any $h \in L^p(0,1;H)$ and $\alpha \in (0,1]$ define an operator R_α by

$$R_\alpha h(t) = \int_0^t (t-s)^{\alpha-1} S(t-s)h(s)ds.$$

LEMMA 2 ([4, Prop. 1]). *Let W be a Wiener process with covariance operator Q. Then for any $0 < p^{-1} < \alpha < \frac{1}{2}$ and any predictable $\phi \in L^p(\Omega \times [0,1]; L(H;H))$:*

$$\pi \sin(\pi\alpha)^{-1} \int_0^t S(t-s)\phi(s)dW(s) = R_\alpha Y(t),$$

where

$$Y(t) = \int_0^t (t-s)^{-\alpha} S(t-s)\phi(s)dW(s).$$

LEMMA 3 ([4, Prop. 1]). *The operator $R_\alpha : L^p(0,1;H) \to C(0,1;H)$ is compact for any $\alpha \in (p^{-1}, 1]$.*

Proof of the theorem. The proof follows [6]. Let $P \in \mathcal{P}(M)$. By Proposition 1, there exists a Wiener process W on the space H with covariance operator Q, such that X satisfies equation (2). By assumption (vi),

$$E^P \int_0^1 \int_{\mathcal{O}} \int_V |u|^2 q(du,dx,dt) \leq \gamma^{-1} M. \tag{4}$$

By the Ito formula, (2) and (4),

$$E^P \|X(t)\|^2 \leq E^P \|X_0\|^2 + K \int_0^t E^P \|X(s)\|^2 ds + K\gamma^{-1} M.$$

By the Gronwall lemma,

$$\sup_{P\in\mathcal{P}(M)}\sup_{0\le t\le 1} E^P\|X(t)\|^2 \le C < +\infty. \tag{5}$$

Let $F(t) = f(X,q,t)$ and

$$Y(t) = \int_0^t (t-s)^{-\alpha} S(t-s)g(X(s),s)dW(s).$$

Fix $p = 2\beta^{-1}$ and $p^{-1} < \alpha < \frac{1}{2}$. By the Young inequality,

$$\begin{aligned} E^P\int_0^1 \|Y(t)\|^p dt &= E^p \int_0^1 \left\| \int_0^t (t-s)^{-\alpha} S(t-s)g(X(s),s)dW(s)\right\|^p dt \\ &\le M_0 \int_0^1 E^P \left(\int_0^t (t-s)^{-2\alpha}\|g(X(s),s)\|^2 ds\right)^{\beta^{-1}} ds \\ &\le M_1 1 + M_1 E^p \int_0^1 \|X(s)\|^2 ds \le C. \end{aligned}$$

Therefore,

$$\sup_{P\in\mathcal{P}(M)} \int_0^1 E^P\|Y(t)\|^p dt \le C < +\infty. \tag{6}$$

Denote, for any $0 < r^{-1} < \delta \le 1$,

$$\Lambda(R,\delta,r) = \left\{ w \in C(0,1;H) : w = R_\delta u, u \in L^r(0,1;H), \int_0^1 \|u(s)\|^r ds \le R \right\}$$

and

$$\begin{aligned} &\Xi(R) \\ &= \{w \in C(0,1;H) : w(t) = S(t)X_0 + w_1(t) + w_2(t), w_1 \in \Lambda(R,1,2), w_2 \in \Lambda(R,\alpha,p)\}. \end{aligned}$$

By Proposition 1 the sets $\Lambda(R,\delta,r)$ and $\Xi(R)$ are relatively compact in $C(0,1;H)$. Recall that $X(t) = S(t)X_0 + \pi^{-1}\sin(\pi\alpha)R_\alpha Y(t) + R_1 F(t)$. By (5), (6), and the Chebyshev inequality, for any $\varepsilon > 0$, $P(X \in \Xi(R)) \ge 1-\varepsilon$ for sufficiently large R for any $P \in \mathcal{P}(M)$. By the Prokhorov theorem the projection of $\mathcal{P}(M)$ on $\mathcal{M}(C(0,1;H))$ is tight.

By [11], the set

$$\left\{ q \in U : \int_0^1 \int_{\mathcal{O}} \int_V |u|^2 q(du,dx,dt) \le K \right\}$$

is relatively compact in the stable topology. Hence the projection of $\mathcal{P}(M)$ on $\mathcal{M}(U)$ is tight. We will show that $\mathcal{P}(M)$ is closed. Let the functional ϕ be as in Definition 1.

Let $C_t(\phi,X,q) = \phi(X(t),t) - L_1\phi(X,t) - L_2\phi(X,q,t)$. Notice that C_t is a continuous function on $C(0,1;H)\times U$. Let $P^n \to P$ weakly and $P^n \in \mathcal{P}(M)$ for any $n \ge 0$. Let $\psi \in C_b(\Omega;\mathbf{R})$ be an $\mathcal{F}_s$-measurable function. Compute that

$$0 = E^{P^n}\psi\{C_t(\phi,X,q) - C_s(\phi,X,q)\} \to E^P\psi\{C_t(\phi,X,q) - C_s(\phi,X,q)\}.$$

Hence $E^P\psi\{C_t(\phi,X,q) - C_s(\phi,X,q)\} = 0$ for any ψ $\mathcal{F}_s$ measurable and therefore $C_t(\phi,X,q)$ is P martingale. Since the functional J is l.s.c. then $J(P) \leq M$. Hence $\mathcal{P}(M)$ is compact. □

COROLLARY. *If* $\inf_{P\in\mathcal{P}} J(P) < +\infty$ *then there exists an optimal policy* P^*,

$$J(P^*) = \inf_{P\in\mathcal{P}} J(P).$$

Proof. It suffices to notice that J is l.s.c. and $\mathcal{P}(M)$ is nonempty and compact for a sufficiently large M. □

4. Example. Let $\mathcal{O} \subseteq \mathbf{R}$ be a bounded area with smooth boundary. Let $H = L(\mathcal{O}), V = [0,\infty)$ and $A = \Delta$ be the Laplace operator on $\mathcal{O}$. Let $f : \mathbf{R}\times[0,\infty) \to \mathbf{R}$ be a function satisfying assumption (i) and $g : \mathbf{R} \to \mathbf{R}$ be a bounded continuous function.

We consider the following control problem:

$$\frac{\partial}{\partial t}X^u(x,t) = \Delta X^u(x,t) + f(X^u(x,t),u(x,t)) + g(X^u(x,t))\dot{W}(t)$$

with

$$X^u(x,0) = \xi(x) \quad \text{for } x \in \mathcal{O},$$

$$X^u(x,t) = 0 \quad \text{for } x \in \partial\mathcal{O} \quad \text{and} \quad t \geq 0,$$

where $\xi \in H$ and W is a one-dimensional Brownian motion. This is a reaction-diffusion equation with randomly disturbed source. We minimize the following functional:

$$J(u) = E\int_0^1\int_{\mathcal{O}} |X^u(x,t) - \zeta(x,t)|^2 + C|u(x,t)|^2\}dxdt + E\int_{\mathcal{O}} |X^u(x,1) - \eta(x)|^2dx,$$

where $\zeta \in L^2(\mathcal{O}\times[0,1]), \eta \in L^2(\mathcal{O})$ and C is a positive number. By our theorem we have the existence of an optimal control for this problem.

REFERENCES

[1] H. Becker and V. Mandrekar, *On the existence of optimal random controls,* J. Math. Mech., 12 (1969), pp. 1151–1166.

[2] V. F. Beneš, *Existence of optimal stochastic controls laws,* SIAM J. Control, 9 (1971), pp. 446–472.

[3] J. N. Curtland, *Internal controls and relaxed controls,* J. London Math. Soc., 27 (1983), pp. 130–140.

[4] G. Da Prato, S. Kwapień, and J. Zabczyk, *Regularity of solutions of linear stochastic equations in Hilbert spaces,* Stochastic, 23 (1987), pp. 1–23.

[5] M. H. A. Davis, *On the existence of optimal policies in stochastic control,* SIAM J. Control, 11 (1973), pp. 587–594.

[6] N. El Karoui, Du'Huu Nguyen, and M. Jeanblanc-Picque, *Compactification methods in the control of degenerate diffusion: Existence of an optimal control,* Stochastics, 23 (1987), pp. 169–219.

[7] W. H. Fleming and N. Nisio, *On the existence of optimal stochastic controls,* J. Math. Mech., 15 (1966), pp. 777–794.

[8] W. H. Fleming and N. Nisio, *On stochastic relaxed control for partially observed diffusions,* Nagoya Math. J., 93 (1984), pp. 71–108.

[9] D. Gątarek and B. Goldys, *On solving stochastic evolution equations by change of drift with application to optimal control,* Appl. Math. Optim., submitted.

[10] ———, *On weak solutions of stochastic equations in Hilbert spaces,* Stochastics, to appear.

[11] U. G. Haussmann and J. P. Lepeltier, *On the existence of optimal controls,* SIAM J. Control Optim., 28 (1990), pp. 851–902.

[12] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes,* North-Holland, Amsterdam, 1981.

[13] N. V. Krylov, *Controlled Diffusion Processes,* Springer-Verlag, Berlin, New York, 1980.

[14] H. Kushner, *Existence result for optimal stochastic controls,* J. Optim. Theory Appl., 15 (1975), pp. 347–360.

[15] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes,* Springer-Verlag, Berlin, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 176–186, January 1994

011

ERGODIC CONTROL OF MARKOV CHAINS WITH CONSTRAINTS—THE GENERAL CASE*

VIVEK S. BORKAR†

Abstract. The problem of controlling a Markov chain on a countable state space with ergodic or 'long run average' cost is studied in the presence of additional constraints, requiring finitely many (say, m) other ergodic costs to satisfy prescribed bounds. Under extremely general conditions, it is proved that an optimal stationary randomized strategy can be found that requires at most m randomizations. This generalizes a result of Ross.

Key words. controlled Markov chains, control under constraints, ergodic control, randomized strategy, ergodic occupation measures

AMS subject classifications. 93E20, 90C40

1. Introduction.

The study of controlled Markov chains with constraints goes back to [10]. There has been a recent upsurge of interest due to possible applications to control of queuing networks ([2], [3], [5], [12], [13], [18], [19], among others). See also [17] for a related result. An important result in this domain is due to Ross [18] who proved that for the ergodic constrained problem with finite state and action spaces, an optimal stationary randomized strategy can be found which requires at most as many randomizations as the number of constraints. In [7 (Chap. 7)] and [8], the author extended this result to countable state space and compact action space for the single constraint case. It was claimed in [7 (Chap. 7)] and [8] that the multiple constraint problem can be handled simply by iterating the argument for a single constraint. This happens to be incorrect. The present work gives a direct proof for the multiple constraint case which is simpler and more elegant than that of [7], [8] even for the single constraint case. Moreover, it does not require the "single communicating class" condition used in [7], [8].

Our approach is to treat the control problem as a constrained optimization problem on a suitably defined closed convex set G of "ergodic occupation measures." This can also be viewed as an abstract (infinite-dimensional) linear programming problem, in which case it becomes a special instance of the so-called "moment problem" of LP (see [1, p. 85], [2], [4], [14], [15], among others). These works share in common with the present work the use of some variant of Dubins' lemma (Lemma 3.1) which immediately tells us that the optimal ergodic occupation measure is a convex combination of at most $m + 1$ extreme points of G, m being the number of constraints. The identification of these extreme points with stationary nonrandomized strategies, and the fact that their convex combination (which a priori means a mixed strategy that randomizes between them) is equivalent (in the sense of yielding the same costs) to a single strategy that requires at most m randomizations, are the main contributions of this paper. Both these results are specific to controlled Markov chains and do not follow from the general theory of the aforementioned moment problem.

The remainder of this section sets up the notation and formulates the problem. The next section contains some preliminary results concerning "ergodic occupation measures." This has some overlap with [7], [8]. The full details are included here, not just to make this account self-contained, but also because they are crucially required in the proofs of the main results. The latter are proved in §3. Section 4 concludes with some relevant remarks.

Let $X_n, n \geq 0$, be a controlled Markov chain on a countable state space $S = \{1, 2, \ldots\}$, with transition matrix $P_u = [[p(i,j,u_i)]], i, j \in S$, indexed by the control vector $u = [u_1, u_2, \ldots]$. Here $u_i \in D(i)$ for prescribed compact metric spaces $D(i), i \in S$.

* Received by the editors September 9, 1991; accepted for publication (in revised form) June 30, 1992.
† Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India.

By replacing $D(i)$ by $\Pi_j D(j)$ and each $p(i,j,.)$ by its composition with the projection $\Pi_j D(j) \to D(i)$, we may suppose that $D(i)$'s are replicas of a fixed compact metric space D. Let $L = D^\infty$. The maps $p(i,j,.)$ are assumed to be continuous.

For any Polish space Y, let $P(Y)$ = the space of probability measures with Prohorov topology [6]. If Y is countable, say $\{1,2,\ldots\}$, write $\mu \in P(Y)$ as a row vector $[\mu(\{1\}), \mu(\{2\}), \ldots]$, or simply $[\mu(1), \mu(2), \ldots]$. Let $P_0(L) \subset P(L)$ be the compact set consisting of product measures on L.

A control strategy (CS for short) is a sequence $\{\xi_n\}, \xi_n = [\xi_n(1), \xi_n(2), \ldots]$ of L-valued random variables such that for $i \in S, n \geq 0$,

$$P(X_{n+1} = i / X_m, \xi_m, m \leq n) = p(X_n, i, \xi_n(X_n)). \tag{1.1}$$

We say that the chain $\{X_n\}$ is governed by the CS $\{\xi_n\}$. If $\{\xi_n\}$ are independently and identically distributed with a common law $\Phi \in P(L)$, call it a stationary randomized strategy, SRS for short. As argued in [7, p. 21], we may assume that $\Phi \in P_0(L)$ and write $\Phi = \Pi_i \hat{\Phi}_i$ with $\hat{\Phi}_i \in P(D)$. Denote this SRS by $\gamma[\Phi]$. Under $\gamma[\Phi]$, $\{X_n\}$ is a Markov chain with a stationary transition matrix

$$P[\Phi] = [[p_\Phi(i,j)]], \qquad p_\Phi(i,j) = \int p(i,j,y)\hat{\Phi}_i(dy).$$

If Φ is a Dirac measure at $\xi \in L$, call $\gamma[\Phi]$ a stationary strategy (SS for short), denoted by $\gamma\{\xi\}$. The corresponding transition matrix will be denoted by $P\{\xi\} = P_\xi$.

If $\pi \in P(S)$ is an invariant probability measure under an SRS $\gamma[\Phi]$, we associate with the pair $(\pi, \gamma[\Phi])$ an "ergodic occupation measure" $\nu \in P(S \times D)$ defined by

$$\nu(\{i\}, du) = \pi(i)\hat{\Phi}_i(du), \qquad i \in S. \tag{1.2}$$

The set of all ergodic occupation measures will be denoted by G. The ergodic or long run average cost control problem is to almost surely minimize

$$\lim_{n\to\infty} \sup \frac{1}{n} \sum_{m=0}^{n-1} k_0(X_m, \xi_m(X_m)) \tag{1.3}$$

for a prescribed $k_0 \in C(S \times D; R^+)$. If an SRS $\gamma[\Phi]$ is being used and the initial state is in the support of a probability measure π which is ergodic under $\gamma[\Phi]$ (i.e., is an extreme point of the simplex of invariant probability measures under $\gamma[\Phi]$), then (1.3) almost surely equals

$$\int k_0 d\nu \tag{1.4}$$

for ν defined as in (1.2). We will consider the following constrained ergodic control problem: Given $k_i \in C(S \times D; R^+), 0 \leq \beta_i < \alpha_i \in R, 1 \leq i \leq m$, minimize $\int k_0 d\nu$ subject to

$$\beta_i \leq \int k_1 d\nu \leq \alpha_i, \qquad 1 \leq i \leq m, \tag{1.5}$$

where $\nu \in G$ (i.e., over H = the subset of G satisfying (1.5) which we assume to be nonempty).

LEMMA 1.1. *G is closed.*

Proof. Let $\nu_n \to \nu$ in $P(S \times D)$. From the definition of ergodic occupation measures, we have

$$\int p(.,j,.)d\nu_n = \nu_n(\{j\} \times D), \quad j \in S, \quad n \geq 1.$$

Letting $n \to \infty$, we get the same under ν. (Recall that $p(.,j,.)$ are continuous and $\{j\} \times D$ is both open and closed in $S \times D$). Disintegrating ν as in (1.2), we have $\nu \in P(S \times D) \Longrightarrow \pi \in P(S)$ and $\pi P[\Phi] = \pi$, i.e., $\pi \in P(S)$ is invariant under $\gamma[\Phi]$. Thus $\nu \in G$. □

We will consider two cases.

Case 1 (the stable case). Here we assume that G is compact and H is closed (and hence compact). Several sufficient conditions for compactness of G can be given; see, e.g., §V.3 of [7]. Since the upper inequality in (1.5) is preserved in any case under sequential limits in G, the closedness hypothesis of H requires that the lower inequality also does so. This would be the case, e.g., if $\beta_i = 0$ for all i or if $k_1, \ldots, k_m$ are bounded.

Case 2 (the near-monotone case). Here we assume that $\beta_i = 0$ for all i and the following "near-monotonicity" condition holds:

$$\liminf_{j \to \infty} \inf_u k_i(j,u) > \alpha_i, \quad 0 \leq i \leq m, \tag{1.6}$$

where

$$\alpha_0 = \inf_{\nu \in H} \int k_0 d\nu.$$

Equation (1.6) is satisfied by $\{k_i\}$ of the form $k_i(j,u) = f_i(j)$ with $f_i(j)$ increasing in j; hence the word "near-monotone."

We conclude this section with a statement of Choquet's theorem which will play a crucial role in what follows. Let E be a Hausdorff locally convex topological vector space and $X \subset E$ a convex compact metrizable subset thereof. Given a probability measure μ on X, call x its barycenter (or "resultant") if $f(x) = \int f d\mu$ for all continuous affine $f : X \to R$. Choquet's theorem states that each $x \in X$ is the barycenter of probability measure supported on the set of extreme points of X ([9, pp. 140–141]). Metrizability of X ensures that the latter set is measurable (in fact, G_δ—see [9, p. 138]), whereas compactness of X ensures that it is nonempty ([9, p. 105]).

2. Preliminary results. We start with some properties of sets G, H.

LEMMA 2.1. *G, H are convex.*

Proof. Since (1.5) is preserved under convex combinations, it suffices to prove the claim for G. Let $\nu_k \in G$, with

$$\nu_k(\{i\}, du) = \pi_k(i)\hat{\Phi}_{ki}(du), \qquad 1 \leq k \leq n.$$

Let $a_i \in (0,1)$ with $\sum_{i=1}^n a_i = 1$ and $\nu = \sum_{k=1}^n a_k \nu_k$. Set $\pi = \sum_{k=1}^n a_k \pi_k$ and define $\Phi = \Pi_i \hat{\Phi}_i \in P_0(L)$ by

$$\hat{\Phi}_i = \sum_{k=1}^n \left[a_k \pi_k(i) \hat{\Phi}_{ki} / \left(\sum_{j=1}^n a_j \pi_j(i) \right) \right] \tag{2.1}$$

for $i \in$ support (π), arbitrary otherwise. For each k,

$$\int p(.,j,.)d\nu_k = \nu_k(\{j\} \times D), \qquad j \in S. \tag{2.2}$$

Multiply (2.2) by a_k on both sides and sum over k. Rearranging terms, we obtain

$$\int p(.\,, j, .)d\nu = \nu(\{j\} \times D), \qquad j \in S,$$

for ν as in (1.2). Thus π is invariant under $\gamma[\Phi]$ and hence $\nu \in G$. □

The next lemma is a technical fact needed later to characterize extreme points of G.

LEMMA 2.2. *Let $\nu \in G$ be as in* (1.2). *Suppose that for some $k \in$ support (π)(say, $k = 1$), $\hat{\Phi}_k = \hat{\Phi}_1 = a\varphi_1 + (1-a)\varphi_2$ for some $a \in (0,1)$ and $\varphi_1 \neq \varphi_2$ in $P(D)$. Define $\Phi', \Phi'' \in P_0(L)$ by*

$$\Phi' = \varphi_1 \times \prod_{i=2}^{\infty} \hat{\Phi}_i, \qquad \Phi'' = \varphi_2 \times \prod_{i=2}^{\infty} \hat{\Phi}_i. \tag{2.3}$$

Then $\gamma[\Phi'], \gamma[\Phi'']$ both admit invariant probability measures containing $k\,(=1)$ in their supports.

Proof. Changing $\gamma[\Phi]$ to $\gamma[\Phi']$ or $\gamma[\Phi'']$ affects only the probabilities of transitions leaving k. Letting T, T_1, T_2 denote the mean return times to k under $\gamma[\Phi], \gamma[\Phi'], \gamma[\Phi'']$, respectively, it is clear that $T = aT_1 + (1-a)T_2$. Since $T < \infty$ and $1 > a > 0, T_1, T_2 < \infty$, and the claim follows. □

LEMMA 2.3. *The extreme points of G correspond to* SS.

Proof. Let ν as in (1.2) be an extreme point of G. For $i \notin$ support (π), we may set $\hat{\Phi}_i =$ some Dirac measure without affecting ν. Let $i \in$ support (π) (say, $i = 1$). Suppose

$$\hat{\Phi}_1 = a\varphi_1 + (1-a)\varphi_2 \tag{2.4}$$

for some $a \in (0,1)$ and $\varphi_1 \neq \varphi_2$ in $P(D)$. Define Φ', Φ'' by (2.3) and let π_1, π_2 be ergodic probability measures under $\gamma[\Phi'], \gamma[\Phi'']$, respectively containing 1 in their supports. Pick $b \in (0,1)$ such that

$$a = b\pi_1(1)/(b\pi_1(1) + (1-b)\pi_2(1)).$$

This is possible because $\pi_1(1), \pi_2(1) > 0$. Let $\pi' = b\pi_1 + (1-b)\pi_2$ and define $\nu' \in P(S \times D)$ by $\nu'(\{i\}, du) = \pi'(i)\hat{\Phi}_i(du)$. A computation similar to that of Lemma 2.1 shows that

$$\int p(.\,, j, .)d\nu' = \nu'(\{j\} \times D), \qquad j \in S.$$

Thus π' is invariant under $\gamma[\Phi]$. Also, support (π') = support $(\pi_1) \cup$ support (π_2). Now

$$p_\Phi(i,j) = ap_{\Phi'}(i,j) + (1-a)p_{\Phi''}(i,j), \qquad i,j \in S. \tag{2.5}$$

Since π_1 (respectively, π_2) is an ergodic probability measure under $\gamma[\Phi']$ (respectively, $\gamma[\Phi'']$), any two states in its support communicate under $\gamma[\Phi']$ (respectively, $\gamma[\Phi'']$) and therefore under $\gamma[\Phi]$ in view of (2.5) and the fact that $1 > a > 0$. Since supports of π_1, π_2 have a nonempty intersection containing "1," support (π') is a single communicating class under $\gamma[\Phi]$. Thus π' is an ergodic probability measure under $\gamma[\Phi]$. If π is also ergodic, we must have $\pi = \pi'$ and $\nu = \nu'$. Then it is easily checked (as in the proof of Lemma 2.1) that

$$\nu' = b\nu_1 + (1-b)\nu_2$$

where

$$\nu_1(\{i\}, du) = \pi_1(i)\hat{\Phi}'_i(du), \qquad \nu_2(\{i\}, du) = \pi_2(i)\hat{\Phi}''_i(du).$$

Since ν_1, ν_2 are clearly distinct and $1 > b > 0, \nu$ cannot be an extreme point of G, a contradiction. Suppose π is not ergodic. Then $\pi = c\overline{\pi} + (1-c)\tilde{\pi}$ for some $1 > c > 0$ and $\overline{\pi}, \tilde{\pi} \in P(S)$, which are distinct invariant probability measures under $\gamma[\Phi]$. Then

$$\nu = c\overline{\nu} + (1-c)\tilde{\nu}$$

where

$$\overline{\nu}(\{i\}, du) = \overline{\pi}(i)\hat{\Phi}_i(du), \qquad \tilde{\nu}(\{i\}, du) = \tilde{\pi}(i)\hat{\Phi}_i(du).$$

Since $\overline{\nu}, \tilde{\nu}$ are distinct and $1 > c > 0$, once again ν cannot be an extreme point of G. Thus (2.4) must not be possible. That is, for all $i \in$ support (π), $\hat{\Phi}_i$ must be Dirac. This completes the proof. □

Remark. We proved in passing that if ν is an extreme point of G and ν is as in (1.2), π must be an ergodic probability measure under $\gamma[\Phi]$.

We characterize the extreme points of H in the next section. The remainder of this section is devoted to proving that (1.4) does attain its minimum at an extreme point of H. Let $\overline{S} = S \cup \{\infty\}$ denote the one point compactification of S. We view $S, P(S), P(S \times D)$ as subsets of $\overline{S}, P(\overline{S}), P(\overline{S} \times \overline{D})$, respectively, via the natural embedding. Let $\overline{G}, \overline{H}$ be closures of G, H in $P(\overline{S} \times D)$. (For Case 1, $G = \overline{G}, H = \overline{H}$). Let $H_e, \overline{H}_e, G_e, \overline{G}_e$ denote the sets of their extreme points, respectively.

LEMMA 2.4. $H_e \subset \overline{H}_e, G_e \subset \overline{G}_e$.

Proof. Any $\nu \in H_e \backslash \overline{H}_e$ must be a convex combination of two distinct elements of $\overline{H}$ at least one of which must assign strictly positive probability to $\{\infty\} \times D$. But then so would ν, a contradiction. Thus $H_e \subset \overline{H}_e$. A similar argument proves the second claim. □

LEMMA 2.5. *Any $\nu \in H$ is the barycenter of a probability measure on H_e.*

Proof. In Case 1, H is compact and hence H_e nonempty ([9, p. 105]). The claim follows from Choquet's theorem. In Case 2, Choquet's theorem implies that ν is the barycenter of a probability measure Φ on $\overline{H}_e$ which is nonempty. If $\Phi(\overline{H}_e / H_e) > 0$, we must have $\nu(\{\infty\} \times D) > 0$, a contradiction. Thus $\Phi(H_e) = 1$ and we are done. (It follows, incidentally, that H_e is nonempty.) □

LEMMA 2.6. *Each $\nu \in \overline{H}$ is of the form*

$$\nu(A) = \delta\nu'(A \cap (S \times D)) + (1-\delta)\nu''(A \cap (\{\infty\} \times D)) \tag{2.6}$$

for all A Borel in $\overline{S} \times D$ with $\delta \in [0,1], \nu' \in G$ and $\nu'' \in P(\{\infty\} \times D)$.

Proof. Equation (2.6) obviously holds for some $\nu' \in P(S \times D)$. The claim is trivial for $\delta = 0$ and for $\delta = 1, \nu = \nu' \in H \subset G$ (e.g., in Case 1). Let $\delta \in (0,1]$. Pick $\nu_n \in H, n \geq 1$, such that

$$\nu_n(\{i\}, du) = \pi_n(i)\hat{\Phi}_{ni}(du), \qquad i \in S,$$

and $\nu_n \to \nu$ in $\overline{H}$. For $j \in S$,

$$\int p(.\,, j, .)d\nu_n = \nu_n(\{j\} \times D), \qquad n \geq 1.$$

Since $\{j\} \times D$ is both open and closed in $\overline{S} \times D$,

$$\nu_n(\{j\} \times D) \to \nu(\{j\} \times D) = \delta\nu'(\{j\} \times D).$$

Also, letting $S(N) = \{1, \ldots, N\} \subset S$,

$$\begin{aligned}
\liminf_{n\to\infty} \int p(.,j,.)d\nu_n &\geq \lim_{n\to\infty} \int_{S(N)\times D} p(.,j,.)d\nu_n \\
&= \int_{S(N)\times D} p(.,j,.)d\nu \\
&= \delta \int_{S(N)\times D} p(.,j,.)d\nu' \\
&\to \delta \int p(.,j,.)d\nu'
\end{aligned}$$

as $N \to \infty$. Combining the two,

$$\int p(.,j,.)d\nu' \leq \nu'(\{j\} \times D).$$

Both sides add up to one when summed over j. Thus equality must hold for all j, implying $\nu' \in G$ as in the proof of Lemma 1.1. □

LEMMA 2.7. *Equation* (1.4) *attains its minimum in* H.

Proof. Let $\{\nu_n\} \in H$ be such that

$$\int k_0 d\nu_n \downarrow \alpha_0.$$

In Case 1, H is compact and therefore $\nu_n \to \nu$ along a subsequence (denoted $\{n\}$ again by abuse of terminology) for some $\nu \in H$. Thus

$$\alpha_0 = \liminf_{n\to\infty} \int k_0 d\nu_n \geq \int k_0 d\nu \geq \alpha_0,$$

i.e., (1.4) attains its minimum at ν. In Case 2, let $\nu_n \to \nu \in \overline{H}$ by dropping to a subsequence if necessary. Pick $\epsilon > 0, k \geq 1$ such that

$$\inf_u k_j(i,u) \geq \alpha_j + \epsilon \quad \text{for } i \geq k, \quad 0 \leq j \leq m.$$

For $n \geq 1$, let

$$k_{jn}(i,u) = k_j(i,u)I\{i \leq k+n\} + (\alpha_j + \epsilon)I\{i > k+n\}.$$

Then for each j,

$$\begin{aligned}
\alpha_j &\geq \liminf_{n\to\infty} \int k_j d\nu_n \\
&\geq \lim_{n\to\infty} \int k_{jl} d\nu_n \\
&= \delta \int k_{jl} d\nu' + (1-\delta)(\alpha_j + \epsilon).
\end{aligned}$$

Letting $l \to \infty$ on the right,

$$\alpha_j \geq \delta \int k_j d\nu' + (1-\delta)(\alpha_j + \epsilon),$$

which is possible only if $\delta > 0$ and $\int k_j d\nu' \leq \alpha_j$ for all j. Thus $\nu' \in H$ and $\int k_0 d\nu' \leq \alpha_0$, implying $\int k_0 d\nu' = \alpha_0$. That is, (1.4) attains its minimum at ν'. □

THEOREM 2.1. *Equation* (1.4) *attains its minimum on* H *at an extreme point of* H.

Proof. Let the minimum of (1.4) on H be attained at ν and let Φ be as in Lemma 2.5. Then

$$\alpha_0 = \int k_0 d\nu = \int_{H_e} \left(\int k_0 d\rho \right) \Phi(d\rho).$$

Since $\int k_0 d\rho \geq \alpha_0$ for $\rho \in H_e$, $\int k_0 d\rho = \alpha_0$ for Φ-almost surely ρ. □

3. Main results. Let $\nu_0 \in H_e$ be such that $\int k_0 d\nu_0 = \alpha_0$. In this section we prove that ν_0 corresponds to an SRS with at most m randomizations. View H, G (respectively $\overline{H}, \overline{G}$) as subsets of the topological vector space of finite signed measures on $S \times D$ (respectively, $\overline{S} \times D$). A key lemma required will be the following specialization of a result of Dubins [11]. (See also [20, p. 265]).

LEMMA 3.1. ν_0 *can be expressed as a strict convex combination (i.e., convex combination with nonzero weights) of* k *points in* G_e *for some* $k \leq m+1$.

Proof. Consider Case 1. Suppose the claim is false. For simplicity, suppose that ν_0 can be expressed as a convex combination of $k = m+2$ distinct points in G_e, but not less. (For higher k, a similar proof works.) Then ν_0 must lie in the interior of an $(m+2)$-simplex A formed by these points in G_e. Let M be the $(m+1)$-dimensional affine space (i.e., translate a linear subspace) generated by A and let B be an open ball in M centered at ν_0 and contained in the interior of A. Thus $B \subset A \subset G$. Consider the intersections of the constraint hyperplanes $\{\nu | \int k_j d\nu = \alpha_j \text{ (or } \beta_j)\}$, $1 \leq j \leq m$, with M. Since at most m distinct constraint hyperplanes can intersect each other at a time, the intersections of their intersections with M must have a codimension of at most m in M and thus cannot have a corner in the interior of B. Thus ν_0 cannot be in H_e, a contradiction. The claim follows for Case 1. For Case 2, argue as above with $\overline{G}_e$ in place of G_e and then observe that if any of the points of $\overline{G}_e$ thus obtained were not in G_e, ν_0 would assign a strictly positive probability to $\{\infty\} \times D$, a contradiction.

In case ν_0 cannot be expressed as a convex combination of finitely many extreme points of G, a simple adaptation of the above proof works. In such a case, we claim that for any $j \geq 1$, we can find j linearly independent finite line segments in G which have ν_0 at the center. If this were not so for say, $j = j_0 + 1$, ν_0 would be in a j_0-dimensional face G' of G and therefore expressible as a convex combination of $j_0 + 1$ extreme points of G' and hence of G (see [9, p. 106]). This goes against the hypothesis, proving the claim. Now take $j \geq m+2$, consider the polytope generated by the end points of these line segments, and argue as above. □

Write ν_0 as

$$\nu_0(\{i\}, du) = \pi_0(i)\hat{\Phi}_i(du), \qquad i \in S, \tag{3.1}$$

corresponding to the SRS $\gamma[\Phi]$ with $\Phi = \prod_i \hat{\Phi}_i$. By Lemmas 2.3 and 3.1, it follows that ν_0 is a strict convex combination of some $\nu_1, \nu_2, \ldots, \nu_k \in G_e, k \leq m+1$, such that

$$\nu_j(\{i\}, du) = \pi_j(i)\delta_{\xi_j}(du), \quad i \in S, \quad 1 \leq j \leq k, \tag{3.2}$$

(where δ_u is the Dirac measure at u) with π_j an ergodic probability measure under $\gamma\{\xi_j\}$ for each j (cf. the remarks following Lemma 2.3). (It does not, however, follow that π_0 is ergodic under $\gamma[\Phi]$.)

In general, let $\pi_0 = \sum_{i=1}^{n} a_i \pi_{0i}$, where $a_i \in (0,1)$ with $\sum_{i=1}^{n} a_i = 1, n \geq 1$ (possibly $+\infty$) and $\{\pi_{0i}\}$ are ergodic probability measures under $\gamma[\Phi]$ with disjoint supports, denoted $\{S_i\}$, respectively.

LEMMA 3.2. *For each $j, 1 \leq j \leq k$, support (π_j) is contained in one of the S_i's.*

Proof. If the claim were false, two states in two distinct S_i's would communicate with each other under $\nu\{\xi_j\}$, and hence under $\gamma[\Phi]$ (cf. the proof of Lemma 2.3), a contradiction. □

For each $i \in S$, define finite subsets $N(i)$ of D by $N(i) = \{u \in D | u = \xi_j(i)$ for some $j, 1 \leq j \leq k$, satisfying $\pi_j(i) > 0\}$ and let $n(i) =$ (the cardinality of $N(i)) - 1$. From (2.1), it follows that for each $i, n(i)$ is the number of randomizations at i in the SRS $\gamma[\Phi]$. (By convention, $n(i) = 0$ if $N(i)$ is empty.)

Pick $i \in S$ (if any) such that $n(i) > 0$. Let $N(i) = \{u(1), u(2), \ldots, u(n(i)+1)\}$. Then by (2.1), $\hat{\Phi}_i$ is a strict convex combination of the Dirac measures at $u(j)$'s. Define $\gamma[\psi(i,j)]$ by: $\psi(i,j) = \prod_l \hat{\psi}_l(i,j)$ with

$$\hat{\psi}_l(i,j) = \hat{\Phi}_l, \quad \text{for } l \neq i,$$
$$= \text{the Dirac measure at } u(j), \quad \text{for } l = i.$$

LEMMA 3.3. *ν_0 is a strict convex combination of distinct elements $\mu_1, \ldots, \mu_{n(i)+1}$ of G such that μ_j is an ergodic occupation measure corresponding to $\gamma[\psi(i,j)]$ for each j. Furthermore, $\mu_1, \ldots, \mu_{n(i)+1}$ form the corners of an $(n(i)+1)$ simplex.*

Proof. If π_0 is ergodic, the first claim follows by iterating the argument used in the proof of Lemma 2.3 to show that ν is a strict convex combination of ν_1, ν_2. If not, let $\{\pi_{0l}\}$ be as above and apply the same argument to the π_{0l} for which $\pi_{0l}(i) > 0$. Suppose the second claim is false. Then ν_0 can be expressed as a strict convex combination of elements from $\{\mu_1, \ldots, \mu_{n(i)+1}\}$ in at least two distinct ways. But then (2.1) allows us to express the finitely supported probability measure $\hat{\Phi}_i$ as a strict convex combination of Dirac measures in two distinct ways, which is not possible. The claim follows. □

Call this $(n(i)+1)$ simplex the perturbation simplex at i and denote it by $Q(i)$.

LEMMA 3.4. *Let $\nu_1, \nu_2 \in Q(i)$ be distinct, with disintegrations*

$$\nu_j(\{l\}, du) = \pi_j(l)\varphi_{jl}(du), \quad l \in S, \quad j = 1,2.$$

Let $\pi_j(l) > 0$. Then $\varphi_{1l} = \varphi_{2l} = \hat{\Phi}_l$ for $l \neq i$ and $\varphi_{1i} \neq \varphi_{2i}$.

Proof. The claim for $l \neq i$ is immediate from (2.1). That $\varphi_{1i} \neq \varphi_{2i}$ follows from (2.1) and the easily verifiable fact that for $n > 1, b_1, \ldots, b_n > 0$, the map that maps

$$(a_1, \ldots, a_n) \in \left\{(x_1, \ldots, x_n) | x_i \in (0,1) \text{ for all } i, \sum x_i = 1\right\}$$

to

$$(a_1 b_1/c, a_2 b_2/c, \ldots, a_n b_n/c) \in (0,1)^n$$

with $c = \sum a_i b_i$ is one-one. □

The $(n(i)+1)$-simplex $Q(i)$ generates an $n(i)$-dimensional affine space in the space of finite signed measures on $S \times D$. Denote it by $Y(i)$.

LEMMA 3.5. *Let $j \neq i$ be another element of S such that $n(j) > 0$. Then $Y(i) \cap Y(j) = \{\nu_0\}$.*

Proof. Suppose not. Then we must have a $\overline{\nu} \in Q(i) \cap Q(j), \overline{\nu} \neq \nu_0$. Consider the line segment Z joining $\overline{\nu}, \nu_0$. $Z \subset Q(i) \cap Q(j)$ by the convexity of the latter. Show a typical $\nu \in Z$ as

$$\nu(\{i\}, du) = \pi(i)\varphi_i(du), \qquad i \in S.$$

As ν moves along Z, Lemma 3.4 and the fact that $Z \subset Q(i)$ imply that $\varphi_j = \hat{\Phi}_j$ all along, but φ_i keeps changing. Using $Z \subset Q(j)$, a similar conclusion can be drawn with the roles of i and j interchanged, leading to a contradiction. The claim follows. □

LEMMA 3.6. *Let $i_1, i_2, \ldots, i_{r+1}$ be different states with $n(i_j) > 0$ for all j and $\alpha_1, \alpha_2, \ldots, \alpha_r \in [0, 1]$ with $\sum_{i=1}^{r} \alpha_i = 1$. Then,*

$$(\alpha_1 Q(i_1) + \cdots + \alpha_r Q(i_r)) \cap Q(i_{r+1}) = \{\nu_0\}.$$

Proof. For $r = 1$, it reduces to the preceding lemma. The general case follows by induction on r using an argument analogous to the above at each step. □

COROLLARY 3.1. *Letting $L(i) = Y(i) - \nu_0$ when $n(i) > 0$, $\dim(L(i_1) + \cdots + L(i_r)) = \dim L(i_1) + \cdots + \dim L(i_r)$ whenever $n(i_1), \ldots, n(i_r) > 0, r \geq 1$ (i.e., $L(i_1) + \cdots + L(i_r)$ is a direct sum).*

Proof. The proof is immediate from the preceding lemma. □

LEMMA 3.7. $\sum_{i \in S} n(i) \leq m$.

Proof. Suppose not. Then there exists a finite subset $\{i_1, \ldots, i_l\} \subset S$ such that $n(i_1) + \cdots + n(i_l) \geq m + 1$. By the foregoing, the corners of $Q(i_1), \ldots, Q(i_l)$ together form a polytope with $n(i_1) + \cdots + n(i_l) \geq m + 1$-dimensional interior that contains ν_0. Now argue as in the proof of Lemma 3.1 to obtain a contradiction. □

Summarizing, we have the following.

THEOREM 3.1. *In both Case 1 and 2, there exists an optimal* SRS $\gamma[\Phi]$ *that requires at most m randomizations. Furthermore, if H has nonempty relative interior in G, there exist $\lambda_1, \ldots, \lambda_{2m} \geq 0$ such that for all $\nu \in G$ and $\mu_1, \ldots, \mu_{2m} \geq 0$,*

$$\begin{aligned}
&\int k_0 d\nu + \sum_{i=1}^{m} \lambda_i \left(\alpha_i - \int k_i d\nu \right) + \sum_{i=1}^{m} \lambda_{m+i} \left(\int k_i d\nu - \beta_i \right) \\
&\quad \geq \int k_0 d\nu_0 + \sum_{i=1}^{m} \lambda_i \left(\alpha_i - \int k_i d\nu_0 \right) + \sum_{i=1}^{m} \lambda_{m+i} \left(\int k_i d\nu_0 - \beta_i \right) \\
&\quad \geq \int k_0 d\nu_0 + \sum_{i=1}^{m} \mu_i \left(\alpha_i - \int k_i d\nu_0 \right) + \sum_{i=1}^{m} \mu_{m+i} \left(\int k_i d\nu_0 - \beta_i \right).
\end{aligned}$$

Proof. The first two claims are immediate in view of the foregoing. The last follows from standard Lagrange multiplier theory ([16, pp. 216–219]). □

4. Concluding remarks. (i) We have a priori restricted our attention to SRS rather than consider the optimization problems over all CS. If $\{k_i\}$ are bounded and S is a single communicating class under all CS, the following "pathwise" extension is available. Assume either the near-monotonicity condition or "condition B" of [7, p. 63] (which implies compactness of G, see [7, p. 60]). Define the $P(\overline{S} \times D)$-valued process $\{\nu_n\}$ by $\nu_n(A \times B) = \frac{1}{n} \sum_{m=0}^{n-1} I\{X_m \in A, \xi_m(X_m) \in B\}$ for A, B Borel in $\overline{S}, D$, respectively,

$\{\xi_n\}$ being the CS used. Argue as in [7, Lem 1.1, pp. 55–56] to conclude that, outside a set of zero probability, every limit point ν of $\{\nu_n\}$ in $P(\overline{S} \times D)$ is a convex combination of an element of G and a probability measure on $\{\infty\} \times D$. Under "condition B" mentioned above, we have $\nu \in G$ ([7, Lem. 2.2, p. 63]). The corresponding limits $\int k_i d\nu$ of $\int k_i d\nu_n, n \geq 1, 0 \leq i \leq m$, then clearly cannot outperform the optimal SRS by virtue of considerations analogous to those of §2 above.

(ii) For other classical cost criteria (infinite horizon discounted cost, finite horizon cost, cost up to a first exit time), their "occupation measure" formulations described in Chapters 3 and 4 of [7] allow us to use the above technique to derive analogous results for the corresponding constrained problems. It should be noted that the initial law plays a much more significant role for these problems—see the discussion in [7, p. 96–97], in particular, [7, Ex. 2, p. 97].

(iii) For the optimal cost $\int k_0 d\nu_0$ in the foregoing to be attainable from arbitrary initial law, we need that support (π_0) should be reachable in finite time with probability one from the given initial law under some CS. Use this CS until support (π_0) is hit and then switch over to the optimal SRS $\gamma[\Phi]$.

(iv) We can show the following incidental result which is of some interest: Suppose $\nu_1, \nu_2 \in G_e$ are distinct and correspond to $\gamma\{\xi_1\}, \gamma\{\xi_2\}$, respectively, with $\xi_j = [\xi_{j1}, \xi_{j2}, \ldots], j = 1, 2$. Suppose that $\xi_{1i} \neq \xi_{2i}$ implies $p(i, ., \xi_1(i)) \neq p(i, ., \xi_2(i))$. If the line segment Z joining ν_1, ν_2 is a face of G, then ξ_1, ξ_2 differ at exactly one state. To see this, note that any point in Z corresponds to some $\gamma[\Phi], \Phi = \prod \hat{\Phi}_i$, where each $\hat{\Phi}_i$ is a convex combination of Dirac measures at ξ_{1i}, ξ_{2i}. On the other hand, considerations leading to Theorem 3.1 show that at most one $\hat{\Phi}_i$ need be non-Dirac, proving the claim.

REFERENCES

[1] E. J. Anderson and P. Nash (1987), *Linear Programming in Infinite Dimensional Spaces,* John Wiley, Chichester.

[2] E. Altman and A. Schwartz (1991), *Markov decision problems and state-action frequencies,* SIAM J. Control Optim., 29, pp. 786–806.

[3] ——— (1991), *Adaptive control of constrained Markov chains,* IEEE Trans. Automat. Control, AC-36, pp. 454–462.

[4] E. J. Balder (1979), *On a useful compactification for optimal control problems,* J. Math. Anal. Appl., 72, pp. 391–398.

[5] F. J. Beutler and K. W. Ross (1985), *Optimal policies for controlled Markov chains with a constraint,* J. Math. Anal. Appl., 112, pp. 236–252.

[6] P. Billingsley (1968), *Convergence of Probability Measures,* John Wiley, New York.

[7] V. S. Borkar (1991), *Topics in Controlled Markov Chains,* Pitman Research Notes in Maths No. 240, Longman Scientific and Technical, Harlow, United Kingdom.

[8] ——— (1990), *Controlled Markov chains with constraints,* Sadhana, Indian Academy of Sciences Journal for Engineering Sciences, 15, pp. 405–413.

[9] G. Choquet (1967), *Lectures on Analysis II—Representation Theory,* W. A. Benjamin, Reading, MA.

[10] C. Derman (1970), *Finite State Markovian Decision Processes,* Academic Press, New York.

[11] L. Dubins (1962), *On extreme points of convex sets,* J. Math. Anal. Appl., 5, pp. 237–244.

[12] A. Hordijk and L. C. M. Kallenberg (1984), *Constrained undiscounted stochastic dynamic programming,* Math. Oper. Research, 9, pp. 276–289.

[13] L. C. M. Kallenberg (1983), *Linear Programming and Finite Markovian Control Problems,* Math. Centre Tracts No. 148, Math. Centre, Amsterdam.

[14] J. H. B. Kemperman (1983), *On the role of duality in the theory of moments,* in Lecture Notes in Economics and Math. Systems, 215, A. V. Fiacco and K. O. Kortanek, eds., Springer-Verlag, Berlin, Heidelberg, pp. 63–92.

[15] W. K. Klein Haneveld (1986), *Duality in stochastic linear and dynamic programming,* in Lecture Notes in Economics and Math. Systems, 274, Springer-Verlag, Berlin, Heidelberg.

[16] D. Luenberger (1969), *Optimization by Vector Space Methods,* John Wiley, New York.

[17] A. B. Piunovskii and V. M. Khametov (1991), *Optimal control by random sequences with constraints,* Math. Notes (translation of Mat. Zametki), 49, pp. 654–656.
[18] K. W. Ross (1989), *Randomized and past-dependent policies for Markov decision processes with multiple constraints,* Oper. Res., 37, pp. 474–477.
[19] L. I. Sennott (1991), *Constrained average cost Markov chains,* preprint.
[20] H. Witsenhausen (1980), *Some aspects of convexity useful in information theory,* IEEE Trans. Inform. Theory, IT-26, pp. 265–271.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 187–212, January 1994

012

FURTHER RESULTS ON LEAST SQUARES BASED ADAPTIVE MINIMUM VARIANCE CONTROL*

LEI GUO†

Abstract. Based on the recently established results on self-tuning regulators originally proposed by Åström and Wittenmark, this paper presents various novel and extended results on least squares based adaptive minimum variance control for linear stochastic systems. These results establish self-optimality, self-tuning property, and the best possible convergence rate of the control law in a variety of situations of interest.

Key words. stochastic adaptive control, self-tuning, least squares

AMS subject classifications. 93C40, 93E12, 93E10

1. Introduction.

1.1. System description. Consider the following SISO linear discrete-time stochastic system:

$$A(z)y_n = B(z)u_{n-1} + C(z)w_n, \qquad n \geq 0, \tag{1.1}$$

where $\{y_n\}, \{u_n\}$ and $\{w_n\}$ are the system output, input, and random disturbance sequences, respectively, $y_n = u_n = w_n = 0$ for all $n < 0$, and $A(z), B(z)$, and $C(z)$ are polynomials in backward-shift operator z:

$$\begin{aligned} A(z) &= 1 + a_1 z + \cdots + a_p z^p, \qquad p \geq 0, \\ B(z) &= b_1 + b_2 z + \cdots + b_q z^{q-1}, \qquad q \geq 1, \\ C(z) &= 1 + c_1 z + \cdots + c_r z^r, \qquad r \geq 0, \end{aligned}$$

with known upper bounds p, q, and r for true orders and unknown coefficients a_i, b_j, and c_k.

As usual, for the above model we adopt the following standard assumptions:

(A1) $\{w_n, \mathcal{F}_n\}$ is a martingale difference sequence, i.e., $E[w_{n+1}|\mathcal{F}_n] = 0$, and satisfies

$$\sup_n E[|w_{n+1}|^\beta | \mathcal{F}_n] < \infty, \text{ a.s. for some } \beta > 2, \tag{1.2}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} w_i^2 = \sigma^2 > 0 \quad \text{a.s.} \tag{1.3}$$

(A2) SPR condition: $\max_{|z|=1} |C(z) - 1| < 1$.

(A3) Minimum phase condition: $B(z) \neq 0$, for all $z : |z| \leq 1$.

Condition (A1) implies that the linear minimum variance predictor for y_{i+1} generated by (1.1) coincides with the minimum variance predictor $E[y_{i+1}|\mathcal{F}_i]$ if $\{u_i, \mathcal{F}_i\}$ is an adapted sequence. Condition (A2) is the usual SPR condition

$$\text{Re}\left\{\frac{1}{C(e^{j\lambda})} - \frac{1}{2}\right\} > 0 \quad \forall \lambda \in [0, 2\pi] \qquad (j \triangleq \sqrt{-1}),$$

* Received by the editors April 15, 1992; accepted for publication (in revised form) November 3, 1992. This work was supported by the National Natural Science Foundation of China.

† Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China.

which implies that $\sum_{i=1}^{r} c_i^2 < 1$, and is implied by $\sum_{i=1}^{r} |c_i| < 1$ (cf. Huang and Guo [1, pp. 1731, 1755]). This condition, together with the a priori knowledge about the orders p, q, and r, can be dispensed with for recursive identification of the linear model (1.1). We will not discuss that issue here and instead refer to Huang and Guo [1] for details. Condition (A3) is necessary for internal stability of minimum variance control systems even if the parameters in (1.1) are known (see, e.g., Kumar and Varaiya [2, p. 121]).

1.2. Performance. Our objective is to construct a control sequence $\{u_n\}$ based on the past and current observations, such that the following averaged square tracking error is asymptotically minimized:

$$J_n \triangleq \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2, \tag{1.4}$$

where $\{y_i^*\}$ is a reference sequence to be tracked, which is assumed to satisfy the following condition:

(A4) $\{y_i^*\}$ is bounded almost surely and is independent of $\{w_i\}$.

For convenience of discussions, we may assume without loss of generality that $\mathcal{F}_i = \sigma\{w_j, y_{j+1}^*, j \le i\}$. Then for any adapted input sequence $\{u_i, \mathcal{F}_i\}, y_i - y_i^* - w_i$ is $\mathcal{F}_{i-1}$-measurable for all i, and so by Chow's local convergence theorem for martingales (cf. [3]), it is easy to conclude that

$$J_n = \frac{1}{n}\sum_{i=1}^{n} w_i^2 + \begin{cases} R_n(1+o(1)) & \text{on } [nR_n \xrightarrow[n\to\infty]{} \infty]; \\ O(\frac{1}{n}) & \text{on } [\lim\limits_{n\to\infty}(nR_n) < \infty] \end{cases} \tag{1.5}$$

where R_n denotes the following "averaged regret":

$$R_n = \frac{1}{n}\sum_{i=1}^{n}(y_i = y_i^* - w_i)^2. \tag{1.6}$$

Consequently, by virtue of (1.3), we know that for any adapted sequence $\{u_i, \mathcal{F}_i\}$ the asymptotic lower bound to J_n is σ^2, and that

$$J_n \xrightarrow[n\to\infty]{} \sigma^2 \text{ a.s.} \iff R_n \xrightarrow[n\to\infty]{} 0 \text{ a.s.} \tag{1.7}$$

which justifies the familiar concept "globally convergent" or "self-optimizing" for an adaptive controller that leads to $R_n \xrightarrow[n\to\infty]{} 0$ a.s.. Moreover, from (1.5) it is apparent that R_n is of essential importance for the convergence rate of J_n, since it can be regarded as a second-order quantity (see also Wei [4, p. 1668]). It is also worth noting that once the self-optimality $R_n \xrightarrow[n\to\infty]{} 0$ a.s. is proved, the global stability, i.e., $\sup_n(1/n)\sum_{i=1}^{n}(y_i^2 + u_i^2) < \infty$ a.s., can be derived trivially by using Assumptions (A1), (A3), and (A4).

1.3. Estimation algorithms. Let us denote the unknown parameters in (1.1) by

$$\theta = [-a_1 \cdots - a_p, b_1 \cdots b_q, c_1 \cdots c_r]^\tau. \tag{1.8}$$

Then the model (1.1) can be succinctly written in a regression form:

$$y_{n+1} = \theta^\tau \varphi_n^0 + w_{n+1}, \qquad n \ge 0, \tag{1.9}$$

where φ_n^0 is the regression vector defined by

$$\varphi_n^0 = [y_n \cdots y_{n-p+1}, u_n \cdots u_{n-q+1}, w_n \cdots w_{n-r+1}]^\tau. \tag{1.10}$$

The standard method for estimating θ is the following recursive extended least squares (ELS) algorithm:

$$\theta_{n+1} = \theta_n + a_n P_n \varphi_n (y_{n+1} - \theta_n^\tau \varphi_n) \tag{1.11}$$

$$P_{n+1} = P_n - a_n P_n \varphi_n \varphi_n^\tau P_n, \quad a_n = (1 + \varphi_n^\tau P_n \varphi_n)^{-1}, \tag{1.12}$$

$$\varphi_n = [y_n \cdots y_{n-p+1}, \quad u_n \cdots u_{n-q+1}, \hat{w}_n \cdots \hat{w}_{n-r+1}]^\tau, \tag{1.13}$$

$$\hat{w}_n = y_n - \theta_n^\tau \varphi_{n-1} \tag{1.14}$$

with arbitrary initial values $\theta_0, \varphi_0 \neq 0$ and $P_0 \geq 0$.

There is a vast literature on strong consistency of the above ELS algorithm (see, e.g., Caines [5], Chen and Guo [6], and the references therein). In a Bayesian framework assuming Gaussianity of both the noise $\{w_n\}$ and the parameter θ, it was shown by Sternby [7] that in the white noise case (i.e., $C(z) = 1$), the necessary and sufficient condition for strong consistency of the least squares (LS) estimate θ_n is that

$$\lambda_{\min}(n) \xrightarrow[n\to\infty]{} \infty \text{ a.s.} \tag{1.15}$$

where $\lambda_{\min}(n)$ denotes the minimum eigenvalue of P_{n+1}^{-1}, i.e.,

$$\lambda_{\min}(n) \triangleq \lambda_{\min} \left\{ \sum_{i=0}^{n} \varphi_i \varphi_i^\tau + P_0^{-1} \right\}. \tag{1.16}$$

In the non-Bayesian framework where θ is an unknown constant vector as the case considered here, Lai and Wei [8] succeeded in showing that in the white noise case, strong consistency of the LS estimate still holds if (1.15) is strengthened into

$$\lambda_{\min}(n) \xrightarrow[n\to\infty]{} \infty, \qquad \frac{\log \lambda_{\max}(n)}{\lambda_{\min}(n)} \xrightarrow[n\to\infty]{} 0 \text{ a.s.} \tag{1.17}$$

where $\lambda_{\max}(n)$ denotes the maximum eigenvalue of P_{n+1}^{-1}. They also presented an example showing that relaxing the second part of (1.17) can result in a loss of strong consistency of the LS algorithm. The above consistency result can be easily generalized to colored noise and multivariable cases by resorting to the standard SPR condition (A2), and by using the standard recursions for the Lyapunov function studied earlier in (e.g., Ledwich and Moore [9], Solo [10], and Chen [11]) together with Chow's local convergence theorem for martingales (see [12] and [13]).

Despite the celebrated convergence properties of the ELS algorithm, the basic stability issue of adaptive minimum variance control constructed by using the ELS algorithm has been a long-standing problem over the past two decades. The main difficulty is that we do not know if the condition (1.17) really holds for the closed-loop systems. In fact, over the past decade, most of the results in stochastic adaptive control theory have been established for adaptive control laws that are not based on ELS algorithm but based on a stochastic gradient (SG) algorithm (or its variant). This algorithm is formed by simply replacing the matrix gain $\{a_n P_n\}$ in (1.11) by a scalar gain $\{\mu / r_n\}$ with $\mu > 0$, where

$$r_n \triangleq 1 + \sum_{i=0}^{n} \|\varphi_i\|^2. \tag{1.18}$$

Goodwin, Ramadge, and Caines [14] obtained the first stability and optimality result on SG-based adaptive tracking algorithms, which stimulated considerable research efforts afterwards. However, as is observed in simulations, the SG algorithm exhibits much slow convergence rate as compared with the ELS algorithm. Chen and Guo [15], [16], [6] have given a comprehensive theoretical study for the convergence of SG algorithm and justified the convergence phenomena known by simulations. To be precise, for strong consistency of SG, the following condition was introduced by Chen and Guo [15]:

$$r_n \xrightarrow[n\to\infty]{} \infty, \quad \frac{\lambda_{\max}(n)}{\lambda_{\min}(n)} = O(\{\log r_n\}^{\alpha}) \text{ a.s.}, \qquad \alpha \geq 0. \tag{1.19}$$

They showed that for the SG algorithm, if (1.19) holds with $\alpha \leq \frac{1}{4}$, then $\theta_n \xrightarrow[n\to\infty]{} \theta$, a.s. (see [15, Thm. 1], [16, Thm. 2], and [6, Thm. 4.5]). They also presented an example showing that in (1.19) the constant α is not allowed to be greater than 1 for strong consistency of SG (see [6, pp. 124–129]).

Hence for strong consistency, the SG algorithm requires much more excitation than the LS algorithm does (note that (1.19) is much stronger that (1.17)). Moreover, in the white noise case under the condition (1.19) with $\alpha = \frac{1}{4}$, the guaranteed convergence rate for the SG algorithm is only of the order $O(1/\log^{\lambda} r_n)$, i.e.,

$$\|\theta_n - \theta\|^2 = O\left(\frac{1}{\log^{\lambda} r_n}\right) \text{ a.s.} \quad \text{for some } \gamma > 0 \tag{1.20}$$

(cf. [15, p. 141] or [6, p. 132]), while under the same conditions, the convergence rate for the LS algorithm is much faster: $\|\theta_n - \theta\|^2 = O(\log^{\frac{5}{4}} r_n / r_n)$ a.s. (see, e.g., [6, p. 96] or [8, p. 155]).

1.4. Background. The standard adaptive minimum variance tracking control is constructed by simply identifying the adaptive predictor with the target value, i.e.,

$$\theta_n^{\tau}\varphi_n = y_{n+1}^{*}, \qquad n \geq 1, \tag{1.21}$$

where $\{\theta_n\}$ is generated by the ELS algorithm (1.11)–(1.14).

Åström and Wittenmark [17] were, apparently, the first to attempt an analysis of adaptive minimum variance control constructed by using LS-type estimates. They showed that if the LS parameter estimates should converge to some limit with no common factor, then the adaptive controller must necessarily be optimal. However, a difficult problem is whether these estimates are indeed convergent. To overcome this difficulty, Kumar [18] considered the case where the additive noise in (1.1) is *i.i.d.* and Gaussian. By using the technique of "Bayesian embedding," he succeeded in showing that, outside an exceptional set of true parameter vectors of Lebesgue measure zero, the LS based self-tuning minimum variance control enjoys various important convergence properties.

Recently, Guo and Chen [19] solved the basic stability and optimality problem of ELS-based adaptive minimum variance control for the general system (1.1) under the standard conditions (A1)–(A3). The following was shown in [19]:

(i) If the "high frequency" gain b_1 is known, then the standard ELS-based self-tuning tracker is globally stable and self-optimizing, with a rate of convergence for the regret: $R_n = O(d_n/n^{1-\varepsilon})$ a.s. for all $\varepsilon > 0$, where $\{d_n\}$ is a positive sequence satisfying $d_n \leq d_{n+1}, \sup_{n\geq 0}(d_{n+1}/d_n) < \infty$, and

$$\|w_n\|^2 = O(d_n) \quad \text{a.s.} \tag{1.22}$$

(ii) If b_1 is unknown, instead of using a fixed a priori estimate $\hat{b}_1$ for b_1 in designing the control law as in Åström and Wittenmark [17], a natural approach is to update this estimate with the current and past data. This was done in [19] by setting the on-line estimate (say $\hat{b}_1(n)$) to be

$$\hat{b}_1(n) = \begin{cases} b_1(n) & \text{if } |b_1(n)| \geq \dfrac{1}{\sqrt{\log r_{n-1}}}, \\ b_1(n) + \operatorname{sgn}(b_1(n)) \dfrac{1}{\sqrt{\log r_{n-1}}} & \text{otherwise,} \end{cases} \tag{1.23}$$

where $\operatorname{sgn}(\cdot)$ is the sign function, r_n is defined by (1.18) and $b_1(n)$ is the $(p+1)$th component of θ_n generated by the ELS algorithm (1.11)–(1.14). Then the resulting ELS-based adaptive control law is again shown to be stable and self-optimizing, with an implicitly established convergence rate $R_n = O(1/\log n)$ a.s..

The purpose of this paper is to give further results on ELS-based adaptive minimum variance control, with emphases placed on the convergence rate of R_n. We will improve the convergence rate obtained in [19] and show that in some cases the limit of $(n/\log n)R_n$ actually exists and is finite. We will also study the standard control law (1.21) (with no modifications on $b_1(n)$) and address the consistency issue of parameter estimates.

2. Preliminaries. To begin with, consider the regulation problem where $y_i^* \equiv 0$. Let $\lambda_{\min}(X)$ denote the minimum eigenvalue of a square matrix X. Then, from (1.9) it follows that

$$\lambda_{\min}\left(\sum_{i=0}^{n-1} \varphi_i^0 \varphi_i^{0\tau}\right) \|\theta\|^2 \leq \sum_{i=0}^{n-1} (\theta^\tau \varphi_i^0)^2 = \sum_{i=0}^{n-1} (y_{i+1} - w_{i+1})^2 \tag{2.1}$$

and so by (1.6),

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=0}^{n-1} \varphi_i^0 \varphi_i^{0\tau}\right) \|\theta\|^2 \leq R_n, \tag{2.2}$$

which implies that the "self-optimality" and "persistency of excitation" cannot hold simultaneously in general for the closed-loop system resulting from regulation (see also [20, pp. 372–373] for a related discussion). Moreover, from (2.2) it is clear that the better the convergence rate of the regret R_n, the poorer the excitation of the regressor φ_i^0 will have. This explains the familiar dilemma between estimation and control. From the following theorem, we will see which kind of excitation results we may have and how the degree of excitation of $\{\varphi_i^0\}$ depends on $\{y_i^*\}$ in a general asymptotically optimal tracking system.

For future reference, we list the following identifiability conditions.

(A5) The polynomials $B(z)$ and $A(z) - C(z)$ are coprime, and either $\partial B(z) = q - 1$ or $\partial(A(z) - C(z)) = \max(p, r)$, where and hereafter $\partial X(z)$ denotes the degree for a given polynomial $X(z)$ in dummy variable z.

(A6) The polynomials $A(z)$ and $B(z)$ are coprime with $|a_p| + |b_q| \neq 0$.

The following theorem extends some related results in [22].

THEOREM 2.1. *Consider the linear model* (1.1). *Let the regret* R_n *be defined by* (1.6), *and the Assumptions* (A1) *and* (A4) *be satisfied. Suppose that* $\{\tau_n\}$ *is a strictly increasing sequence of random integers such that* $R_{\tau_n+1} \xrightarrow[n\to\infty]{} 0$ *holds on a set* D *of positive probability; then the following two conclusions hold:*

(i)

$$\liminf_{n\to\infty} \lambda_{\min}\left\{\frac{1}{\tau_n}\sum_{i=0}^{\tau_n}\psi_i\psi_i^{\tau}\right\} > 0 \qquad a.s.\ on\ D, \tag{2.3}$$

provided that (A5) *holds, where*

$$\psi_i = [y_i \cdots y_{i-p^*+1}, u_{i-1}\cdots u_{i-q+1}]^{\tau}, \qquad p^* \triangleq \max(p,r). \tag{2.4}$$

(ii)

$$\liminf_{n\to\infty} \frac{\lambda_{\min}\left(\sum_{i=0}^{\tau_n}\varphi_i^0\varphi_i^{0\tau}\right)}{\lambda^*_{\min}(\tau_n)} > 0 \qquad a.s.\ on\ D, \tag{2.5}$$

provided that (A6) *holds, and that*

$$\sqrt{R_{\tau_n+1}+\frac{\log\log\tau_n}{\tau_n}} = o\left(\frac{\lambda^*_{\min}(\tau_n)}{\tau_n}\right) \qquad a.s.\ on\ D, \tag{2.6}$$

where φ_i^0 *is defined by* (1.10), *and*

$$\lambda^*_{\min}(n) = \lambda_{\min}\left(\sum_{i=1}^{n} Y_i^* Y_i^{*\tau}\right), \qquad Y_i^* = [y_i^* y_{i-1}^* \cdots y_{i-p-q+1}^*]^{\tau}. \tag{2.7}$$

We remark that Theorem 2.1 holds irrespective of the control law structure and the minimum phase condition (A3). Following some proof ideas used in Chen and Guo [22], we preface the proof of the theorem by four simple facts, which are stated as lemmas since they will be frequently referenced in the sequel.

For any polynomial $F(z)$, denote its L_2-norm $\|F(z)\|_2$ by

$$\|F(z)\|_2^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |F(e^{i\lambda})|^2 d\lambda.$$

In the sequel, we shall sometimes suppress the argument (z) for simplicity.

LEMMA 2.1. *Let* $F(z)$ *and* $G(z)$ *be two coprime polynomials, and* S_d *be a set of polynomials* $(M(z), N(z))$, *defined by*

$$S_d = \{(M(z),N(z)) : \|M(z)\|_2^2 + \|N(z)\|_2^2 = 1; \partial M + \partial N \le d;$$

$$\textit{and either } \partial M < \partial G \textit{ or } \partial N < \partial F\}.$$

Then for any integer $d \ge 0, \inf_{(M,N)\in S_d} \|FM + GN\|_2 > 0$.

Proof. Suppose that the converse assertion were true; then it would necessarily imply that

$$FM + GN = 0 \tag{2.8}$$

for some polynomial (M, N) in S_d and some integer $d \geq 0$. By the coprimeness of F and G, there exist polynomials L and H such that $FL + GH = 1$. If $\partial M < \partial G$, then $G \not\equiv 0$, and we have by (2.8)

$$M = M(FL + GH) = L(-GN) + MGH = G(MH - LN).$$

From this it is easy to conclude that $M = 0$. By (2.8), we then have $N = 0$ since G is a nonzero polynomial, and so $\|M\|_2 + \|N\|_2 = 0$. Similarly, if $\partial N < \partial F$, again we have $\|M\|_2 + \|N\|_2 = 0$. This contradicts with $(M, N) \in S_d$. □

LEMMA 2.2. *Let F and G be two coprime polynomials. For any integers $m \geq 0, n \geq 0$, and any sequence $\{z_k\}$, define for any $k \geq 0$,*

$$Z_k = [F(z), zF(z) \cdots z^m F(z), G(z), zG(z) \cdots z^n G(z)]^\tau z_k.$$

If either $m < \partial G$ or $n < \partial F$, then with $c \triangleq \inf_{(M,N)\in S_{m+n}} \|FM + GN\|_2^2 > 0$,

$$\lambda_{\min}\left(\sum_{i=0}^{k} Z_i Z_i^\tau\right) \geq c\lambda_{\min}\left(\sum_{i=0}^{k} Z_i^0 Z_i^{0\tau}\right) \quad \forall k \geq 1,$$

where S_{m+n} is defined as in Lemma 2.1, *and*

$$Z_k^0 = [z_k, z_{k-1} \cdots z_{k-s}]^\tau, \qquad s \triangleq \max\{m + \partial F, n + \partial G\}.$$

Proof. We first note that $c > 0$ is guaranteed by Lemma 2.1. For any $x \in \mathbb{R}^{n+m+2}$, $\|x\| = 1$, with $x = [\alpha_0 \cdots \alpha_m, \beta_0 \cdots \beta_n]^\tau$, set $M(z) = \alpha_0 + \cdots + \alpha_m z^m$ and $N(z) = \beta_0 + \cdots + \beta_n z^n$. We have for all $k \geq 1$,

$$\begin{aligned}
\lambda_{\min}\left(\sum_{i=0}^{k} Z_i Z_i^\tau\right) &= \inf_{\|x\|=1} \sum_{i=0}^{k} (x^\tau Z_i)^2 \\
&= \inf_{\|x\|=1} \sum_{i=0}^{k} [(M(z)F(z) + N(z)G(z))z_i]^2 \\
&\geq \inf_{\|x\|=1} \|MF + NG\|_2^2 \lambda_{\min}\left(\sum_{i=0}^{k} Z_i^0 Z_i^{0\tau}\right) \\
&\geq \inf_{(M,N)\in S_{m+n}} \|MF + NG\|_2^2 \lambda_{\min}\left(\sum_{i=0}^{k} Z_i^0 Z_i^{0\tau}\right). \qquad \square
\end{aligned}$$

LEMMA 2.3. *Let $x_k \in \mathbb{R}^d, (d > 0), k \geq 0$, be a vector sequence, $x_k = 0$, for all $k < 0$, and $F(z)$ be a polynomial with $\|F(z)\|_2 \not\equiv 0$. Set $\overline{x}_k = F(z)x_k$. Then we have for all $n > 0$,*

$$\lambda_{\min}\left(\sum_{k=0}^{n} x_k x_k^\tau\right) \geq \frac{1}{(\partial F + 1)\|F(z)\|_2^2} \lambda_{\min}\left(\sum_{k=0}^{n} \overline{x}_k \overline{x}_k^\tau\right).$$

Proof. Let the coefficients of $F(z)$ be $f_i, i = 0, \cdots, \partial F$. Then by the Schwarz inequality,

$$\lambda_{\min}\left(\sum_{k=0}^{n}\overline{x}_k\overline{x}_k^\tau\right) = \inf_{\|x\|=1}\sum_{k=0}^{n}(x^\tau\overline{x}_k)^2$$

$$= \inf_{\|x\|=1}\sum_{k=0}^{n}[F(z)x^\tau x_k]^2 = \inf_{\|x\|=1}\sum_{k=0}^{n}\left[\sum_{i=0}^{\partial F}f_i(x^\tau x_{k-i})\right]^2$$

$$\le \sum_{i=0}^{\partial F}f_i^2\inf_{\|x\|=1}\sum_{k=0}^{n}\sum_{i=0}^{\partial F}(x^\tau x_{k-i})^2 \le \|F(z)\|_2^2(\partial F+1)\lambda_{\min}\left(\sum_{k=0}^{n}x_kx_k^\tau\right). \qquad \square$$

We also need a simple corollary of the laws of the iterated logarithm for martingales established in Jain, Jogdeo, and Stout [21].

LEMMA 2.4. *Let* $\{w_i, \mathcal{F}_i\}$ *satisfy condition* (A1), *and* $\{f_i, \mathcal{F}_i\}$ *be an adapted sequence satisfying*

$$\sum_{i=1}^{n}f_i^2 = O(n) \quad a.s., \quad f_n^2 = O(n^\delta) \quad a.s., \text{ for some } \delta \in [0,1).$$

Then as $n \to \infty$,

$$\sum_{i=1}^{n}f_iw_{i+1} = O(\sqrt{n\log\log n}) \quad a.s.$$

Proof. We first consider the case $|f_i| \ge 1$ a.s., for all i. By the martingale convergence theorem in [3] and the Kronecker lemma it follows that

$$\sum_{i=1}^{n}(E[w_{i+1}^2|\mathcal{F}_i] - w_{i+1}^2) = o(n) \quad \text{a.s.}$$

So by (1.3)

$$\sum_{i=1}^{n}E[w_{i+1}^2|\mathcal{F}_i] = (1+o(1))\sigma^2n \quad \text{a.s.}$$

Consequently, by noting $|f_i| \ge 1$ a.s.,

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}f_i^2E[w_{i+1}^2|\mathcal{F}_i] \ge \sigma^2 > 0 \quad \text{a.s.}$$

Hence by applying Theorem 3.1 in [21] it is easy to see that the lemma is true.

In the general case, noting that $f_i = [f_i + \text{ sgn }(f_i)] - \text{ sgn}(f_i)$, and applying the just proved result to $\sum_{i=1}^{n}[f_i + \text{ sgn}(f_i)]w_{i+1}$ and $\sum_{i=1}^{n}\text{ sgn}(f_i)w_{i+1}$, respectively, we see that the desired result is also true. $\square$

We are now in a position to prove Theorem 2.1.

Proof of Theorem 2.1. Following Chen and Guo [22] or [16], set $\xi_i = y_i - y_i^* - w_i$, $z_i = \xi_i + y_i^*$. Then by the assumption we have

$$\frac{1}{\tau_n}\sum_{i=0}^{\tau_n+1}\xi_i^2 \xrightarrow[n\to\infty]{} 0 \quad \text{on } D. \tag{2.9}$$

Note that

$$y_i = w_i + y_i^* + \xi_i = w_i + z_i, \tag{2.10}$$

and then by (1.1),

$$B(z)u_i = [A(z)y_{i+1} - C(z)w_{i+1}] = [A(z) - C(z)]w_{i+1} + A(z)z_{i+1}. \tag{2.11}$$

Part (i). By Lemma 2.3 we need only to consider

$$\overline{\psi}_i \triangleq B(z)\psi_i \triangleq \psi_i^w + \psi_i^z, \tag{2.12}$$

where, by (2.3), (2.9), and (2.10),

$$\psi_i^w = [B(z), zB(z), \cdots z^{p^*-1}B(z), A(z) - C(z), \cdots z^{q-2}[A(z) - C(z)]]^\tau w_i,$$
$$\psi_i^z = [B(z), zB(z), \cdots z^{p^*-1}B(z), A(z), \cdots z^{q-2}A(z)]^\tau z_i.$$

By Lemma 2.2 we know that there exists $c > 0$ such that

$$\lambda_{\min}\left\{\sum_{i=0}^{n} \psi_i^w \psi_i^{w\tau}\right\} \geq c\lambda_{\min}\left\{\sum_{i=0}^{n} [w_i \cdots w_{i-s}]^\tau [w_i \cdots w_{i-s}]\right\}$$

holds for all $n > 0$, where $s = p^* + q - 2$. Consequently, by (A1),

$$\liminf_{n\to\infty} \frac{1}{n}\lambda_{\min}\left(\sum_{i=0}^{n} \psi_i^w \psi_i^{w\tau}\right) > 0 \quad \text{a.s.} \tag{2.13}$$

Let $\psi_i^{y^*}$ and ψ_i^{ξ} be defined in the same way as ψ_i^z (i.e., in the definition of ψ_i^z replace z_i by y_i^* and ξ_i, respectively). Then by the Schwarz inequality and (2.9), it is clear that

$$\frac{1}{\tau_n}\sum_{i=0}^{\tau_n} \psi_i^{\xi}\psi_i^{w\tau} \xrightarrow[n\to\infty]{} 0 \quad \text{on } D.$$

Also, by Assumptions (A1) and (A4),

$$\frac{1}{\tau_n}\sum_{i=0}^{\tau_n} \psi_i^{y^*}\psi_i^{w\tau} \xrightarrow[n\to\infty]{} 0 \quad \text{a.s.}$$

Hence we have

$$\frac{1}{\tau_n}\sum_{i=0}^{\tau_n} \psi_i^{z}\psi_i^{w\tau} \xrightarrow[n\to\infty]{} 0 \quad \text{a.s. on } D \tag{2.14}$$

Therefore, by (2.12)–(2.14) it is easy to see that

$$\liminf_{n\to\infty} \lambda_{\min}\left(\frac{1}{\tau_n}\sum_{i=0}^{\tau_n} \overline{\psi}_i \overline{\psi}_i^\tau\right) > 0 \quad \text{a.s. on } D.$$

From this and Lemma 2.3, the assertion (i) follows immediately.

Part (ii). Similarly, we consider the transformation $\overline{\varphi}_i^0 = B(z)\varphi_i^0$. By (2.10) and (2.11), $\overline{\varphi}_i^0$ can be decomposed as $\overline{\varphi}_i^0 = \varphi_i^w + \varphi_i^z$, where

$$\varphi_i^w = [zB(z)\cdots z^pB(z), A(z) - C(z)\cdots z^{q-1}[A(z) - C(z)], zB(z)\cdots z^\tau B(z)]^\tau w_{i+1},$$
$$\varphi_i^z = [zB(z),\cdots, z^pB(z), A(z),\cdots z^{q-1}A(z), 0,\cdots, 0]^\tau z_{i+1}.$$

Let $\varphi_i^{y^*}$ and φ_i^ξ be defined in the same way as for φ_i^z. For $x = w, z, y^*$ and ξ, let $\overline{\varphi}_i^x$ be the vector composed of the first $(p+q)$ elements of φ_i^x. Then by (A1), (A4), and Lemma 2.4 it is easy to see that

$$\sum_{i=0}^{\tau_n} \varphi_i^w \overline{\varphi}_i^{y^*\tau} = O(\sqrt{\tau_n \log\log \tau_n}) \quad \text{a.s.} \tag{2.15}$$

Let $x \in \mathbb{R}^{p+q+r}$ be any (random) vector, $\|x\| = 1$. Put $x = (\alpha^\tau, \beta^\tau)^\tau, \alpha \in \mathbb{R}^{p+q}, \beta \in \mathbb{R}^r$. Then by the Schwarz inequality, (2.15) and the fact that $\overline{\varphi}_i^z = \overline{\varphi}_i^\xi + \overline{\varphi}_i^{y^*}$,

$$\begin{aligned} x^\tau \sum_{i=0}^{\tau_n} \overline{\varphi}_i^0 \overline{\varphi}_i^{0\tau} x &= \sum_{i=0}^{\tau_n} (x^\tau \varphi_i^w + x^\tau \varphi_i^z)^2 = \sum_{i=0}^{\tau_n} (x^\tau \varphi_i^w + \alpha^\tau \overline{\varphi}_i^z)^2 \\ &= \sum_{i=0}^{\tau_n} (x^\tau \varphi_i^w)^2 + O\left(\sqrt{\sum_{i=0}^{\tau_n} w_i^2 \sum_{i=0}^{\tau_n+1} \xi_i^2}\right) + O(\sqrt{\tau_n \log\log \tau_n}) \\ &\quad + \sum_{i=0}^{\tau_n} (\alpha^\tau \overline{\varphi}_i^z)^2 \\ &\geq \sum_{i=0}^{\tau_n} (x^\tau \varphi_i^w)^2 + O(\tau_n \sqrt{R_{\tau_n+1}}) + O(\sqrt{\tau_n \log\log \tau_n}) \\ &\quad + \sum_{i=0}^{\tau_n} (\alpha^\tau \overline{\varphi}_i^{y^*})^2 \\ &\geq \sum_{i=0}^{\tau_n} (x^\tau \varphi_i^w)^2 + (c\|\alpha\|^2 + o(1))\lambda_{\min}^*(\tau_n) \quad \text{a.s. on } D \end{aligned} \tag{2.16}$$

where for the last inequality we have used the assumption (2.6) and Lemma 2.2, and where the quantities $c > 0$ and "$o(1)$" are independent of the vector x.

Now, suppose that the converse assertion of (2.5) were true; then by Lemma 2.3 we know that there would be a set $D_1 \subset D$ with $P(D_1) > 0$ such that

$$\liminf_{n\to\infty} \frac{\lambda_{\min}\left(\sum_{i=0}^{\tau_n} \overline{\varphi}_i^0 \overline{\varphi}_i^{0\tau}\right)}{\lambda_{\min}^*(\tau_n)} = 0 \quad \text{on } D_1.$$

From this and (2.16) it is not difficult to find vectors $x_n \in \mathbb{R}^{p+q+r}, \|x_n\| = 1, x_n = (\alpha_n^\tau, \beta_n^\tau)^\tau, \alpha_n \in \mathbb{R}^{p+q}$, and a subsequence of $\{\tau_n\}$, which is also denoted by $\{\tau_n\}$, such that

$$\|\alpha_n\| \xrightarrow[n\to\infty]{} 0 \quad \text{a.s. on } D_1, \tag{2.17}$$

and that

$$\frac{1}{\tau_n}\sum_{i=0}^{\tau_n}(x_n^\tau\varphi_i^w)^2 = \frac{1}{\tau_n}\sum_{i=0}^{\tau_n}\{\alpha_n^\tau\overline{\varphi}_i^w + \beta_n^\tau B(z)[w_i\cdots w_{i-r+1}]^\tau\}^2 \xrightarrow[n\to\infty]{} 0 \quad \text{a.s. on } D_1. \tag{2.18}$$

From (2.17) and (2.18), it is obvious that

$$\frac{1}{\tau_n}\sum_{i=0}^{\tau_n}\{\beta_n^\tau B(z)[w_i\cdots w_{i-r+1}]^\tau\}^2 \xrightarrow[n\to\infty]{} 0 \quad \text{a.s. on } D_1.$$

Consequently, from this and (A1), it follows that

$$\begin{aligned}
0 &= \lim_{n\to\infty}\frac{1}{\tau_n}\sum_{i=0}^{\tau_n}\{\beta_n^\tau B(z)[1,\cdots,z^{r-1}]^\tau w_i\}^2 \\
&\geq \lim_{n\to\infty}\|\beta_n^\tau B(z)[1,\cdots,z^{r-1}]^\tau\|_2^2\lambda_{\min}\left\{\frac{1}{\tau_n}\sum_{i=0}^{\tau_n}[w_i,\cdots,w_{i-q-r+2}]^\tau[w_i,\cdots,w_{i-q-r+2}]\right\} \\
&= \sigma^2\lim_{n\to\infty}\|\beta_n^\tau B(z)[1,\cdots,z^{r-1}]^\tau\|_2^2 \quad \text{a.s. on } D_1,
\end{aligned}$$

which obviously implies that $\beta_n \xrightarrow[n\to\infty]{} 0$ a.s. on D_1, and so by (2.17), $\|x_n\| \xrightarrow[n\to\infty]{} 0$ a.s. on D_1. This contradicts with $\|x_n\| = 1$, and hence assertion (ii) is also true. □

Before concluding this section, we list some basic properties of the ELS algorithm here, which will be used frequently in later sections.

LEMMA 2.5. *For the system* (1.1) *and the* ELS *algorithm* (1.11)–(1.14), *if Conditions* (A1) *and* (A2) *hold and* u_n *is* $\mathcal{F}_n$*-measurable for* $n \geq 1$, *then*

(i) $$\tilde{\theta}_{n+1}^\tau P_{n+1}^{-1}\tilde{\theta}_{n+1} = O(\log r_n) \quad a.s.,$$

(ii) $$\sum_{i=1}^{n+1}\|\hat{w}_i - w_i\|^2 = O(\log r_n) \quad a.s.,$$

(iii) $$\sum_{i=1}^{n}\frac{\|\tilde{\theta}_i^\tau\varphi_i\|^2}{1+\varphi_i^\tau P_i\varphi_i} = O(\log r_n) \quad a.s.,$$

where $\tilde{\theta}_n \triangleq \theta - \theta_n$, *and* r_n *is defined by* (1.18).

Except (i), this lemma is the same as Lemma 1 in [19], but (i) is actually also established in the proof of that lemma.

COROLLARY 2.1. *Under the same conditions and notations as in Lemma 2.5, the following property holds:*

$$\|\tilde{\theta}_{n+1}\|^2 + \|\hat{w}_{n+1} - w_{n+1}\|^2 + \frac{\|\tilde{\theta}_n^\tau\varphi_n\|^2}{1+\varphi_n^\tau P_n\varphi_n} = O(\log r_n) \quad a.s. \tag{2.19}$$

Proof. We need only to note that by (1.12) and the choice of the initial condition, $P_{n+1}^{-1} \geq P_0^{-1} > 0$, for all $n \geq 0$. □

3. Adaptive minimum variance control (with b_1 fixed). Throughout this section we assume that the "high-frequency" gain b_1 in the model (1.1) is known. The main consideration behind this is that results obtained in this case are relatively complete, which can indicate the greatest expected achievement in the general case.

Similar to (1.8)–(1.10), we rewrite (1.1) in the regression form

$$y_{n+1} - b_1 u_n = \theta^\tau \varphi_n^0 + w_{n+1}, \qquad n \geq 0, \tag{3.1}$$

but here θ and φ_n^0 should be defined as

$$\theta = [-a_1 \dots - a_p \quad b_2 \dots b_q \quad c_1 \dots c_r]^\tau, \tag{3.2}$$

$$\varphi_n^0 = [y_n \dots y_{n-p+1}, u_{n-1} \dots u_{n-q+1}, w_n \dots w_{n-r+1}]^\tau. \tag{3.3}$$

The standard ELS algorithm for estimating θ is as follows:

$$\theta_{n+1} = \theta_n + a_n P_n \varphi_n (y_{n+1} - b_1 u_n - \theta_n^\tau \varphi_n), \tag{3.4}$$

$$P_{n+1} = P_n - a_n P_n \varphi_n \varphi_n^\tau P_n, \quad a_n = (1 + \varphi_n^\tau P_n \varphi_n)^{-1}, \tag{3.5}$$

$$\varphi_n = [y_n \dots y_{n-p+1}, u_{n-1} \dots u_{n-q+1}, \hat{w}_n \dots \hat{w}_{n-r+1}]^\tau, \tag{3.6}$$

$$\hat{w}_n = y_n - b_1 u_{n-1} - \theta_n^\tau \varphi_{n-1}, n \geq 0, \hat{w}_n = 0, n < 0, \tag{3.7}$$

with arbitrary initial values θ_0, φ_0, and $P_0 > 0$.

We note that Lemma 2.5 and Corollary 2.1 also hold for the present algorithm, and in what follows we shall use them directly without explanations.

The "certainty equivalent" minimum variance adaptive control is defined by

$$u_n = b_1^{-1}(y_{n+1}^* - \theta_n^\tau \varphi_n). \tag{3.8}$$

We first treat the white noise case.

THEOREM 3.1. *Consider the system* (1.1) *with* $r = 0$ *and* $E[w_{n+1}^2 | \mathcal{F}_n] = \sigma^2 > 0$, *a.s. for all* $n \geq 0$. *Suppose that* (A1) *and* (A3)–(A5) *hold. If the control law* (3.4)–(3.8) *is applied, then the closed-loop system has the following properties:*

$$\lim_{n \to \infty} \left(\frac{n}{\log n} \right) R_n = (p + q - 1)\sigma^2 \quad a.s., \tag{3.9}$$

and

$$\|\theta_n - \theta\|^2 = O\left(\frac{\log \log n}{n} \right) \quad a.s., \quad as\ n \to \infty, \tag{3.10}$$

where R_n *is defined by* (1.6), *and* θ *is given by* (3.2) *with* $r = 0$.

Proof. By Theorem 1 of Guo and Chen [19] we know that $R_n \xrightarrow[n \to \infty]{} 0$ a.s., and that

$$\sum_{i=0}^{n} \|\varphi_i\|^2 = O(n) \quad \text{a.s.} \tag{3.11}$$

Hence by Theorem 2.1 (i), we have the following persistency of excitation property:

$$\liminf_{n\to\infty} \lambda_{\min}\left(\frac{1}{n}\sum_{i=0}^{n}\varphi_i\varphi_i^\tau\right) > 0 \quad \text{a.s.} \tag{3.12}$$

Also, by combining Lemma 2, (2.9), and Theorem 1 of Guo and Chen [19] we know that

$$\|\varphi_n\|^2 = O(n^\delta), \quad \text{a.s. } \forall \delta \in \left(\frac{2}{\beta}, 1\right), \tag{3.13}$$

where β is defined in (1.2). Hence, by (3.12) and (3.13),

$$\varphi_n^\tau P_{n+1}\varphi_n \xrightarrow[n\to\infty]{} 0 \quad \text{a.s.} \tag{3.14}$$

By (3.11), (3.12), and (3.14) we know that Theorem 3 of Wei [4] is applicable (there is a slight difference between the LS estimates defined there and here due to initial conditions, but that is not essential since (3.12) has been established), and hence we have

$$\sum_{i=0}^{n-1}(\theta^\tau\varphi_i - \theta_i^\tau\varphi_i)^2 \sim \sigma^2 \log\det\left(\sum_{i=0}^{n-1}\varphi_i\varphi_i^\tau\right) \quad \text{a.s.} \tag{3.15}$$

But by (3.11) and (3.12) it is easy to verify that $\log\det\left(\sum_{i=1}^{n-1}\varphi_i\varphi_i^\tau\right) \sim (p+q-1)\log n$. Hence, by combining (1.6), (3.1), (3.8), and (3.15) we see that (3.9) holds.

As for the second assertion of the theorem, by (3.4) and (3.5) we can express the estimation error as

$$\theta_n - \theta = P_n P_0^{-1}(\theta_0 - \theta) + P_n\sum_{i=0}^{n-1}\varphi_i w_{i+1}. \tag{3.16}$$

By (3.11), (3.13), and Lemma 2.4, we know that

$$\left\|\sum_{i=0}^{n-1}\varphi_i w_{i+1}\right\| = O(\sqrt{n\log\log n}) \quad \text{a.s.} \tag{3.17}$$

Finally, combining (3.12), (3.16), and (3.17) it is easy to see that (3.10) holds. □

Remark 3.1. The property (3.9) asserts, among other things, that $O(\log n/n)$ is the best convergence rate for the regret R_n generated by LS-based adaptive control. The convergence rate $O(\log\log n/n)$ in (3.10) is also obviously the best possible for the estimation error, since it is the same rate as that in the laws of the iterated logarithm. In a Bayesian framework, assuming that $\{w_i\}$ is *i.i.d.* with a Gaussian distribution $N(0,\sigma^2)$ and that θ has a certain truncated Gaussian prior distribution π, Lai [23] showed that under some stability conditions on the system and some regularity conditions on the input sequence $\{u_n\}$, the order $(p+q-1)\sigma^2(1+o(1))\log n/n$ is a *lower bound* to the expected regret $E_\pi[R_n]$ in the regulation problem. According to Lai's definition in [23, p. 37], the control law of Theorem 3.1 is "asymptotically efficient". It is also interesting to note that when the system orders p and q are increasing with the time (or data size) n, similar results as (3.15) are also obtainable (see, Huang and Guo [1]).

Next, we consider the general colored noise case $r > 0$. Let us write θ_n defined by (3.4)–(3.7) in its component form:

$$\theta_n = [-a_{1n}, \cdots, -a_{pn}, b_{2n}, \cdots, b_{qn}, c_{1n}, \cdots, c_{rn}]^\tau, \tag{3.18}$$

and set

$$\theta_n^* = [c_{1n} - a_{1n}, \cdots, c_{p^*n} - a_{p^*n}, b_{2n}, \cdots, b_{qn}]^\tau, \qquad p^* = \max(p, r), \tag{3.19}$$

where by definition $c_{in} = a_{jn} = 0$, for $i > r, j > p$.

Similarly, denote ($c_i = a_j = 0$, for all $i > r, j > p$),

$$\theta^* = [c_1 - a_1, \cdots, c_{p^*} - a_{p^*}, b_2, \cdots, b_q]^\tau. \tag{3.20}$$

It is easy to see that (cf. [2, p. 122]) for the regulation problem $y_i^* \equiv 0$ with b_1 known, to construct the nonadaptive (asymptotically) optimal control law, it is sufficient to know only θ^*, and hence θ^* may be regarded as the "true parameter."

THEOREM 3.2. *Let* (A1)–(A4) *hold, and let the adaptive control law* (3.4)–(3.8) *be applied to the system* (1.1).

(i) *For the regulation problem* $y_i^* \equiv 0$, *if* (A5) *holds, then*

$$\|\theta_n^* - \theta^*\|^2 + R_n = O\left(\frac{d_n}{n^{1-\varepsilon}}\right) \quad a.s. \forall \varepsilon > 0, \tag{3.21}$$

where d_n *is defined as in* (1.22), *and* θ_n^* *and* θ^* *are respectively defined by* (3.19) *and* (3.20).

(ii) *For the general tracking problem, if* (A6) *holds and* $\{y_i^*\}$ *satisfies*

$$n^{\frac{1+\delta}{2}}\sqrt{d_n} = O(\lambda_{\min}^*(n)) \quad a.s. \text{ for some } \delta > 0, \tag{3.22}$$

where d_n *and* $\lambda_{\min}^*(n)$ *are defined in* (1.22) *and* (2.7), *respectively, then as* $n \to \infty$,

$$R_n = O\left(\frac{\log n}{n}\right) a.s., \qquad \|\theta_n - \theta\|^2 = O\left(\frac{\log n}{\lambda_{\min}^*(n)}\right) a.s., \tag{3.23}$$

where θ_n *and* θ *are respectively given by* (3.4) *and* (3.2).

Proof. (i) By Theorem 1 in [19] we know that $R_n = 0(d_n/n^{1-\varepsilon})$, a.s., for all $\varepsilon > 0$. So for (3.21) we need only to consider the convergence rate of the estimation error. By Lemma 2.5 (i) we know that

$$\tilde{\theta}_{n+1}^\tau P_{n+1}^{-1} \tilde{\theta}_{n+1} = O(\log r_n) = O(\log n), \text{ a.s.} \tag{3.24}$$

where $\tilde{\theta}_{n+1} = \theta_{n+1} - \theta$. By (3.19) and (3.20) and the fact that $P_{n+1}^{-1} = \sum_{i=0}^n \varphi_i \varphi_i^\tau + P_0^{-1}$, we can rewrite (3.24) as

$$\sum_{i=0}^n \left[\psi_i^\tau \tilde{\theta}_{n+1}^* + \sum_{j=1}^{p^*}(c_{jn+1} - c_j)(\hat{w}_{i-j+1} - y_{i-j+1})\right]^2 + \tilde{\theta}_{n+1}^\tau P_0^{-1} \tilde{\theta}_{n+1} = O(\log n), \text{ a.s.},$$

where $\tilde{\theta}_{n+1}^* = \theta_{n+1}^* - \theta^*$, and ψ_i and p^* are defined by (2.4). By Lemma 2.5(ii), Corollary 2.1 and the fact that

$$\sum_{i=0}^n (y_i - w_i)^2 = O(n^\varepsilon d_n) \quad \text{a.s., } \forall \varepsilon > 0,$$

it is easy to see that

$$\sum_{i=0}^n \left[\sum_{j=1}^{p^*}(c_{jn+1} - c_j)(\hat{w}_{i-j+1} - y_{i-j+1})\right]^2 = O(n^\varepsilon d_n) \quad \text{a.s. } \forall \varepsilon > 0.$$

Therefore, we have

$$\sum_{i=0}^{n}(\psi_i^\tau\tilde{\theta}_{n+1}^*)^2 = O(n^\varepsilon d_n) \quad \text{a.s.}, \forall\varepsilon > 0.$$

From this and Theorem 2.1 (i), we obtain $\|\tilde{\theta}_{n+1}^*\|^2 = O(d_n/n^{1-\varepsilon})$, a.s.. This proves assertion (i).

(ii). Again, by [19], $R_n = O\left(\frac{d_n}{n^{1-\varepsilon}}\right)$, a.s., $\forall\varepsilon > 0$. Hence, by (3.22) we know that Theorem 2.1 (ii) is applicable, and so we have

$$\liminf_{n\to\infty}\frac{\lambda_{\min}\left(\sum_{i=0}^{n}\varphi_i^0\varphi_i^{0\tau}\right)}{\lambda_{\min}^*(n)} > 0 \quad \text{a.s.}$$

Consequently, by Lemma 2.5 (ii) and the fact that

$$\lambda_{\min}\left(\sum_{i=0}^{n}\varphi_i^0\varphi_i^{0\tau}\right) \le 2\lambda_{\min}\left(\sum_{i=0}^{n}\varphi_i\varphi_i^\tau\right) + 2\sum_{i=0}^{n}\|\varphi_i^0 - \varphi_i\|^2,$$

we have

$$\liminf_{n\to\infty}\frac{\lambda_{\min}\left(\sum_{i=0}^{n}\varphi_i\varphi_i^\tau\right)}{\lambda_{\min}^*(n)} > 0 \quad \text{a.s.}, \tag{3.25}$$

which in conjunction with (3.24) yields the second assertion in (3.23). By (3.22), (3.25) and Lemma 2 in [19] it is not difficult to see that $\varphi_n^\tau P_n\varphi_n \xrightarrow[n\to\infty]{} 0$. Therefore, by Lemma 2.5(iii), $\sum_{i=0}^{n}\|\tilde{\theta}_i^\tau\varphi_i\|^2 = O(\log n)$, and hence the first assertion in (3.23) is also true. □

Remark 3.2. For the regulation problem, the one degree of freedom identifiability problem as shown in Becker, Kumar, and Wei [24] does not occur in Theorem 3.2, since b_1 is not estimated. For the general tracking problem, it is clear that in Theorem 3.2, $\{y_i^*\}$ is not necessarily required to be "sufficiently rich" or "persistently exciting." Condition (3.22) is considerably weaker than the corresponding nonpersistence of excitation condition used in [22] and [16] for the SG-based algorithm. It would be of interest to establish similar results for a lower-dimensional ELS-based adaptive controller when $\{y_i^*\}$ is generated by a homogeneous finite-order linear difference equation $H(z)y_i^* = 0$, as was done by Kumar and Praly [25] for the SG-based algorithm.

4. Adaptive minimum variance control (the general case). In the general case where b_1 is not available, the analysis becomes much more complicated. Throughout this section, we assume that $\{\theta_n\}$ is generated by the ELS algorithm (1.11)–(1.14).

First, the minimum variance adaptive control law defined from (1.21) can be explicitly written as

$$u_n = \frac{1}{b_1(n)}\{y_{n+1}^* + (b_1(n)u_n - \theta_n^\tau\varphi_n)\}, \tag{4.1}$$

provided that $b_1(n) \neq 0$ a.s., where $b_1(n)$ is the ELS estimate for b_1 given by θ_n.

When (4.1) is applied, the first problem is that the set $\{b_1(n) = 0\}$ may have a positive probability, which is known as the zero divisor problem in stochastic adaptive control (cf. Meyn and Caines [26]). There are at least three ways to deal with this problem.

(a) Guarantee $P\{b_1(n) = 0\} = 0$ by assuming that all finite-dimensional distributions of $\{w_n\}$ are absolutely continuous with respect to Lebesgue measure (see, [26] or [5, pp. 778–782]). The absolute continuity assumption can be weakened to continuity only if $\{w_n\}$ is an independent sequence (cf. [16]).

(b) Guarantee $P\{b_1(n) = 0\} = 0$ by adding an independent random sequence with continuous distributions to the input signal. Such a sequence is preferably decaying with the time so that it does not upset the control performance (cf. [22]).

(c) Replace $b_1(n)$ appearing in the denominator of (4.1) by a quantity (say $\hat{b}_1(n)$), which is close to $b_1(n)$ but does not equal to zero (see, e.g., (1.23) or [19]).

In the sequel, whenever the control law (4.1) is concerned we always assume that $P\{b_1(n) = 0\} = 0$. The following lemma plays a key role in this section.

LEMMA 4.1. *For the system* (1.1) *assume that* (A1)–(A4) *are satisfied. At each time instant* n, *let the control law* u_n *be defined from the following equation:*

$$y_{n+1}^* = \theta_n^\tau \varphi_n + \Delta\hat{b}_{1n} u_n, \tag{4.2}$$

where $\{\theta_n\}$ *is given by the* ELS *algorithm* (1.11)–(1.14), *and* $\Delta\hat{b}_{1n} \in \mathcal{F}_n$ *is such that either* $\Delta\hat{b}_{1n} \equiv 0, \forall n$ *or* $\Delta\hat{b}_{1n} \xrightarrow[n\to\infty]{} 0$ *a.s. Then for any strictly increasing random sequence* $\{\tau_n\}$ *satisfying*

$$\inf_n |b_1(\tau_n + 1) - b_1| > 0 \quad \text{a.s. on } D, \tag{4.3}$$

with $P(D) > 0$, *and with* $b_1(n)$ *being the component of* θ_n *estimating* b_1, *the following properties hold as* $n \to \infty$:

$$\sup_{k \le \tau_n} \|\varphi_k\|^2 = O(\tau_n^\varepsilon d_{\tau_n}) \quad \text{a.s. on } D, \forall \varepsilon > 0, \tag{4.4}$$

and

$$r_{\tau_n} = O(\tau_n) \quad \text{a.s. on } D, \tag{4.5}$$

where r_n *and* d_n *are defined by* (1.18) *and* (1.22), *respectively.*

Proof. Before starting the proof, we remark that the case $\Delta\hat{b}_{1n} \equiv 0$ corresponds to the control law (4.1), while the case $\Delta\hat{b}_{1n} \neq 0$ corresponds to a (slight) modification of $b_1(n)$.

We first prove (4.4). By (1.9) and (4.2) we have with $\tilde{\theta}_k \triangleq \theta - \theta_k$,

$$\begin{aligned} y_{k+1} &= \theta^\tau \varphi_k + \theta^\tau(\varphi_k^0 - \varphi_k) + w_{k+1} \\ &= \tilde{\theta}_k^\tau \varphi_k + y_{k+1}^* - \Delta\hat{b}_{1n} u_k + \theta^\tau(\varphi_k^0 - \varphi_k) + w_{k+1}. \end{aligned} \tag{4.6}$$

Following Guo and Chen [19], denoting $\delta_k = tr(P_k - P_{k+1}), \alpha_k = \|\tilde{\theta}_k^\tau \varphi_k\|^2/(1 + \varphi_k^\tau P_k \varphi_k)$, and using Corollary 2.1 and the fact that $\varphi_k^\tau P_{k+1} \varphi_k \le 1$, we have by (4.6)

$$\begin{aligned} y_{k+1}^2 &\le 3\|\tilde{\theta}_k^\tau \varphi_k\|^2 + 3(\Delta\hat{b}_{1k})^2 u_k^2 + O(\log r_k + d_k) \\ &\le 3\alpha_k\{1 + \varphi_k^\tau P_{k+1} \varphi_k + \varphi_k^\tau (P_k - P_{k+1})\varphi_k\} \\ &\qquad + 3(\Delta\hat{b}_{1k})^2 u_k^2 + O(\log r_k + d_k) \\ &\le 3\alpha_k\{2 + \delta_k \|\varphi_k\|^2\} + 3(\Delta\hat{b}_{1k})^2 u_k^2 + O(\log r_k + d_k) \\ &= 3\alpha_k \delta_k \|\varphi_k\|^2 + 3(\Delta\hat{b}_{1k})^2 u_k^2 + O(\log r_k + d_k). \end{aligned} \tag{4.7}$$

By the stability of $B(z)$ and (1.1) there exists a constant $\lambda \in (0,1)$ such that

$$u_k^2 = O\left(\sum_{i=0}^{k+1} \lambda^{k-i} y_i^2\right) + O\left(\sum_{i=0}^{k+1} \lambda^{k-i} w_i^2\right). \tag{4.8}$$

Consequently,

$$\begin{aligned}[\|\varphi_k\|^2 - u_k^2] &= O\left(\sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + O\left(\sum_{i=0}^{r} (\hat{w}_{k-i})^2\right) + O\left(\sum_{i=0}^{k} \lambda^{k-i} w_i^2\right) \\ &= O\left(\sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + O(\log r_k + d_k) \quad \text{a.s.},\end{aligned} \tag{4.9}$$

where for the last relationship we have used Lemma 2.5 (ii).

Note that $P_{n+1}^{-1} = \sum_{i=0}^{n} \varphi_i \varphi_i^\tau + P_0^{-1}$, and we have by Lemma 2.5 (i),

$$\sum_{i=0}^{n} \|\tilde{\theta}_{n+1}^\tau \varphi_i\|^2 = O(\log r_n) \quad \text{a.s.}, \tag{4.10}$$

and consequently,

$$\max_{i \le n} \|\tilde{\theta}_{n+1}^\tau \varphi_i\|^2 = O(\log r_n) \quad \text{a.s.} \tag{4.11}$$

For simplicity of statements, we shall omit the phrase "a.s. on D" in the remainder of the proof, and unless otherwise stated all relationships hold on D with a possible exception set of probability zero. Denote $\tilde{b}_1(\tau_n+1) = b_1 - b_1(\tau_n+1)$, we have by (4.3), $\inf_n |\tilde{b}_1(\tau_n+1)| > 0$. Consequently, by (4.11) and the fact that $\|\tilde{\theta}_{n+1}\|^2 = 0(\log r_n)$ a.s., we have for all $k \le \tau_n$ and all $n \ge 1$,

$$\begin{aligned} u_k^2 &= \frac{1}{(\tilde{b}_1(\tau_n+1))^2} \{\tilde{b}_1(\tau_n+1) u_k\}^2 \\ &= \frac{1}{(\tilde{b}_1(\tau_n+1))^2} \{[\tilde{\theta}_{\tau_n+1}^\tau \varphi_k - \tilde{b}_1(\tau_n+1) u_k] - \tilde{\theta}_{\tau_n+1}^\tau \varphi_k\}^2 \\ &\le \frac{2}{(\tilde{b}_1(\tau_n+1))^2} \{\|\tilde{\theta}_{\tau_n+1}^\tau \varphi_k - \tilde{b}_1(\tau_n+1) u_k\|^2 + \|\tilde{\theta}_{\tau_n+1}^\tau \varphi_k\|^2\} \\ &= O((\log r_{\tau_n})[\|\varphi_k\|^2 - u_k^2]) + O(\log r_{\tau_n}) \\ &= O((\log r_{\tau_n}) \sum_{i=0}^{k} \lambda^{k-i} y_i^2) + O(\log^2 r_{\tau_n} + d_{\tau_n} \log r_{\tau_n}), \end{aligned} \tag{4.12}$$

where for the last relationship we have used (4.9), and where and hereafter the "O" constant depends neither on k nor on n.

Combining (4.9) and (4.12) we get for all $k \le \tau_n$,

$$\|\varphi_k\|^2 = O([\log r_{\tau_n}] \sum_{i=0}^{k} \lambda^{k-i} y_i^2) + O(\log^2 r_{\tau_n} + d_{\tau_n} \log r_{\tau_n}). \tag{4.13}$$

Substituting (4.8) and (4.13) into (4.7) and noticing $\Delta \hat{b}_{1n} \xrightarrow[n \to \infty]{} 0$, it is easy to see that for all $k \le \tau_n$, and all large n,

$$y_{k+1}^2 = O\left(\alpha_k \delta_k \log r_{\tau_n} \sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + o\left(\sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + O(\log^3 r_{\tau_n} + d_{\tau_n} \log^2 r_{\tau_n}). \tag{4.14}$$

Now, following [19] we set $L_k = \sum_{i=0}^{k} \lambda^{k-i} y_i^2$. Then by (4.14) there are constants $\delta > 0$ and $c > 0$ such that $(1+\delta)\lambda < 1$ and that

$$y_{k+1}^2 \le \lambda[(1+\delta)c\alpha_k\delta_k(\log r_{\tau_n}) + \delta]L_k + O(\log^3 r_{\tau_n} + d_{\tau_n}\log^2 r_{\tau_n})$$

holds for all suitably large n and all $k \le \tau_n$. Consequently, by denoting $\gamma \triangleq (1+\delta)\lambda < 1$, we obtain for large n and for all $k \le \tau_n$,

$$L_{k+1} = \lambda L_k + y_{k+1}^2 \le \gamma(1 + c\alpha_k\delta_k \log r_{\tau_n})L_k + O(\log^3 r_{\tau_n} + d_{\tau_n}\log^2 r_{\tau_n}).$$

Hence, iterating this inequality k times we get for all large n and for all $k \le \tau_n$,

$$\begin{aligned} L_{k+1} &\le \gamma^{k+1}\prod_{i=0}^{k}(1 + c\alpha_i\delta_i \log r_{\tau_n})L_0 \\ &\quad + O\left(\sum_{i=0}^{k}\gamma^{k-i}\prod_{j=i+1}^{k}(1 + c\alpha_j\delta_j \log r_{\tau_n})[\log^3 r_{\tau_n} + d_{\tau_n}\log^2 r_{\tau_n}]\right). \end{aligned} \tag{4.15}$$

By Lemma 2.5 (iii) and the convergency of the series $\sum_{i=1}^{\infty}\delta_i$, we know that for any small $\varepsilon > 0$, there exists $i > 0$ large enough such that

$$\varepsilon^2\sum_{j=i+1}^{k}\alpha_j \le \frac{\varepsilon}{2}\log r_k, \qquad \varepsilon^{-2}c\sum_{j=i+1}^{\infty}\delta_j < \frac{\varepsilon}{2}.$$

Hence we have for all $i \le k \le \tau_n$,

$$\begin{aligned} \prod_{j=i+1}^{k}(1 + c\alpha_j\delta_j \log r_{\tau_n} &\le \prod_{j=i+1}^{k}(1+\varepsilon^2\alpha_j)\prod_{j=i+1}^{k}(1 + c\varepsilon^{-2}\delta_j \log r_{\tau_n}) \\ &\le \exp\left\{\varepsilon^2\sum_{j=i+1}^{k}\alpha_j + \sum_{j=i+1}^{k}c\varepsilon^{-2}\delta_j \log r_{\tau_n}\right\} \le r_{\tau_n}^{\varepsilon}. \end{aligned} \tag{4.16}$$

Substituting this into (4.15), it is easy to conclude that for large n,

$$L_{k+1} = O(r_{\tau_n}^{\varepsilon} d_{\tau_n}), \quad \forall k \le \tau_n, \quad \forall \varepsilon > 0. \tag{4.17}$$

Substituting this into (4.13) we know that $\sup_{k\le\tau_n}\|\varphi_k\|^2 = O(r_{\tau_n}^{\varepsilon} d_{\tau_n})$ for all $\varepsilon > 0$, and hence (4.4) will be true if (4.5) is proved.

We now prove (4.5). By (4.17) and the assumption $\Delta\hat{b}_{1n} \xrightarrow[n\to\infty]{} 0$, it follows from (4.6) that

$$\begin{aligned} \sum_{k=1}^{\tau_n} y_{k+1}^2 &= O\left(\sum_{k=1}^{\tau_n}\alpha_k(1 + \varphi_k^{\tau}P_k\varphi_k)\right) + o\left(\sum_{k=1}^{\tau_n}u_k^2\right) + O(\log r_{\tau_n} + \tau_n) \\ &= O\left(r_{\tau_n}^{\varepsilon} d_{\tau_n}\sum_{k=1}^{\tau_n}\alpha_k\right) + o(r_{\tau_n}) + O(\log r_{\tau_n} + \tau_n), \quad \forall \varepsilon > 0. \end{aligned}$$

From this, Lemma 2.5, and (4.8), it is easy to see that

$$r_{\tau_n} = O(r_{\tau_n}^{\varepsilon} d_{\tau_n}\log r_{\tau_n}) + o(r_{\tau_n}) + O(\tau_n), \quad \forall \varepsilon > 0 \tag{4.18}$$

But as noted in ([19, p. 804]), d_k can be taken as $d_k = k^\delta$ for all $\delta \in (\frac{2}{\beta}, 1)$. Hence, from (4.18) it is easy to conclude (4.5), and hence the proof is complete. □

Let D_1 be a set defined by

$$D_1 = \left\{ w : \liminf_{n\to\infty} |b_1(n)| \neq 0 \right\}, \tag{4.19}$$

where $b_1(n)$ denotes the component of θ_n estimating b_1.

For any constant $a \in (0, |b_1|)$, define a sequence $\{\tau_n\}$ recursively by

$$\tau_n = \inf\{k > \tau_{n-1} : |b_1(k+1)| < a\}, \qquad \tau_0 = 0, \quad n \geq 1. \tag{4.20}$$

Note that (A3) implies $b_1 \neq 0$, and so the interval $(0, |b_1|)$ is not empty.

On the complement set of D_1, D_1^c, it is obvious that $\tau_n < \infty$ for all $n \geq 1$. Hence, if we set

$$\sigma_n = \begin{cases} n, & w \in D_1, \\ \tau_n, & w \in D_1^c, \end{cases} \tag{4.21}$$

then $\sigma_n < \infty$ a.s. for all n, and $\sigma_n \xrightarrow[n\to\infty]{} \infty$ a.s..

THEOREM 4.1. *For the system* (1.1) *assume that* (A1)–(A4) *are satisfied, and that the control law defined by* (4.1) *is applied. Then the following hold:*

(i) *For the sequence* $\{\sigma_n\}$ *defined by* (4.19)–(4.21), *as* $n \to \infty$

$$r_{\sigma_n} = O(\sigma_n) \text{ a.s.} \quad \text{and} \quad R_{\sigma_n+1} = O\left(\frac{d_{\sigma_n}}{\sigma_n^{1-\varepsilon}}\right) \quad \text{a.s., } \forall \varepsilon > 0,$$

where R_n, r_n *and* d_n *are defined by* (1.6), (1.18), *and* (1.22), *respectively.*

(ii)

$$R_n = O\left(\frac{1}{n^{1-\delta}}\right) \quad \text{a.s. on } D, \quad \forall \delta \in \left(\frac{2}{\beta}, 1\right),$$

where β *is defined in* (1.2) *and* $D = D_1 \cup D_2$ *with* D_1 *defined by* (4.19) *and* D_2 *defined by*

$$D_2 = \left\{ w \in D_1^c : \sup_n \frac{\tau_{n+1}}{\tau_n} < \infty \right\};$$

here $\{\tau_n\}$ *is defined by* (4.20).

Proof. (i). On the set D_1, by a completely similar argument as that used for Theorem 2 in [19], it is known that $R_n = O(d_n/n^{1-\varepsilon})$ a.s. on D_1, for all $\varepsilon > 0$. So we need only to consider the complement set D_1^c. By the definition of τ_n we have

$$\inf_n |b_1(\tau_n + 1) - b_1| \geq |b_1| - a > 0 \quad \text{on } D_1^c.$$

Therefore, by Lemma 4.1 we know that $r_{\tau_n} = O(\tau_n)$ and $\sup_{i\leq\tau_n} \|\varphi_i\|^2 = O(\tau_n^\varepsilon d_{\tau_n})$ a.s. on D_1^c. Hence, by (4.6) (with $\Delta\hat{b}_{1n} \equiv 0$) and Lemma 2.5, we have

$$\begin{aligned} R_{\tau_n+1} &= \frac{1}{\tau_n+1} \sum_{i=0}^{\tau_n} (y_{i+1} - y_{i+1}^* - w_{i+1})^2 \\ &= O\left(\frac{1}{\tau_n} \sum_{i=0}^{\tau_n} \|\tilde{\theta}_i^\tau \varphi_i\|^2\right) + O\left(\frac{1}{\tau_n} \sum_{i=0}^{\tau_n} \|\varphi_i - \varphi_i^0\|^2\right) \\ &= O\left(\frac{1}{\tau_n} \sum_{i=0}^{\tau_n} \frac{\|\tilde{\theta}_i^\tau \varphi_i\|^2}{1 + \varphi_i^\tau P_i \varphi_i} (1 + \|\varphi_i\|^2)\right) + O\left(\frac{\log r_{\tau_n}}{\tau_n}\right) \\ &= O\left(\frac{d_{\tau_n} \log \tau_n}{\tau_n^{1-\varepsilon}}\right) \quad \text{a.s.} \quad \text{on } D_1^c, \forall \varepsilon > 0. \end{aligned} \tag{4.22}$$

Hence the conclusion (i) holds.

(ii) As is just mentioned above (ii) holds on D_1, since d_n can be taken as n^δ for all $\delta \in (2/\beta, 1)$. Hence we need only to consider the set D_2. By the definition of $\{\tau_n\}$, we know that on $D_2, \tau_n \to \infty, \sup_n(\tau_{n+1}/\tau_n) < \infty$, and by (4.22), $R_{\tau_n} = O(1/\tau_n^{1-\delta})$ a.s. on D_2 for all $\delta \in (2/\beta, 1)$. Consequently,

$$\begin{aligned}\sup_n[n^{1-\delta}R_n] &= \sup_k \sup_{n\in[\tau_k,\tau_{k+1}]}[n^{1-\delta}R_n] \\ &\le \sup_k \frac{\tau_{k+1}}{\tau_k}[\tau_{k+1}^{1-\delta}R_{\tau_{k+1}}] < \infty \quad \text{a.s. on } D_2.\end{aligned}$$

Hence assertion (ii) is also true. □

Remark 4.1. From Guo and Chen [19], we know that under conditions of Theorem 4.1, if $\liminf_{n\to\infty}|b_1(n)| \neq 0$ a.s., then $R_n \to 0$ a.s. Theorem 4.1 asserts, among other things, that if $\lim_{n\to\infty}|b_1(n)| = 0$ a.s. does hold, then since $P(D_2) = 1$ and again we have $R_n \to 0$ a.s.. This result is rather interesting since $b_1(n)$ appears as the divisor in the control law (4.1), and small $b_1(n)$ seems to yield large input signal u_n (but actually does not). The key idea behind the proof of Theorem 4.1 (or Lemma 4.1) is as follows: if $\lim_{n\to\infty}|b_1(n)| = 0$, then $|\tilde{b}_1(n+1)| \ge |b_1|/2 > 0$ for all suitably large n. Thus, for each fixed large n, and for all $i \le n, u_i$ will have a significant contribution to $\tilde{\theta}_{n+1}^\tau\varphi_i$ if it is not small. But by (4.11) we know that $\|\tilde{\theta}_{n+1}^\tau\varphi_i\|^2 = O(\log\ r_n)$ a.s. for all $i \le n$. Hence, for all $i \le n, u_i^2$ will be dominated by a "linear combination" of $\{y_i^2, \cdots, y_{i-p+1}^2, u_{i-1}^2 \cdots u_{i-q+1}^2, \hat{w}_i^2 \cdots \hat{w}_{i-r+1}^2\}$, and thus we can successfully sidestep the difficult "small divisor" problem in the analysis. Certainly, in this approach, it would be of considerable interest to preclude the case where the sequence $\{b_1(n)\}$ visits the interval $(-a, a)$ with $0 < a < |b_1|$ in a very scattering way (i.e., $P(D^c) > 0$).

We now consider the case where the set D defined in Theorem 4.1 does have probability one.

THEOREM 4.2. *Consider the system* (1.1), *the* ELS *algorithm* (1.11)–(1.14), *and the control law* (4.1). *If* (A1)–(A4) *and* (A6) *hold, and in addition, the reference signal* $\{y_i^*\}$ *satisfies*

$$n^{\frac{1+\varepsilon}{2}}\sqrt{d_n} = O(\lambda_{\min}^*(n)) \quad a.s., \quad \text{for some } \varepsilon > 0, \tag{4.23}$$

where d_n *and* $\lambda_{\min}^*(n)$ *are defined in* (1.22) *and* (2.7), *respectively, then as* $n \to \infty$

$$R_n = O\left(\frac{\log\ n}{n}\right) \ a.s., \qquad \|\theta_n - \theta\|^2 = O\left(\frac{\log\ n}{\lambda_{\min}^*(n)}\right) \ a.s., \tag{4.24}$$

where R_n *is the regret defined by* (1.6). *Furthermore, if* (4.23) *is replaced by* $n = O(\lambda_{\min}^*(n))$ *a.s., and* $E[w_{n+1}^2|\mathcal{F}_n] = \sigma^2$ *a.s., then for the white noise case* ($r = 0$), (4.24) *can be strengthened into*

$$\lim_{n\to\infty}\left(\frac{n}{\log\ n}\right)R_n = (p+q)\sigma^2 \quad a.s., \qquad \|\theta_n - \theta\|^2 = O\left(\frac{\log\log\ n}{n}\right) a.s.$$

Proof. By Theorem 4.1 (i) and (4.23) we know that Theorem 2.1 (ii) is applicable to the sequence $\{\sigma_n\}$, and hence we have

$$\liminf_{n\to\infty}\frac{\lambda_{\min}\left(\sum_{i=0}^{\sigma_n}\varphi_i^0\varphi_i^{0\tau}\right)}{\lambda_{\min}^*(\sigma_n)} > 0 \quad \text{a.s.}$$

Consequently, similar to the proof of (3.25), it is easy to see that

$$\liminf_{n\to\infty} \frac{\lambda_{\min}\left(\sum_{i=0}^{\sigma_n} \varphi_i\varphi_i^\tau\right)}{\lambda^*_{\min}(\sigma_n)} > 0 \quad \text{a.s.} \tag{4.25}$$

By this and Lemma 2.5 (i) it is easy to see that

$$\|\tilde{\theta}_{\sigma_n+1}\|^2 = O\left(\frac{\log r_{\sigma_n}}{\lambda^*_{\min}(\sigma_n)}\right) \quad \text{a.s.} \tag{4.26}$$

By Theorem 4.1 (i), we know that $\log r_{\sigma_n} = O(\log \sigma_n)$ a.s., and so by (4.23) and (4.26) we conclude that $\tilde{\theta}_{\sigma_n+1} \to 0$ a.s., and in particular,

$$b_1(\sigma_n + 1) \xrightarrow[n\to\infty]{} b_1 \quad \text{a.s.} \tag{4.27}$$

We now prove that $P(D_1) = 1$ where D_1 is defined by (4.19). Otherwise, we would have $P(D_1^c) > 0$, and on D_1^c by the definition of σ_n we know that $\sigma_n < \infty$ for all n, and that

$$|b_1(\sigma_n + 1)| < a \quad \forall n \geq 1 \quad \text{on } D_1^c, \tag{4.28}$$

which clearly contradicts with (4.27) since $a < |b_1|$. Hence $P(D_1) = 1$ and we have $\liminf_{n\to\infty} |b_1(n)| \neq 0$ a.s. Therefore, by a similar means as in the proof of Theorem 2 in Guo and Chen [19], we obtain $R_n = O(d_n/n^{1-\varepsilon})$ a.s. and $\|\varphi_n\|^2 = O(n^\varepsilon d_n)$ a.s. for all $\varepsilon > 0$. Using this and a similar argument as for (4.25) and (4.26), we know that (4.25) and (4.26) also hold with $\{\sigma_n\}$ replaced by $\{n\}$. Hence we have proved the second assertion in (4.24). Since (4.25) holds with $\{\sigma_n\}$ replaced by $\{n\}$ and $\|\varphi_n\|^2 = O(n^{\delta/2}\sqrt{d_n})$ a.s., for all $\delta \in (2/\beta, 1)$, we know that $\varphi_n^\tau P_n \varphi_n \xrightarrow[n\to\infty]{} 0$ a.s. By this and Lemma 2.5 it follows from (4.6) (with $\Delta\hat{b}_{1n} \equiv 0$) that the first assertion in (4.24) is also true. Finally, the last two assertions of the theorem can be proved in exactly the same way as in Theorem 3.1, and the details are not repeated. □

Remark 4.2. (i) Again, the best possible convergence rate $O(\log n/n)$ is established for the regret R_n. It is worth noting that this result is established without introducing any modifications to the standard minimum variance control law (4.1). This fact makes Theorem 4.2 differ essentially from the existing results including those in the recent work [19].

(ii) The (nonpersistent) excitation condition (4.23) on the reference signal $\{y_n^*\}$ can be easily verified for a large class of deterministic and/or stochastic signals. In principle, we can always make this condition satisfied by use of the "continually disturbed demand method" of Caines and Lafortune [27] or the "diminishing excitation technique" in Chen and Guo [16]. To be precise, for any desired trajectory $\{y_{nd}^*\}$ that is bounded and independent of $\{w_n\}$, we may take the reference signal in (4.1) to be

$$y_n^* = y_{nd}^* + v_n, \tag{4.29}$$

where $\{v_n\}$ is a zero mean independent bounded random sequence which is independent of $\{w_n, y_{nd}^*\}$. Then with some suitable moment conditions on $\{v_n\}$ it is easy to see that (4.23) holds. In order that the "dither" does not influence the self-optimality the variance of $\{v_n\}$ must be chosen to satisfy $Ev_n^2 \xrightarrow[n\to\infty]{} 0$. This is possible since the excitation requirement

(4.23) is not necessarily persistent. The disadvantage of adding the "dither" $\{v_n\}$ in such a way is that it may influence the convergence rate of tracking.

Next, we consider the case where (4.23) fails. As a typical example, we shall only consider the regulation problem ($y_n^* \equiv 0$). Similar to (1.23), we set for $n \geq 1$,

$$\hat{b}_1(n) = \begin{cases} b_1(n) & \text{if } |b_1(n)| \geq \dfrac{1}{\sqrt{n \log (n+1)}}; \\ b_1(n) + \dfrac{\operatorname{sgn}(b_1(n))}{\sqrt{n \log(n+1)}} & \text{otherwise.} \end{cases} \tag{4.30}$$

Instead of (4.1), we define the control u_n by

$$u_n = \frac{1}{\hat{b}_1(n)}\{b_1(n)u_n - \theta_n^\tau \varphi_n\}, \qquad n \geq 1, \tag{4.31}$$

which obviously has the form of (4.2):

$$\theta_n^\tau \varphi_n + \Delta\hat{b}_{1n} u_n = 0, \tag{4.32}$$

where $\Delta\hat{b}_{1n} \triangleq \hat{b}_1(n) - b_1(n)$. By (4.30) it is clear that

$$|\hat{b}_1(n)|^2 \geq \frac{1}{n \log (n+1)}, \qquad |\Delta\hat{b}_{1n}|^2 \leq \frac{1}{n \log(n+1)} \xrightarrow[n\to\infty]{} 0. \tag{4.33}$$

THEOREM 4.3. *Consider the model* (1.1) *with white additive noise (i.e., $r = 0$). Assume that* (A1), (A3), *and* (A5) *are satisfied, and that in* (1.22), $d_n = O(n^\varepsilon)$ *for all $\varepsilon > 0$. If the control law defined by* (4.30) *and* (4.31) *is applied, then as $n \to \infty$,*

$$\sum_{i=1}^{n} (y_i - w_i)^2 = O(n^\varepsilon) \quad \text{a.s.,} \quad \forall \varepsilon > 0, \tag{4.34}$$

Proof. Let D_1 and $\{\tau_n\}$ be defined by (4.19) and (4.20), respectively. As explained earlier, (4.34) holds on D_1, and so we need only to consider D_1^c. In the remainder of the proof all relationships are established on D_1^c with a possible exception set of probability zero, and we shall omit the phrase "a.s. on D_1^c" for simplicity.

By (4.20) we know that on D_1^c, $\tau_n < \infty$, for all $n \geq 1, \lim_{n\to\infty} \tau_n = \infty$, and $\inf_n |b_1(\tau_n + 1) - b_1| \geq |b_1| - a > 0$. Hence, by Lemma 4.1 we know that

$$r_{\tau_n} = O(\tau_n) \quad \text{and} \quad \sup_{k \leq \tau_n} \|\varphi_k\|^2 = O(\tau_n^\varepsilon d_{\tau_n}). \tag{4.35}$$

Consequently, by (4.6) and (4.33),

$$\begin{aligned} \sum_{i=1}^{\tau_n} (y_{i+1} - w_{i+1})^2 &= O\left(\sum_{i=1}^{\tau_n} \|\tilde{\theta}_i^\tau \varphi_i\|^2\right) + O\left(\sum_{i=1}^{\tau_n} (\Delta\hat{b}_{1i} u_i)^2\right) \\ &= O\left(\tau_n^\varepsilon d_{\tau_n} \sum_{i=1}^{\tau_n} \frac{\|\tilde{\theta}_i^\tau \varphi_i\|^2}{1 + \varphi_i^\tau P_i \varphi_i}\right) + O\left(\sum_{i=1}^{\tau_n} \frac{1}{i \log (i+1)} u_i^2\right) \\ &= O(\tau_n^\varepsilon d_{\tau_n} \log \tau_n) + O\left(\tau_n^\varepsilon d_{\tau_n} \sum_{i=1}^{\tau_n} \frac{1}{i \log (i+1)}\right) \quad \forall \varepsilon > 0, \\ &= O(\tau_n^\varepsilon) \quad \forall \varepsilon > 0. \end{aligned} \tag{4.36}$$

Hence Theorem 2.1 (i) is applicable, and we then have

$$\liminf_{n\to\infty} \lambda_{\min}\left(\frac{1}{\tau_n}\sum_{i=1}^{\tau_n}\psi_i\psi_i^\tau\right) > 0, \tag{4.37}$$

where $\{\psi_i\}$ is defined by (2.4) with $p^* = p$.

Let $a_i(n), b_j(n)$ be the estimates for a_i, b_j, $1 \le i \le p$, $1 \le j \le q$, given by θ_n. Set

$$\bar{\theta}_n = [-a_1(n), \cdots, -a_p(n), b_2(n), \cdots, b_q(n)]^\tau. \tag{4.38}$$

Then by (4.31) we have

$$u_n = -\frac{1}{\hat{b}_1(n)}\bar{\theta}_n^\tau\psi_n \tag{4.39}$$

Now, we prove that

$$\left\|\frac{1}{\hat{b}_1(\tau_n+1)}\bar{\theta}_{\tau_n+1}\right\|^2 = 0(\tau_n^\varepsilon) \quad \forall \varepsilon > 0. \tag{4.40}$$

By (4.36) we know that $\sum_{i=1}^{\tau_n}\|\theta^\tau\varphi_i\|^2 = O(\tau_n^\varepsilon)$, which in conjunction with (4.10) and (4.35) gives

$$\sum_{i=1}^{\tau_n}(\theta_{\tau_n+1}^\tau\varphi_i)^2 \le 2\sum_{i=1}^{\tau_n}(\tilde{\theta}_{\tau_n+1}^\tau\varphi_i)^2 + 2\sum_{i=1}^{\tau_n}(\theta^\tau\varphi_i)^2 = O(\tau_n^\varepsilon).$$

From this, by noting that $\theta_{\tau_n+1}^\tau\varphi_i = \bar{\theta}_{\tau_n+1}^\tau\psi_i + b_1(\tau_n+1)u_i$, we have for all $\varepsilon > 0$,

$$\begin{aligned}\sum_{i=1}^{\tau_n}(\bar{\theta}_{\tau_n+1}^\tau\psi_i)^2 &\le 2\sum_{i=1}^{\tau_n}(\theta_{\tau_n+1}^\tau\varphi_i)^2 + 2b_1^2(\tau_n+1)\sum_{i=1}^{\tau_n}u_i^2 \\ &\le O(\tau_n^\varepsilon) + 4\{[\hat{b}_1(\tau_n+1)]^2 + [\Delta\hat{b}_1(\tau_n+1)]^2\}\sum_{i=1}^{\tau_n}u_i^2.\end{aligned}$$

Multiplying $1/[\hat{b}_1(\tau_n+1)]^2$ from both sides of this inequality, and noticing (4.33) and (4.35) we get

$$\frac{1}{[\hat{b}_1(\tau_n+1)]^2}\sum_{i=1}^{\tau_n}(\bar{\theta}_{\tau_n+1}^\tau\psi_i)^2 = O\left(\frac{\tau_n^\varepsilon}{[\hat{b}_1(\tau_n+1)]^2}\right) + O\left(\sum_{i=1}^{\tau_n}u_i^2\right) = O(\tau_n^{1+\varepsilon}\log(\tau_n+1)+\tau_n).$$

From this and (4.37) we see that for all suitably large n,

$$\begin{aligned}\left\|\frac{1}{\hat{b}_1(\tau_n+1)}\bar{\theta}_{\tau_n+1}\right\|^2 &\le O\left(\left\|\frac{1}{\hat{b}_1(\tau_n+1)}\bar{\theta}_{\tau_n+1}\right\|^2\lambda_{\min}\left(\frac{1}{\tau_n}\sum_{i=1}^{\tau_n}\psi_i\psi_i^\tau\right)\right) \\ &\le O\left(\frac{1}{\tau_n[\hat{b}_1(\tau_n+1)]^2}\sum_{i=1}^{\tau_n}(\bar{\theta}_{\tau_n+1}^\tau\psi_i)^2\right) = O(\tau_n^\varepsilon\log(\tau_n+1)+1) \quad \forall\varepsilon > 0.\end{aligned}$$

Hence (4.40) holds.

Next, we prove that

$$\|\varphi_n\|^2 = O([nr_n]^\varepsilon d_n) \quad \forall \varepsilon > 0. \tag{4.41}$$

Note that by (4.20), we know that on D_1^c,

$$|b_1(k+1)| \geq a > 0, \quad \forall k \in [\tau_n + 1, \tau_{n+1} - 1], \quad \forall n \geq 1. \tag{4.42}$$

Hence from (4.39), (4.40), (4.42), and the fact that $\|\bar{\theta}_{k+1}\|^2 = O(\log r_k)$, it follows that on D_1^c

$$u_{k+1}^2 = \begin{cases} O(\tau_n^\varepsilon \|\psi_{\tau_n+1}\|^2), & \forall \varepsilon > 0, \quad k = \tau_n; \\ O((\log r_k)\|\psi_{k+1}\|^2), & k \in [\tau_n + 1, \tau_{n+1} - 1]. \end{cases} \tag{4.43}$$

Similar to (4.9) it is easy to see by (A3) that

$$\|\psi_k\|^2 = O\left(\sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + O(d_k) \quad \forall k \geq 1.$$

From this and (4.43) we have, for all $k \in [\tau_n + 2, \tau_{n+1}]$,

$$\begin{aligned} \|\varphi_k\|^2 &= \|\psi_k\|^2 + u_k^2 \\ &= O\left([\log r_k] \sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + O(d_k \log r_k). \end{aligned} \tag{4.44}$$

Substituting this together with (4.8) into (4.7) and noting that $\Delta \hat{b}_{1n} \xrightarrow[n\to\infty]{} 0$, we get for all $k \in [\tau_n + 2, \tau_{n+1}]$, and all suitably large n,

$$y_{k+1}^2 = O\left(\alpha_k \delta_k (\log r_k) \sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + o\left(\sum_{i=0}^{k} \lambda^{k-i} y_i^2\right) + O(d_k \log^2 r_k). \tag{4.45}$$

Set $L_k = \sum_{i=0}^{k} \lambda^{k-i} y_i^2$. Similar to the proof of (4.15), from (4.45) we have for some $\gamma \in (0, 1), \forall k \in [\tau_n + 2, \tau_{n+1}]$, and all large n,

$$\begin{aligned} L_{k+1} \leq{}& \gamma^{k-\tau_n-1} \prod_{i=\tau_n+2}^{k} (1 + c\alpha_i \delta_i \log r_i) L_{\tau_n+2} \\ &+ O\left(\sum_{i=\tau_n+2}^{k} \gamma^{k-i} \prod_{j=i+1}^{k} (1 + c\alpha_j \delta_j \log r_j) d_i \log^2 r_i\right), \end{aligned} \tag{4.46}$$

where $c > 0$ is a constant. Similar to the proof of (4.16) we know that for all small $\varepsilon > 0$ and all $k \geq i$, with i suitably large, $\prod_{j=i+1}^{k} (1 + c\alpha_j \delta_j \log r_j) \leq r_k^\varepsilon$. Substituting this into (4.46) yields for large n,

$$L_{k+1} = O(r_k^\varepsilon L_{\tau_n+2}) + O(r_k^\varepsilon d_k), \quad \forall k \in [\tau_n + 2, \tau_{n+1}], \quad \forall \varepsilon > 0. \tag{4.47}$$

By (4.35), (1.1), and (A3), it is easy to see that

$$L_{\tau_n+1} + \|\psi_{\tau_n+1}\|^2 = O(\tau_n^\varepsilon d_{\tau_n}), \quad \forall \varepsilon > 0. \tag{4.48}$$

Consequently by (4.43), $u^2_{\tau_n+1} = O(\tau_n^{\varepsilon} d_{\tau_n})$, for all $\varepsilon > 0$. From this, (1.1), and (4.35) again, we obtain $L_{\tau_n+2} = O(\tau_n^{\varepsilon} d_{\tau_n})$ for all $\varepsilon > 0$. This in conjunction with (4.47) and (4.48) yields

$$L_{k+1} = O([kr_k]^{\varepsilon} d_k), \quad \forall k \in [\tau_n, \tau_{n+1}], \quad \forall \varepsilon > 0. \tag{4.49}$$

From this it is easy to convince oneself that

$$\|\varphi_k\|^2 = O([kr_k]^{\varepsilon} d_k), \quad \forall k \in [\tau_n, \tau_{n+1}], \quad \forall \varepsilon > 0,$$

holds for all suitably large n. This implies (4.41), since $\tau_n \xrightarrow[n\to\infty]{} \infty$.

By (4.41) and a similar proof as for (4.18) we get $r_n = O([nr_n]^{\varepsilon} d_n \log r_n) + o(r_n) + O(n)$ for all $\varepsilon > 0$. Hence it follows that $r_n = O(n)$. Then, by (4.41) and the assumption that $d_n = O(n^{\varepsilon})$ for all $\varepsilon > 0$, we obtain $\|\varphi_n\|^2 = O(n^{\varepsilon})$ for all $\varepsilon > 0$. Therefore, similar to the proof of (4.36), we get $\sum_{i=1}^{n} (y_{i+1} - w_{i+1})^2 = O(n^{\varepsilon})$ for all $\varepsilon > 0$ a.s. on D_1^c. This completes the proof. □

Remark 4.3. The advantage of the modification (4.30) over (1.23) as used in [19] is clear. When (1.23) is used, the cumulated square errors resulting from the modification of $b_1(n)$ is of the order $O(n/\log n)$, i.e.,

$$\sum_{i=1}^{n} (\Delta \hat{b}_{1i})^2 = O\left(\sum_{i=1}^{n} \frac{1}{\log r_i}\right) = O\left(\frac{n}{\log n}\right).$$

Hence in Theorem 2 of Guo and Chen [19], the guaranteed convergence rate for the averaged regret R_n is only of the order $O(1/\log n)$, which is clearly much slower than the rate $R_n = O(1/n^{1-\varepsilon})$ a.s. for all $\varepsilon > 0$, obtained in Theorem 4.3. Of course, it would be of interest to generalize Theorem 4.3 to the colored noise case and to show that the left-hand side of (4.34) is of the order $O(\log n)$.

5. Concluding remarks. The convergence rate of least-squares–based adaptive algorithm has been observed in practice to be superior to any other type of implementable on-line recursive algorithms including the extensively studied stochastic gradient algorithm. In this paper, we have obtained various new results on the standard ELS-based adaptive minimum variance control for SISO ARMAX systems, and improved on the recent work [19] in many aspects. In particular, we have obtained the best possible convergence rate $O(\log n/n)$ for the averaged regret of tracking in several situations of interest. This rate is not believed to be achievable, for example, for the stochastic gradient based adaptive algorithm. For further study, it is desirable to generalize the result $R_n = O(\log n/n)$ to general tracking problems with arbitrarily bounded reference signal $\{y_n^*\}$, using (preferably) the control law (4.1).

REFERENCES

[1] D. W. Huang and L. Guo, *Estimation of nonstationary* ARMAX *models based on the Hannan–Rissanen method,* Ann. Statist., 18 (1990), pp. 1729–1756.

[2] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control,* Prentice-Hall, Englewood Cliffs, NJ, 1986.

[3] Y. S. Chow, *Local convergence of martingales and the law of large numbers,* Ann. Math. Statist., 36 (1965), pp. 552–558.

[4] C. Z. Wei, *Adaptive prediction by least squares predictors in stochastic regression models with applications to time series,* Ann. Statist., 15 (1987), pp. 1667–1682.

[5] P. E. Caines, *Linear Stochastic Systems,* John Wiley, New York, 1988.

[6] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control,* Birkhäuser, Boston, 1991.
[7] J. Sternby, *On consistency of least squares method using martingale theory,* IEEE Trans. Automat. Control, 22 (1977), pp. 346–352.
[8] T. L. Lai and C. Z. Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems,* Ann. Statist., 10 (1982), pp. 154–166.
[9] G. Ledwich and J. B. Moore, *Multivariable self-tuning filters,* in Differential Games and Control Theory, II. Lecture Notes in Pure and Appl. Math., Marcel-Dekker, New York, 1977, pp. 345–376.
[10] V. Solo, *The convergence of* AML, IEEE Trans. Automat. Control, 24 (1979), pp. 958–962.
[11] H. F. Chen, *Strong consistency and convergence rate of least squares identification,* Scientia Sinica (Ser. A), 25 (1982), pp. 771–784.
[12] H. F. Chen and L. Guo, *Convergence rate of least-squares identification and adaptive control for stochastic systems,* Internat. J. Control, 44 (1986), pp. 1459–1476.
[13] T. L. Lai and C. Z. Wei, *Extended least squares and their applications to adaptive control and prediction in linear systems,* IEEE Trans. Automat. Control, 31 (1986), pp. 898–906.
[14] G. C. Goodwin, P. Ramadge and P. E. Caines, *Discrete time stochastic adaptive control,* SIAM J. Control Optimiz., 19 (1981), pp. 829–853.
[15] H. F. Chen and L. Guo, *Strong consistency of recursive identification by no use of persistent excitation condition,* Acta. Math. Appl. Sinica, 2 (1985), pp. 133–145.
[16] ———, *Asymptotically optimal adaptive control with consistent parameter estimates,* SIAM J. Control Optimiz., 25 (1987), pp. 558–575.
[17] K. J. Åström and B. Wittenmark, *On self-tuning regulators,* Automatica, 9 (1973), pp. 195–199.
[18] P. R. Kumar, *Convergence of adaptive control schemes using least-squares parameter estimates,* IEEE Trans. Automat. Control, 35 (1990), pp. 416–423.
[19] L. Guo and H. F. Chen, *The Åström–Wittenmark self-tuning regulator revisited and* ELS*-based adaptive trackers,* IEEE Trans. Automat. Control, 36 (1991), pp. 802–812.
[20] P. R. Kumar, *A survey of some recent results in stochastic adaptive control,* SIAM J. Control Optimiz., 23 (1985), pp. 329–380.
[21] N. C. Jain, K. Jogdeo and W. F. Stout, *Upper and lower functions for martingales and mixing processes,* Ann. Probability, 3 (1975), pp. 119–145.
[22] H. F. Chen and L. Guo, *Strong consistency of parameter estimates in optimal stochastic adaptive tracking systems,* Sci. Sinica (Ser. A), 29 (1986), pp. 1145–1156.
[23] T. L. Lai, *Asymptotically efficient adaptive control in stochastic regression models,* Adv. Appl. Math., 7 (1986), pp. 23–45.
[24] A. Becker, P. R. Kumar and C. Z. Wei, *Adaptive control with the stochastic approximation algorithm: Geometry and convergence,* IEEE Trans. Automat. Control. 30 (1985), pp. 154–166.
[25] P. R. Kumar and L. Praly, *Self-tuning trackers,* SIAM J. Control Optimiz., 25 (1987), pp. 1053–1071.
[26] S. P. Meyn and P. E. Caines, *The zero divisor problem of multivariable stochastic adaptive control,* System Control Lett., 6 (1985), pp. 235–238.
[27] P. E. Caines and S. La Fortune, *Adaptive control with recursive identification for stochastic linear systems,* IEEE Trans. Automat. Control, 29 (1984), pp. 312–321.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 213–223, January 1994

013

TOPOLOGICAL METHODS FOR THE LOCAL CONTROLLABILITY OF NONLINEAR SYSTEMS*

W. KRYSZEWSKI† AND S. PLASKACZ‡

Abstract. Local controllability of systems using the topological degree of finite-dimensional set-valued maps is studied. For perturbed linear systems a generalization of the Lee–Markus sufficient condition of local controllability is established. For systems given by a finite family of continuous vector fields first order controllability condition is obtained. In both cases, a set of control functions sufficient for the local controllability is described, which is homeomorphic with a d-dimensional unit ball, where d is the dimension of the state space.

Key words. small-time local controllability, topological degree, perturbed linear systems, vector fields

AMS subject classifications. 93B05, 93C10

1. Introduction.

The purpose of this paper is to study local controllability of control systems

$$\begin{aligned} &x' = f(t,x,u);\\ &x(0) = 0,\\ &u(\cdot) \in U, \end{aligned} \tag{1}$$

where U is a family of control functions from [0,1] to a given set Ω.

For a given control function $u(\cdot) \in U$, by $\mathrm{Sol}_f(u)$ we denote the set of all solutions to the Cauchy problem

$$\begin{aligned} &x' = f(t,x,u(t));\\ &x(0) = 0. \end{aligned}$$

Let $T \in [0,1]$ and $U_0 \subset U$. The reachable set of system (1) controlled by U_0 at the time T (respectively, in time $\leq T$) is defined by

$$R_f(T,U_0) = \{x(T) : x \in \mathrm{Sol}_f(u) \text{ and } u \in U_0\}$$

$$\left(\text{resp. } R_f(\leq T,U_0) = \bigcup_{t\in[0,T]} R_f(t,U_0)\right).$$

We say that the system (1) is small time locally controllable (STLC) by means of U_0 if for every $T \in (0,1]$ we have

$$0 \in \mathrm{int}\, R_f(\leq T,U_0). \tag{2}$$

In the present paper we introduce an approach to the controllability of systems based on the topological degree of finite-dimensional set-valued maps used in order to establish (2). The general framework of our attitude looks as follows. First, we choose a subset K of U which is homeomorphic with the d-dimensional unit ball—say $p : B^d \to K$ is a homeomorphism; second, we deform linearly the right-hand side f of (1) to a more regular

* Received by the editors January 27, 1992; accepted for publication (in revised form October 20, 1992.

† Institute of Mathematics, University of Lódź, ul.Banacha 22, 90-238 Lódź, Poland. This author's research was partially supported by the Alexander von Humbold Foundation.

‡ Institute of Mathematics, Nicholas Copernicus University, ul.Chopina 12/18, 87-100 Toruń, Poland. This paper was partially prepared while the second author was visiting Centre de la Recherche de Mathématiques de la Décision, due to kind invitation of Jean-Pierre Aubin.

function g, i.e., we consider a family $h_\lambda = (1-\lambda)f + \lambda^g, \lambda \in [0,1]$ and study a set-valued homotopy $\mathcal{H} : [0,1] \times K \rightsquigarrow R^d$ given by the formula

$$\mathcal{H}(\lambda, u) = \{x(l(u)) : x \in \text{Sol}_{h_\lambda}(u)\},$$

where $l : K \to [0,T]$ is a suitably choosen continuous function. For perturbed linear systems we take l as a constant function $l \equiv T$, while for systems defined by family of vector fields, l is somewhat hidden behind a certain reformulation of the system. The function g and the set K are chosen as to satisfy the following requirements:

- $\mathcal{H}(1,\cdot)$ is a homeomorphism (onto the image);
- $\mathcal{H}$ is an admissible map (in the sense of [9]; see the next section);
- $0 \notin \mathcal{H}(\lambda, u)$ for $\lambda \in [0,1]$ and $u \in \partial K$ (more precisely : $u \in p(\partial B^d)$).

In view of the properties of the topological degree, we obtain that $0 \in \text{int } \mathcal{H}(0,K)$, which implies $0 \in \text{int } R_f(\leq T, K)$.

We obtain two types of results: (i) the description of the set of controls which allow to reach in time $\leq T$ points from a neighbourhood of 0 (i.e., the set K above); (ii) some extensions of known sufficient conditions of STLC to the case of less regular (i.e., continuous instead of Lipschitz continuous or smooth) systems.

The paper is organized as follows. In §2 we briefly recall the topological degree theory of set-valued maps and provide some auxiliary lemmas. In §3 we study perturbed linear systems and establish a generalization of the classical Lee–Markus sufficient condition of local controllability. Our technics are similar to those used in [7] and [15] to study global controllability problems, In §4 we consider control systems given by a finite family of continuous vector fields $f_i : [0,1] \times R^d \to R^d, i = 1, \ldots, k$, which satisfy the following first-order controllability condition:

$$0 \in \text{int } (\text{co}\{f_i(0,0) : i = 1, \ldots, k\}),$$

where co A, int A denote the closed convex hall and the interior of a subset $A \subset R^d$, respectively. For more regular, i.e., lipschitz continuous with respect to the second variable, vector fields Frankowska proved in [6] that the above condition is sufficient for STLC.

The Appendix gives the proof of the auxiliary Lemma 4.1.

2. Preliminaries. We use the standard notation, in particular by $\langle \cdot, \cdot \rangle, |\cdot|, B^d$, and S^{d-1} we denote the scalar product, the norm, the closed unit ball and unit sphere in R^d. For $\varepsilon > 0$, the ε-neighbourhood of a set $A \subset R^d$ is denoted by $N_\varepsilon(A) = \{x \in R^d : \text{dist}(x,A) = \inf_{y \in A}|x-y| < \varepsilon\}$. Moreover, we put $I = [0,1]$. We denote by $\chi_A : I \to \{0,1\}$ the characteristic function of a subset $A \subset I$. The norms in the spaces $L^1(I,R^d), L^\infty(I,R^d), C(I,R^d)$ and $AC(I,R^d)$ of integrable, essentially bounded, continuous, and absolutely continuous functions are denoted by $\|\cdot\|_1, \|\cdot\|_\infty, \|\cdot\|$, and $\|\cdot\|_{AC}$, respectively (recall $\|x\|_{AC} = \|x'\|_1 + \|x\|$).

We say that a metric space K is acyclic if $H^*(K) = H^*(pt)$, where H^* denote the Alexander–Spanier cohomology functor (see [16]) and pt is a one point space. The continuity of the theory H^* implies that any R_δ-set K (i.e., $K = \cap_1^\infty K_i, K_{i+1} \subset K_i, K_i$ is compact and contractible—see [11]), is compact and acyclic.

A set valued map φ from the set X to a set Y will be denoted by $\varphi : X \rightsquigarrow Y$, while the ordinary arrow $\to$ refers to single-valued maps. If X, Y are metric spaces, we say that φ is acyclic whenever φ is upper semicontinuous and, for each $x \in X$, the set $\varphi(x)$ is compact acyclic. In particular an upper semicontinuous map with R_δ values is acyclic.

A map $\varphi : X \rightsquigarrow Y$ is admissible if there is a metric space Γ and continuous maps $\alpha : \Gamma \to X, \beta : \Gamma \to Y$ such that

(i) α is a proper surjection (i.e., $\alpha^{-1}(K)$ is compact for any compact $K \subset X$);
(ii) $\alpha^{-1}(x)$ is acyclic for every $x \in X$;
(iii) $\varphi(x) = \beta(\alpha^{-1}(x))$ for every $x \in X$.

The following facts are readily proven:

- an admissible map is upper semicontinuous;
- an acyclic map is admissible;
- the composition of admissible maps is again admissible.

By $\mathcal{A}$ we denote the set of all admissible maps $\varphi : B^d \rightsquigarrow R^d$ such that $0 \notin \varphi(S^{d-1})$.

THEOREM 2.1 (see [2]). *There is a set-valued map* $\mathrm{Deg} : \mathcal{A} \rightsquigarrow Z$ *such that for* $\varphi \in \mathcal{A}$:

(i) *If* φ *is a single-valued map, then* $\mathrm{Deg}\ \varphi = \{deg\ \varphi\}$, *where deg stands for the ordinary Brouwer degree (see* [5], [14]). *In particular, if* φ *is a homeomorphism (onto the image), then* $\mathrm{Deg}\ \varphi = \{1\}$ *or* $\{-1\}$.
(ii) (*Existence*) *If* $\mathrm{Deg}\ \varphi \neq \{0\}$, *then there exists a neighbourhood* $\mathcal{O}$ *of zero such that* $\mathcal{O} \subset \varphi(B^d)$.
(iii) (*Homotopy invariance*) *Let* $\mathcal{H} : I \times B^d \rightsquigarrow R^d$ *be an admissible map and* $0 \notin \mathcal{H}(I \times S^{d-1})$. *Then* $\mathcal{H}(i, \cdot) \in \mathcal{A}, i = 0, 1$ *and* $\mathrm{Deg}\ \mathcal{H}(0, \cdot) \cap \mathrm{Deg}\ \mathcal{H}(1, \cdot) \neq \emptyset$.

For more details and facts concerning admissible maps and degree theory we recommend [2] and [9].

We say that a function $f : I \times X \to R^d$ (X is a metric space) is measurable-continuous if:

- $t \to f(t, x)$ is measurable for every $x \in X$;
- $x \to f(t, x)$ is continuous for almost all $t \in I$.

In what follows, the following lemmas play an important role.

LEMMA 2.2 (see [1] and [4]). *If* $h : I \times R^d \to R^d$ *is a measurable-continuous function and* $|h(t, x)| \leq \mu(t)(1 + |x|)$, *where* $\mu \in L^1$, *then the set of solutions to the Cauchy problem*

$$\begin{cases} x'(t) = h(t, x(t)), \\ x(0) = 0 \end{cases}$$

is an R_δ*-set in* $C(I, R^d)$.

LEMMA 2.3. *If* X *is a closed convex and bounded subset of a normed space* E, V *is a metric space,* $F : V \times X \to E$ *is a continuous map and* $F(V \times X)$ *is relatively compact subset of* X, *then the map* $\mathcal{F} : V \rightsquigarrow E$ *given by the formula* $\mathcal{F}(v) = Fix\ F(v, \cdot) := \{x \in X : x = F(v, x)\}$, *for* $v \in V$, *is upper semicontinuous.*

Proof. In view of the Schauder fixed point theorem, the set $\mathcal{F}(v)$ is nonempty and compact for any $v \in V$. Take a closed set $K \subset E$ and a sequence $(v_n) \subset \mathcal{F}^{-1}(K)$ converging to $v \in V$. By the definition, there is a sequence $(x_n) \subset K$ such that $x_n = F(v_n, x_n)$ for any n. In view of the compactness of $\overline{F(V \times X)}$, passing to subsequences if necessary, we may assume that $x_n \to x \in K$ as $n \to \infty$. Therefore, $x = F(v, x)$, and hence $v \in \mathcal{F}^{-1}(K)$. □

3. Perturbed linear control systems. We shall study perturbed linear control systems of the form

$$\begin{aligned} &x' = A(t)x + B(t)u + f_1(t, x, u), \\ &x(0) = 0, \\ &u \in L^\infty(I, R^m), \end{aligned} \tag{3}$$

where

(4) $A(\cdot), B(\cdot)$ are time dependent $d \times d$ and $d \times m$, respectively, matrices with measurable coefficients and there exists an integrable function $\mu : I \to R$ such that
$|A(t)x + B(t)u| \leq \mu(t)(|x| + |u|)$ for $t \in I, x \in R^d, u \in R^m$;

(5) $$f_1 : I \times R^{d+m} \to R^d \text{ is measurable-continuous;}$$

(6) $$\lim_{x,u \to 0} \frac{f_1(t,x,u)}{|x| + |u|} = 0, \text{ uniformly for } t \in I.$$

We will reduce the problem of controllability of the system (3) to the one of controllability of its linearization

(7) $$\begin{aligned} &x' = A(t)x + B(t)u, \\ &x(0) = 0, \\ &u \in L^\infty(I, R^m). \end{aligned}$$

Let denote the right-hand side of the systems (3) and (7) by f and g, respectively.

It is well known that, for any $T \in (0,1]$, the map $\Psi_T : L^\infty(I, R^m) \to R^d; \Psi_T(u) = x(T)$, where $x \in \mathrm{Sol}_g(u)$, is linear and continuous (single-valued). For any linear subspace U of $L^\infty([0,T], R^m)$ the following conditions are equivalent:

- $0 \in \operatorname{int} R_g(T,U)$.
- $R_g(T,U) = R^d$.

Fix $T \in (0,1]$ and a linear subspace V of $L^\infty([0,T], R^m)$. In connection with perturbed linear systems of the form (3) some authors (cf. [3], [8]) consider the following operator:

$$\begin{aligned} &\Phi : AC([0,T], R^d) \times V \to L^1([0,T], R^d) \times R^{d+d}; \\ &\Phi(x,u) = (x' - A(\cdot)x - B(\cdot)u, x(0), x(T)). \end{aligned}$$

Using Banach Inverse Mapping Theorem we easily prove the following lemma.

LEMMA 3.1. *If V is an d-dimensional linear subspace of $L^\infty([0,T], R^m)$ and $\Psi(V) = R^d$, then the operator Φ is a topological isomorphism. Consequently, there is a constant $c > 0$ such that:*

$$\|x\|_{AC} + \|u\|_\infty \leq c \left(\int_0^T |x'(s) - A(s)x(s) - B(s)u(s)|ds + |x(0)| + |x(T)| \right)$$

for $x \in AC([0,T], R^d), u \in V$.

LEMMA 3.2. *Suppose that functions $h_i : I \times R^{d+m} \to R^d, i = 0, 1$, are measurable-continuous and satisfy*

$$|h_i(t,x,u)| \leq \mu(t)(1 + |x| + |u|),$$

for all $x \in R^d, u \in R^m$ and almost all $t \in I$, where $\mu \in L^1(I, R)$. Then the set valued map $S : I \times L^\infty(I, R^m) \rightsquigarrow C(I, R^d)$ which associate to any $(\lambda, u) \in I \times L^\infty(I, R^m)$ the set of solutions to the following Cauchy problem

(8) $$\begin{aligned} &x'(t) = \lambda h_1(t, x(t), u(t)) + (1 - \lambda)h_0(t, x(t), u(t)), \\ &x(0) = 0 \end{aligned}$$

is admissible.

Proof. By Lemma 2.2, the set $S(\lambda, u)$ is an R_δ-set for every $\lambda \in I$ and $u \in L^\infty$.

We shall show that the map S is upper semicontinuous on any set $I \times B_R$, where $B_R = \{u \in L^\infty : \|u\|_\infty \leq R\}$. By the Gronwall inequality (see [10, p. 36]) there is a constant r such that $\|x\| < r$ for any $x \in S(\lambda, u), \lambda \in I, u \in B_R$. We set $c = (1 + R + r)\int_0^1 \mu(t)dt, X = \{x \in C(I, R^d) : \|x\| \leq c\}$. We define an integral operator $F : I \times B_R \times X \to C(I, R^d)$

$$F(\lambda, u, x)(t) = \int_0^T \tilde{h}_\lambda(s, x(s), u(s))ds,$$

where

$$\tilde{h}_i(t, x, u) = \begin{cases} h_i(t, x, u) & \text{for } t \in I, |x| \leq r, u \in R^m; \\ h_i(t, rx/|x|, u) & \text{for } t \in I, |x| > r, u \in R^m, \end{cases}$$

$i = 0, 1$ and $\tilde{h}_\lambda = \lambda\tilde{h}_1 + (1 - \lambda)\tilde{h}_0$.

It is easy to check that $F(\lambda, x, u) \in X$ for any $\lambda \in I, u \in B_R, x \in X$ and, moreover, that $Fix\ F(\lambda, u, \cdot) = S(\lambda, u)$. By Lemma 2.3, the map S is upper semicontinuous and, therefore, admissible. □

THEOREM 3.3. *Suppose that a perturbed linear control system* (3) *satisfies* (4), (5), (6), *and V is a linear subspace of $L^\infty(I, R^m)$. Then there is $T_0 \in (0, 1]$ such that for any $T \in (0, T_0)$ if*

$$0 \in \text{int } R_g(T, V), \tag{9}$$

then, for each $\varepsilon > 0$

$$0 \in \text{int } R_f(T, \{u \in V : \|u\chi_{[0,T]}\|_\infty < \varepsilon\}).$$

Proof. By (6), there exists $\delta > 0$ such that

$$|f_1(t, x, u)| < |x| + |u| \text{ for } t \in I, |x| \leq \delta, |u| \leq \delta. \tag{10}$$

We set

$$f_2(t, x, y) = \begin{cases} f_1(t, x, y) & \text{for } t \in I, |x| \leq \delta, |y| \leq \delta, \\ f_1(t, \delta x/|x|, y) & \text{for } t \in I, |x| > \delta, |y| \leq \delta, \\ f_1(t, x, \delta y/|y|) & \text{for } t \in I, |x| \leq \delta, |y| > \delta, \\ f_1(t, \delta x/|x|, \delta y/|y|) & \text{for } t \in I, |x| > \delta, |y| > \delta. \end{cases}$$

By (4), (10), and the definition of f_2, we get that

$$|h_\lambda(t, x, u)| < (\mu(t) + 1)(|x| + |u|) \quad \text{for } t \in I, x \in R^d, u \in R^m, \tag{11}$$

where $h_\lambda(t, x, u) = A(t)x + B(t)u + \lambda f_2(t, x, u)$.

We take $T_0 \in (0, 1]$ such that

$$\int_0^{T_0} (\mu(t) + 1)2\delta\ dt < \delta \tag{12}$$

and fix $T \in (0, T_0]$.

For $\lambda \in I, u \in V$, let $S_T(\lambda, u) = \{x|_{[0,T]} : x \in \text{Sol}_{h_\lambda}(u)$ and $V_T = \{u|_{[0,T]} : u \in V\}$. Evidently S_T can be treated as a map $S_T : I \times V_T \rightsquigarrow C([0, T], R^d)$. By Lemma 3.2, the map S_T is admissible. From now on till the end of this proof all elements $u \in V_T$ and solutions $x \in S_T(\lambda, u)$ are considered on $[0, T]$ and the norms $\|u\|_\infty$ and $\|x\|$ are taken from respective spaces $L^\infty([0, T], R^m)$ and $C([0, T], R^d)$.

By the Gronwall inequality, there is a constant $c_0 > 0$ such that

$$\|x\| \leq c_0 \|u\|_\infty \tag{13}$$

for any $u \in V_T, \lambda \in I$ and $x \in S_T(\lambda, u)$.

It is easy to deduce from (11) and (12) that, if $x \in \mathrm{Sol}_{h_1}(u)$ and $\|u\|_\infty \leq \delta$, then $\|x\| \leq \delta$. By the same arguments, we get that $\|x\| \leq \delta$ for any $x \in \mathrm{Sol}_f(u)$ if $u \in V_T$ and $\|u\|_\infty \leq \delta$. Thus

$$R_f(T, W) = R_{h_1}(T, W) \tag{14}$$

for any subset $W \subset \{u \in V_T : \|u\|_\infty \leq \delta\}$.

By (9), there is a d-dimensional subspace V_0 of V_T (i.e., $V_0 \subset L^\infty([0,T], R^m)$) such that $\Psi(V_0) = R^d$. In view of Lemma 3.1, there is a constant $c > 0$ such that

$$\|x\| + \|u\|_\infty \leq c \left(\int_0^T |f_2(s, x(s), u(s))| ds + |x(T)| \right), \tag{15}$$

for any $x \in S_T(\lambda, u), u \in V_0$, and $\lambda \in I$. Once again, by (6), there is $\delta_1 (0 < \delta_1 < \delta)$ such that

$$|f_2(t, x, u)| \leq \frac{|x| + |u|}{2Tc}, \tag{16}$$

for $t \in I, |x| < \delta_1, |u| < \delta_1$. Let $\delta_2 > 0$ be such that $\delta_2 < \min(\delta_1, \varepsilon), c_0 \delta_2 < \delta_1$ and take $K \subset V_0$ be the closed ball of radius δ_2 centered at zero.

A homotopy $\mathcal{H} : I \times K \rightsquigarrow R^d$ given by the formula

$$\mathcal{H}(\lambda, u) = \{x(T) : x \in S_T(\lambda, u)\}$$

for $\lambda \in I, u \in K$, is an admissible map since it is the composition of the admissible map S_T with the continuous (single-valued) evaluation $e_T : C([0,T], R^d) \to R^d; e_T(x) = x(T)$.

We shall show that if $u \in K, \|u\|_\infty = \delta_2$ then $0 \notin \mathcal{H}(\lambda, u)$ for any $\lambda \in I$. Suppose to the contrary that for some $u \in K, \|u\|_\infty = \delta_2$ and $\lambda \in I$, there is $x \in S_T(\lambda, u)$ such that $x(T) = 0$. In view of (13), $\|x\| \leq c_0 \|u\|_\infty \leq c_0 \delta_2 < \delta_1$. By (15), (16)

$$\|x\| + \|u\|_\infty \leq c \int_0^T |f_2(s, x(s), u(s))| ds \leq 1/2(\|x\| + \|u\|_\infty),$$

a contradiction. By Theorem 2.1(i), (iii), $0 \neq \mathrm{Deg}\ \mathcal{H}(\cdot, 0) \in \mathrm{Deg}\ \mathcal{H}(\cdot, 1)$, hence, by Theorem 2.1(ii), $0 \in \mathrm{int}\ \mathcal{H}(1, K) = R_{h_1}(T, K)$. Equality (14) ends the proof. □

LEMMA 3.4. *Suppose that a linear control system* (7) *satisfies* (4). *Then the system* (7) *is* STLC *by* $L^\infty(I, R^m)$ *if and only if* $0 \in \mathrm{int}\ R_g(T, L^\infty)$ *for every* $T \in (0, 1]$.

Proof. For $t \in (0, 1]$, we have

$$R_g(t, L^\infty) = \{H(t) \int_0^t H^{-1}(s) B(s) u(s) ds : u \in L^\infty\},$$

where H denotes the resolvent of the matrix function A. Let $W(t) = H^{-1}(t)[R_g(t, L^\infty)]$. Obviously, $W(t), R_g(t, L^\infty)$ are linear subspaces of R^d and, for $t_1 < t_2, W(t_1) \subset W(t_2)$ since

$$\int_0^{t_1} H^{-1}(s) B(s) u(s) ds = \int_0^{t_2} H^{-1}(s) B(s) \chi_{[0,t_1]}(s) u(s) ds$$

for any $u \in L^\infty$.

Suppose, to the contrary, that there is $T \in (0,1]$ such that $0 \notin \text{int } R_g(T, L^\infty)$, i.e., $R_g(T, L^\infty) \neq R^d$ and $\dim W(T) < d$. Let $r = \sup_{0\le t\le 1} \|H^{-1}(t)\|$ (here $\|\cdot\|$ denotes the norm of a linear map) and let $0 < \varepsilon < 1/r$. There is $0 < \delta < T$ such that $\|H(t) - I_d\| < \varepsilon$ for $0 \le t \le \delta$, where I_d denotes the unit d-dimensional matrix. Take $a \in R^d, |a| = 1$ such that $\inf_{w\in W(T)}|a - w| = 1$. Since (7) is STLC by means of $L^\infty, R^d = \cup_{0\le t\le\delta} R_g(t, L^\infty)$, so there is $b \in W(s), 0 \le s \le \delta$ such that $a = H(s)(b)$. But $|b| \le \|H^{-1}(s)\||a| \le r$ and $|a - b| = |(H(s) - I_d)(b)| < \varepsilon|b| < 1$, contradiction.

We have proved the "only if" part, the "if" part being self-evident. □

In view of Theorem 3.3 and Lemma 3.4 we get the following corollary.

COROLLARY 3.5. *Suppose that a perturbed linear control system* (3) *satisfies* (4), (5), *and* (6). *If the linearization* (7) *is* STLC *by means of* $L^\infty(I, R^m)$, *then, for every* $\varepsilon > 0$, *the system* (3) *is* STLC *by means of* $\{u \in L^\infty(I, R^m) : \|u\|_\infty < \varepsilon\}$.

4. Control systems given by vector fields. Let $\Omega = \{1, 2, \dots, k\}, U = \{u : I \to \Omega :$ u *is nondecreasing, continuous from the right*$\}$ and $f : I \times R^d \times \Omega \to R^d$ be a continuous function. In this case, the control system (1) can be described by a family $\{f_1, f_2, \dots, f_k\}$ of continuous vector fields $f_i : I \times R^d \to R^d, f_i(t, x) = f(t, x, i)$. There are some higher-order sufficient conditions of STLC for smooth autonomous vector fields (see [17]) and for Lipschitz continuous vector fields (see [6]).

We shall study this system under the following assumption:

$$0 \in \text{int}(co\{f_i(0,0) : i = 1, 2, \dots, k\}). \tag{17}$$

If f_i are Lipschitz continuous with respect to x, then (17) implies STLC of the control system

$$x' = \sum_1^k u_i(t) f_i(t, x), \qquad x(0) = 0,$$

by means of measurable scalar controls u_i such that $\sum_1^k u_i(t) = 1$ almost everywhere in I and $u_i \ge 0$ for $i = 1, 2, \dots k$ (see [6, Cor. 2.1]).

Below, we generalize this fact to the case of continuous vector fields. Moreover, for a given sufficiently small T, we describe a d-dimensional set K_T of controls such that $0 \in$ int $R_f(\le T, K_T)$.

The following combinatorial Lemma plays a crucial role in the sequel. Let $Y = \{y_1, y_2, \dots y_k\} \subset R^d$. By $cone(Y)$ we denote the convex cone spanned by Y, i.e.,

$$\text{cone}(Y) = \left\{ \sum_1^k \alpha_i y_i : \alpha_i \ge 0, 1 \le i \le k \right\}.$$

LEMMA 4.1. *If* $\text{cone}(Y) = R^d$, *then there is a family* $\mathcal{P}$ *of subsets of* Ω *such that:*

(i) *if* $S \in \mathcal{P}$, *then the set* $\{y_i : i \in S\}$ *is linearly independent;*

(ii) *if* $S \in \mathcal{P}$ *and* $R \subset S$, *then* $R \in \mathcal{P}$;

(iii) $\text{cone}\{y_i : i \in S\} \cap \text{cone}\{y_i : i \in R\} = \text{cone}\{y_i : i \in R \cap S\}$ *for any* S, $R \in \mathcal{P}$;

(iv) $\cup_{S\in\mathcal{P}} \text{cone}\{y_i : i \in S\} = \text{cone}\{y_i : i \in \cup_{S\in\mathcal{P}} S\}$;

(v) $\text{cone}\{y_i : i \in \cup_{S\in\mathcal{P}} S\} = R^d$.

The proof of this lemma will be given in Appendix.

To point out the evaluation time of a solution corresponding to a control $u \in U$ we formally reformulate the problem. First, set $\tilde{\Omega} = \Omega \cup \{k+1\}$ and let

$$\tilde{f}(t,x,i) = \begin{cases} f(t,x,i) & \text{for } 1 \le i \le k, \\ 0 & \text{for } i = k+1 \end{cases}$$

and, for $T \in (0,1]$, let $\tilde{U}_T = \{u : [0,T] \to \tilde{\Omega} : u \textit{ is nondecreasing continuous from the right and } u(T) = k+1\}$. There is a bijection between $\tilde{U}_T$ and the set $V_T = \{\alpha = (\alpha_1, \dots \alpha_k) \in R^k : \alpha_i \ge 0 \textit{ and } \sum_1^k \alpha_i \le T\}$. Namely, to any $\alpha \in V_T$ we assign a control $u_\alpha = 1 + \sum_{i=1}^k \chi_{(\sum_{j=1}^i \alpha_j, T]}$.

By Lemma 4.1 there is a family $\mathcal{P}$ associated with $Y = \{y_i = f_i(0,0) : i = 1,2,\dots k\}$. Let us define

$$K_T = \{u_\alpha \in \tilde{U}_T : \alpha \in V_T \text{ and } \{i : \alpha_i > 0\} \in \mathcal{P}\}.$$

THEOREM 4.2. *If a function $f : I \times R^d \times \Omega \to R^d$ is continuous and satisfies* (17), *then there is $T_0 > 0$ such that*

$$0 \in R_{\tilde{f}}(T, K_T) \quad \textit{for any } T \in (0, T_0].$$

The proof of Theorem 4.2 will be proceeded with two lemmas.

To avoid too many symbols, we use the same symbol K_T to denote the set $\{\alpha \in V_T : \{i \in \Omega : \alpha_i > 0\} \in \mathcal{P}\}$.

Let $W = \{(\alpha_1, \dots \alpha_k) \in R_+^k : \{i \in \Omega : \alpha_i > 0\} \in \mathcal{P}\}$. On W and K_T we have the metric inherited from R^k.

LEMMA 4.3. *A map $\eta : W \to R^d$ defined by the formula*

$$\eta(\alpha) = \sum_1^k \alpha_i y_i$$

is a homeomorphism.

Proof. Properties 4.1(iv), (v) of $\mathcal{P}$ show that η is surjective. Suppose that $\eta(\alpha) = \eta(\beta)$. By 4.1(iii), there is $\gamma \in W$ such that $\{i \in \Omega : \gamma_i > 0\} \subset \{i \in \Omega : \alpha_i > 0\} \cap \{i \in \Omega : \beta_i > 0\}$ and $\eta(\alpha) = \eta(\gamma)$. By 4.1(i), the system $\{y_i : \alpha_i > 0\}$ is linearly independent; hence $\gamma = \alpha$. Similarly we show that $\beta = \gamma$. The continuity of η follows from its linearity. For $S \in \mathcal{P}$, η^{-1} restricted to $\operatorname{cone}\{y_i : i \in S\}$ is linear and, therefore, continuous. Since cones $\operatorname{cone}\{y_i : i \in S\}, S \in \mathcal{P}$, are closed and, by 4.1(iv), (v), cover the whole space R^d, η^{-1} is also continuous. □

LEMMA 4.4. *A map $\nu^T : B^d \to \eta(K_T)$ given by the formula*

$$\nu^T(z) = \begin{cases} \left(\sum_1^k \beta_i\right)^{-1} T|z|z & \textit{for } z \ne 0 \textit{ and } \beta = \eta^{-1}(z), \\ 0 & \textit{for } z = 0 \end{cases}$$

is a homeomorphism and $\nu^T(S^{d-1}) = \{\eta(\alpha) : \alpha \in K_T \textit{ and } \sum_1^k \alpha_i = T\}$.

Proof. Observe that a map $\Theta^T : \eta(K_T) \to B^d$ given by the formula

$$\Theta^T(z) = \begin{cases} T^{-1}\left(\sum_1^k \alpha_i\right)|z|^{-1}z & \text{for } z \ne 0 \text{ and } \alpha = \eta^{-1}(z), \\ 0 & \text{for } z = 0 \end{cases}$$

is the inverse to ν^T. Maps ν^T, Θ^T are continuous in view of the continuity of η. □

Proof of Theorem 4.2. For any $S \in \mathcal{P}$, by (4.1), $0 \notin \operatorname{co}\{y_i : i \in S\}$; hence there is $\varepsilon > 0$ such that $0 \notin N_\varepsilon(\operatorname{co}\{y_i : i \in S\})$ and there is $z_S \in R^d$ such that $\langle z_S, z\rangle > 0$ for every $z \in N_\varepsilon(\operatorname{co}\{y_i : i \in S\})$.

There is $\delta \in (0,1]$ such that $|f_i(t,x) - f_i(0,0)| < \varepsilon$ for $|x| \le \delta, 0 \le t \le \delta$ and $i \in \Omega$. Define $c = \sup\{|f_i(t,x)| : 0 \le t \le \delta, |x| \le \delta, i \in \Omega\}$ and put $T_0 = \min(\delta, \delta/c)$. Fix $T \in (0, T_0]$. Let $X = \{x \in C([0,T], R^d) : \|x\| \le \delta\}$ and consider an integral operator $H : V_T \times I \times X \to X$ given by the formula

$$H(\alpha, \lambda, x)(t) = \int_0^t [(1-\lambda)\tilde{f}(s, x(s), u_\alpha(s)) + \lambda \tilde{f}(0,0,u_\alpha(s))]\, ds$$

for $\alpha \in V_T, x \in X, \lambda \in I$. This operator is well-defined on $V_T \times I \times X$, since $|H(\alpha,\lambda,x)(t)| \le Tc < \delta$. Evidently H is continuous and the set $\overline{H(V_T \times I \times X)}$ is compact. Hence, by Lemma 2.3, the map $S : V_T \times I \rightsquigarrow X$ given by the formula $S(\alpha,\lambda) =$ Fix $H(\alpha,\lambda,\cdot)$ is upper semicontinuous. On the other hand, we easily see that $S(\alpha,\lambda) = \mathrm{Sol}_{(1-\lambda)\tilde{f}+\lambda\tilde{f}(0,0,\cdot)}(u_\alpha)$, so, by Lemma 2.2, $S(\alpha,\lambda)$ is an R_δ-set for every $\alpha \in V_T, \lambda \in I$. Hence the map $\mathcal{H} : K_T \times I \rightsquigarrow R^d$ given by $\mathcal{H}(\alpha,\lambda) = \{x(T) : x \in S(\alpha,\lambda)\}$ is admissible.

By Lemmas 4.3 and 4.4, the set K_T is homeomorphic with B^d and the boundary $\partial K_T = \{u_\alpha : \sum_1^k \alpha_i = T\}$. We will show that $0 \notin \mathcal{H}(\alpha,\lambda)$ for every $\alpha \in \partial K_T, \lambda \in I$. Let $S = \{i \in \Omega : \alpha_i > 0\} \in \mathcal{P}$. We easily see that

$$(1-\lambda)f(t,x(t),i) + \lambda f(0,0,i) = (1-\lambda)f_i(t,x(t)) + \lambda y_i \in N_\varepsilon(\mathrm{co}\{y_i : i \in S\})$$

for $i \in S, t \in [0,T]$. Thus, for $i \in S$,

$$\langle z_S, (1-\lambda)f_i(t,x(t)) + \lambda y_i\rangle > 0$$

and $\langle x(T), z_S\rangle > 0$; hence $x(T) \neq 0$.

Now, observe that $\mathcal{H}(\alpha,1) = \sum_1^k \alpha_i y_i = \eta(\alpha)$. By Lemma 4.3 and Theorem 2.1(i), (iii), $0 \neq \mathrm{Deg}\ \mathcal{H}(\cdot,1) \in \mathrm{Deg}\ \mathcal{H}(\cdot,0)$, so, in view of Theorem 2.1(ii), we have the desired conclusion. □

COROLLARY 4.5. *Suppose that a function $f : I \times R^d \times \Omega \to R^d$ is continuous and satisfies* (17). *Then the control system* (1) *is* STLC *by the set of controls $U = \{u : I \to \Omega : u$ is nondecreasing, continuous from the right$\}$.*

5. Concluding remarks. The technique introduced in the paper may also be employed to the study of the point-to-point controllability problem of control systems. Considering the system (1), we say that (1) is controllable from zero to $y \in R^d$ in time $T \in (0,1]$ by means of $U_0 \subset U$ if $y \in R_f(T, U_0)$. We easily see that the controllability in the above sense is equivalent to the solvability of the generalized (set-valued) equation

$$y \in e_T \circ \mathrm{Sol}_f(u); u \in U_0 \tag{18}$$

where $e_T : C([0,1], R^d) \to R^d$ is the evaluation in the time T. The solvability of (18) may be established by the topological methods roughly discussed in the Introduction. These techniques usually yield stronger results: namely, that of local controllability, i.e., the issue states that $y \in \mathrm{int}\ R_f(T,K)$, where $K \subset U$ is a set homeomorphic to the ball B^d.

In this manner one is in a position to provide different proofs of some results of [7] or/and get the precise description of the control sets for some special controllability problems (see [8] and [12]).

6. Appendix. In this section we prove Lemma 4.1. Let $y \in R^d$ and let $y \notin C$; by $s(C,y)$ we denote the *set of points from C that are "seen" from y*, i.e.,

$$s(C,y) = \{c \in C : y + \lambda(c-y) \notin C \text{ for each } 0 < \lambda < 1\}.$$

We start with the following self-evident result.

LEMMA 6.1. *Let C be a closed convex cone in R^d and $y \notin C$.*

(i) *If* $c \in C \backslash s(C,y)$ *and* $\alpha > 0$, *then* $\alpha c \in C \backslash s(C,y)$.
(ii) *If* $c \in C \backslash s(C,y)$ *and* $d \in C$, *then* $c + d \in C \backslash s(C,y)$.

Proof of Lemma 4.1. The family $\mathcal{P}$ will be constructed inductively in a finite number of steps. First, let us choose $i_1, \dots i_N, 1 \le i_j \le k$, such that the system $\{y_{i_j} : 1 \le j \le N\}$ is an (algebraic) basis of R^d; we define $\mathcal{P}_1$ in the following way:

$$\mathcal{P}_1 = \{S : S \subset \{i_j : 1 \le j \le N\}\}.$$

We easily see that $\mathcal{P}_1$ satisfies conditions (i)–(iv). Assume that a constructed family $\mathcal{P}_r$ satisfies conditions (i)–(iv) but fails to satisfy condition (v). We shall define a family $\mathcal{P}_{r+1}$ satisfying (i)–(iv) and such that

$$\operatorname{card}\left(\bigcup_{S\in\mathcal{P}_{r+1}} S\right) = \operatorname{card}\left(\bigcup_{S\in\mathcal{P}_r} S\right) + 1,$$

where $\operatorname{card}(\cdot)$ denotes the cardinality of a set. Hence, after at most $(k - N)$-steps the defined family should satisfy condition (v), too.

Since $\mathcal{P}_r$ does not fulfill (v), there is $1 \le i_0 \le k$ such that $i_0 \in S$ for no $S \in \mathcal{P}_r$ and $y_{i_0} \notin C := \operatorname{cone}\{y_j : j \in \cup_{S\in\mathcal{P}_r} S\}$. Let us put

$$\mathcal{P}_{r+1} = \mathcal{P}_r \cup \{\{i_0\} \cup S : S \in \mathcal{P}_r \quad \text{and} \quad \operatorname{cone}\{y_i : i \in S\} \subset s(C, y_{i_0})\}.$$

We shall show that $\mathcal{P}_{r+1}$ fulfills (i)–(iv).

(i) Let $S \in \mathcal{P}_r$ such that $\{i_0\} \cup S \in \mathcal{P}_{r+1}$ and let $S = \{i_j : 1 \le j \le t\}$. It is sufficient to prove that $y_{i_0} \notin \operatorname{span}\{y_{i_j} : 1 \le j \le t\}$. Suppose to the contrary that $y_{i_0} = \sum_{j=1}^{t} \alpha_j y_{i_j}$. If $x = \sum\{\alpha_j y_{i_j} : 1 \le j \le t \text{ and } \alpha_j \ge 0\}$ and $y = \sum\{-\alpha_j y_{i_j} : 1 \le j \le t \text{ and } \alpha_j < 0\}$, then $y_{i_0} = x - y$ and $x, y \in \operatorname{cone}\{y_i : i \in S\}$. For any $\frac{1}{2} < \lambda < 1, y_{i_0} + \lambda(y - y_{i_0}) = (1-\lambda)x + (2\lambda - 1)y \in C$. Hence $y \in s(C, y_{i_0})$, a contradiction.

(ii) is obvious.

(iii) Let $S, R \in \mathcal{P}_{r+1}$. If $S, R \in \mathcal{P}_r$, then we are done. Assume that $S \in \mathcal{P}_r$ and $R \notin \mathcal{P}_r$. There is $R' \in \mathcal{P}_r$ such that $R = R' \cup \{i_0\}$. We show that $a \notin C$ provided $a \in \operatorname{cone}\{y_i : i \in R\} \setminus \operatorname{cone}\{y_i : i \in R'\}$. Let $a = \alpha y_{i_0} + y$, where $y \in \operatorname{cone}\{y_i : i \in R'\}$ and $\alpha > 0$. Evidently $a' = (2\alpha)^{-1} a = 2^{-1} y_{i_0} + w'$, where $w' \in \operatorname{cone}\{y_i : i \in R'\}$. Since $a' = y_{i_0} + 2^{-1}(2w' - y_{i_0})$ and $w' \in \operatorname{cone}\{y_i : i \in R'\} \subset s(C, y_{i_0})$, we have $a' \in C$ and, consequently, $a \in C$. Hence

$$(\operatorname{cone}\{y_i : i \in R\} \setminus \operatorname{cone}\{y_i : i \in R'\}) \cap \operatorname{cone}\{y_i : i \in S\} = \emptyset$$

and

$$\operatorname{cone}\{y_i : i \in R\} \cap \operatorname{cone}\{y_i : i \in S\} = \operatorname{cone}\{y_i : i \in R'\} \cap \operatorname{cone}\{y_i : i \in S\} = \operatorname{cone}\{y_i : i \in R' \cap S\} = \operatorname{cone}\{y_i : i \in R \cap S\}$$

because $R', S \in \mathcal{P}_r$.

Assume now that $R, S \notin \mathcal{P}_r$; thus $R = R' \cup \{i_0\}, S = S' \cup \{i_0\}$, where $R', S' \in \mathcal{P}_r$. Let $a \in \operatorname{cone}\{y_i : i \in S\} \cap \operatorname{cone}\{y_i : i \in R\}$. Hence $a = \alpha y_{i_0} + w = \beta y_{i_0} + z$, where $\alpha, \beta \ge 0$ and $w \in \operatorname{cone}\{y_i : i \in S'\}, z \in \operatorname{cone}\{y_i : i \in R'\}$. If $\alpha = 0$ or $\beta = 0$, then we are back in the case treated above. Without any loss of generality we may therefore assume that $0 < \alpha, \beta < 1$. For $\lambda_1 = (1 - \alpha)^{-1}$,

$$y_{i_0} + \lambda_1(a - y_{i_0}) = \lambda_1 w \in \operatorname{cone}\{y_i : i \in S'\} \subset s(C, y_{i_0})$$

and, for $\lambda_2 = (1-\beta)^{-1}$,

$$y_{i_0} + \lambda_2(a - y_{i_0}) = \lambda_2 z \in s(C, y_{i_0}).$$

Consequently, $\lambda_1 w = \lambda_2 z$ and

$$w \in \text{cone}\{y_i : i \in S'\} \cap \text{cone}\{y_i : i \in R'\} = \text{cone}\{y_i : i \in R' \cap S'\}.$$

This implies that $a \in \text{cone}\{y_i : i \in R \cap S\}$. The reverse inclusion is obvious.

(iv) Suppose that $a \in \text{cone}\{y_i : i \in S$ where $S \in \mathcal{P}_{r+1}\}\backslash C$. There is $w \in C$ and $\alpha > 0$ such that $a = \alpha y_{i_0} + w$. We may assume that $\alpha < 1$. Let $\lambda_0 = \inf\{\lambda > 0 : y_{i_0} + \lambda(a - y_{i_0}) \in C\}$. The number λ_0 is well-defined since $y_{i_0} + (1-\alpha)^{-1}(a - y_{i_0}) \in C$ and $\lambda_0 > 1$. Put $w_0 = y_{i_0} + \lambda_0(a - y_{i_0})$. Evidently,

$$w_0 \in s(C, y_{i_0}). \tag{19}$$

By (iv) for $\mathcal{P}_r$, there is $\hat{S} \in \mathcal{P}_r$ such that $w_0 \in \text{cone}\{y_i : i \in \hat{S}\}$. In view of (ii) for $\mathcal{P}_r$, we may assume that $\hat{S}$ is minimal in the sense that, for any $S \subset \hat{S}, S \neq \hat{S}, w_0 \notin \text{cone}\{y_i : i \in S\}$. We shall show that $\{i_0\} \cup \hat{S} \in \mathcal{P}_{r+1}$, i.e. that $\text{cone}\{y_i : i \in \hat{S}\} \subset s(C, y_{i_0})$. Take $w' \in \text{cone}\{y_i : i \in \hat{S}\}$. The minimality of $\hat{S}$ implies the existence of $\alpha_i > 0, \beta_i \geq 0, i \in \hat{S}$, such that

$$w_0 = \sum\{\alpha_i y_i : i \in \hat{S}\}, w' = \sum\{\beta_i y_i : i \in \hat{S}\}.$$

Take $\alpha_0 > 0$ such that $\alpha_0\beta_i \leq \alpha_i$ for $i \in \hat{S}$ and put $\gamma_i = \alpha_i - \alpha_0\beta_i$ and $w'' = \sum\{\gamma_i y_i : i \in \hat{S}\}$. Evidently $w'' \in \text{cone}\{y_i : i \in \hat{S}\}$ and $w_0 = \alpha_0 w' + w''$. If $w' \in C\backslash s(C, y_{i_0})$, then, by Lemma 6.1, $w_0 \in C\backslash s(C, y_{i_0})$ contrary to (19). Hence $\hat{S} \cup \{i_0\} \in \mathcal{P}_{r+1}$ and

$$a = \lambda_0^{-1}(w_0 - (1-\lambda_0)y_{i_0}) \in \text{cone}\{y_i : i \in \hat{S} \cup \{i_0\}\}$$

which ends the proof. □

REFERENCES

[1] N. Aronszajn, *Le corrèspondent topologique de l'unicitè dans la théorie des equations differentielles*, Ann. Math., 43 (1942), pp. 730–738.

[2] J. Bryszewski and L. Górniewicz, *Multi-valued maps of subsets of euclidean spaces*, Fund. Math., 90 (1976), pp. 233–251.

[3] R. Conti, *Linear differential equations and control*, Academic Press, London, 1976.

[4] F. S. De Blasi and J. Myjak, *On the solution sets for differential inclusions*, Bull. Acad. Polon. Sci., 33 (1985), pp. 17–23.

[5] A. Dold, *Lectures on algebraic topology*, Springer-Verlag, Berlin, 1972.

[6] H. Frankowska, *Local controllability and infinitesimal generators of semigroups of set-valued maps*, SIAM J. Control Optim., 25 (1987), pp. 412–432.

[7] M. Furi, P. Nistri, M. P. Pera, and P. L. Zezza, *Linear controllability by piecewise constant controls with assigned switching times*, J. Optim. Theory Appl., 45 (1985), pp. 219–229.

[8] ———, *Topological methods for the global controllability of nonlinear systems*, J. Optim. Theory Appl., 45 (1985), pp. 231–256.

[9] L. Górniewicz, *Homological methods in fixed point theory of multi-valued maps*, Discuss. Math. 129 (1975), pp. 1–71.

[10] J. K. Hale, *Ordinary differential equations*, Wiley-Interscience, New York, 1969.

[11] D. M. Hyman, *On decreasing sequences of compact absolute retracts*, Fund. Math., 64 (1969), pp. 91–97.

[12] W. Kryszewski and P. L. Zezza, *Remarks on the relay controllability of control systems*, to appear in J. Math. Anal. Appl.

[13] E. B. Lee and L. Markus, *Foundations of optimal control*, John Wiley, New York, 1964.

[14] N. G. Lloyd, *Degree theory*, Cambridge Univ. Press, Cambridge, 1978.

[15] P. Nistri, *On a general notion of controllability for nonlinear systems*, Boll. Un. Mat. Ital. Ser. VI, (1986), pp. 383–403.

[16] E. H. Spanier, *Algebraic Topology*, McGraw-Hill, New York, 1966.

[17] H. J. Sussman, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 224–260, January 1994

014

ON THE OPTIMAL CONTROL OF SYSTEMS DESCRIBED BY EVOLUTION EQUATIONS*

T. E. BAKER† AND E. POLAK†

Abstract. The authors present a mathematical foundation for the algorithmic solution of free- and fixed-time optimal control problems with evolution equation dynamics, finite-dimensional controls, and constraints on the controls and end points. In particular, (i) expressions for the derivatives of the solutions of the evolution equations are developed with respect to controls in $L_2^m[0, 1]$ and to the final time; (ii) the solutions of the relaxed evolution equations are shown to have a certain kind of directional derivative; (iii) algorithmic optimality conditions are developed with respect to both ordinary and relaxed controls and the final time; and (iv) an approximation theory is presented that shows that finite-dimensional minimax, and methods of centers-type algorithms can be used to obtain arbitrarily good approximations to stationary controls for optimal control problems with evolution equation dynamics and various constraints.

Key words. optimal control algorithms, approximation theory, relaxed controls, optimality conditions, evolution equations

AMS subject classifications. 49M27, 49M30, 49M39, 35B30

1. Introduction.

The results presented in this paper, dealing with the optimal control of evolution equations, were largely motivated by optimal slewing problems arising in the control of large, flexible, aerospace structures and in the control of various earthbound mechanisms with flexible links, which are naturally modeled by coupled systems of partial differential equations. Since, in practice, only finite element plant models may be available (which are in the form of ordinary differential equations (ODEs)), and since it is much easier to work with a canonical system representation, we assume that the plant dynamics are in evolution equation form, which permits us to treat both cases in a unified manner.

The majority of optimal control algorithms (see, for example, [May.1], [May.5], [Pir.2], [Teo.1], [Teo.2], [War.2]–[War.4], [Won.1]) are presented in *conceptual form*, i.e., the effects of numerical integration of the differential equations are ignored. In [Kle.1] we find an approximation theory for unconstrained optimal problems with ODE dynamics, in the form of an implementation of the method of steepest descent. More generally, this theory provides guidelines for adaptively increasing the precision of numerical integration so as to ensure that the numerical scheme retains the convergence properties of the conceptual one. It was later used by [Dun.1] to implement a conditional gradient method for optimal control problems with ODE dynamics. As far as optimal control problems with partial differential equation (PDE) dynamics are concerned, in [Gib.1], [Gib.2], [Gib.3], we find a detailed solution of the linear quadratic regulator problem, including conditions for the convergence of modal approximation schemes. However, for more general optimal control problems with PDE dynamics, the prevailing approach has been to use some method for constructing a *particular finite-dimensional* approximating optimal control problem and then to solve this problem by some method or other, see, e.g., [Jun.1], [Chu.1], [Ben.1], [Bur.1], [Flo.1]. The relationship between the solutions and stationary points of the approximating optimal control problem and those of the original optimal control problem is not established in these papers.

In this paper, we deal with the numerical solution of optimal control problems not by *adaptive implementation* of conceptual algorithms, but by *adaptive diagonalization*,

* Received by the editors November 28, 1989; accepted for publication (in revised form) August 4, 1992. This research was sponsored in part by National Science Foundation grant ECS-8121149, Air Force Office of Scientific Research grant AFOSR-83-0361, and the State of California MICRO Program.

† Department of Electrical Engineering and Computer Science, University of California, Berkeley, California 94720.

which requires less restrictive assumptions and, in our experience, seems to produce more efficient computational schemes. In any diagonalization approach, an original optimal control problem, **P**, is decomposed into an infinite sequence of *finite-dimensional* problems $\mathbf{P}_n, n = 1, 2, 3, \ldots$, which are solvable by nonlinear programming or nonsmooth optimization algorithms. These problems $\mathbf{P}_n$ must satisfy the following minimal consistency condition. Since, in the absence of convexity, finite-dimensional optimization algorithms can only be shown to compute stationary points, rather than optimal points, the problems $\mathbf{P}_n$ must be such that not only do their solutions converge to a solution of **P**, but also their (first-order) stationary points converge to a stationary point of **P**. Next, there is considerable empirical evidence to suggest that from a computational point of view, the most efficient approach is to proceed gradually, iterating toward a solution of a problem $\mathbf{P}_n$ until some test is satisfied and then carry over the last iterate as a starting point for problem $\mathbf{P}_{n+1}$, until the value of n is increased to some preassigned maximum value n^*, rather than to solve $\mathbf{P}_{n*}$ directly. In an adaptive diagonalization scheme, we can expect to find tests which determine not only when the solution of problem $\mathbf{P}_n$ should be arrested, but also the next value if n, which may be larger than $n + 1$. In return, as we will show later, the use of adaptive tests results in stronger convergence properties for the diagonalization method.

In developing an adaptive diagonalization scheme for the numerical solution of free- and fixed-time optimal control problems with evolution equation dynamics, finite-dimensional controls, and constraints on the controls and end points, we had to deal with (i) the differentiability of solutions of PDEs with respect to controls; (ii) optimality conditions for optimal control problems, which relate to those used in finite-dimensional nonlinear programming and nonsmooth optimization;[1] (iii) relaxed control theory in a PDE setting; (iv) conditions on the numerical methods for integrating the dynamical equations, to ensure consistent discretization, and (v) tests for progressing from $\mathbf{P}_n$ to $\mathbf{P}_{n+1}$.

The results presented in this paper extend and generalize the results in [Kle.1], [Wil.1]. In particular, the results in [Kle.1] do not apply to constrained problems and hence a new generation of tests had to be invented; furthermore, the results in [Kle.1], [Wil.1] apply only to problems with ODE dynamics. Nor were algorithms for constrained minimax optimal control problems, such as those considered in this paper, addressed in [Kle.1], [Wil.1].

In §2, we give a formulation of the problems that we will consider. In §3, we develop expressions for the derivatives of the solutions of the evolution equations with respect to controls in $(L_2^m[0, 1], \|\cdot\|_2)$ and the final time, and we establish first-order optimality conditions for minimax optimal control problems with control constraints and for optimal control problems with constraints on the control and inequality constraints on the final point. In §4 we introduce relaxed controls, extensions of the optimal control problems under consideration and develop appropriate extensions of the optimality conditions introduced in §3. In §5 we present our approximation theory and our adaptive diagonalization schemes. We show that these can be combined with a finite-dimensional minimax algorithm [Pir.1], [Psh.1], [Pol.1], and a new phase I–phase II method of feasible directions [Pol.2] to obtain arbitrarily good approximations to optimal controls for optimal control problems with evolution equation dynamics and various constraints. In §6 we present computational examples.

2. Formulation of optimal control problems. Many optimal control algorithms, including the ones to be presented in this paper, are extensions of finite-dimensional optimiza-

[1] It should be clear that, because the optimality conditions for finite-dimensional problems are in terms of "weak variations," in the absence of convexity, stationary controls of finite-dimensional approximations to an optimal control problem can only converge to a control satisfying a "weak" optimality condition. Hence the Maximum Principle is generally an inappropriate optimality condition within the particular numerical approximation framework considered in this paper.

tion algorithms that deal with problems defined in the Hilbert space $\mathbb{R}^n$. Now, the natural space for establishing differentiability of solutions of a differential equation with respect to m-dimensional controls is $L_\infty^m[0,1]$. However, adoption of $L_\infty^m[0,1]$ as the space for analysis leads to the somewhat awkward situation that the extensions of the finite-dimensional algorithms do not appear to be natural, because they require that we also use the $L_2^m[0,1]$ norm, $\|\cdot\|_2$, and $L_2^m[0,1]$ scalar product, $\langle\cdot,\cdot\rangle_2$.

Fortunately, we can also establish differentiability of solutions of a differential equation with respect to controls in the Hilbert space $L_2^m[0,1]$, provided that we impose a *growth condition* on the velocity function, as we will do shortly. In the case of control constrained optimal control problems, such as the ones treated in this paper, the imposition of a growth condition on the velocity function does not restrict the class of problems that can be considered and amounts to no more than a mathematically convenient device.

Finally, we recall that for any $u \in L_2^m[0,1]$, $\|u\|_2 \triangleq [\int_0^1 \|u(t)\|^2 dt]^{1/2}$, and for any $u, \nu \in L_2^m[0,1]$, $\langle u, \nu\rangle_2 \triangleq \int_0^1 \langle u(t), \nu(t)\rangle dt$, where $\|\cdot\|$ denotes the norm on $\mathbb{R}^m$ and $\langle\cdot,\cdot\rangle$ denotes the scalar product on $\mathbb{R}^m$.

We are now ready to proceed. For any $0 < \tau < \infty$, let $G(\tau)$ be the set of admissible controls defined by

$$G(\tau) \triangleq \{u \in L_2^m[0,\tau] | u(t) \in U, \text{for almost all } t \in [0,\tau]\}, \tag{2.1}$$

where U is a compact convex subset of $\mathbb{R}^m$.

Let X denote a Hilbert space with inner product $\langle\cdot,\cdot\rangle_X$ and corresponding norm $\|\cdot\|_X$. Let $A : D(A) \to X$ be the infinitesimal generator of a strongly continuous semigroup $\{T(t)\}_{t\geq 0}$; let $F : X \times \mathbb{R} \to X$ be a nonlinear operator that is Lipschitz continuous on bounded sets. We will consider dynamical systems of the following form:

$$\frac{d}{dt}\tilde{z}(t,\tilde{u}) = A\tilde{z}(t,\tilde{u}) + F(\tilde{z}(t,\tilde{u}),\tilde{u}(t)), \quad \tilde{z}(0,\tilde{u}) = z_0 \in D(A), \quad \tilde{u} \in G(\tau), \tag{2.2a}$$

where $\tilde{z}(t,\tilde{u}) \in X$, for all $t \in [0,\tau]$.

Because the set $U \subset \mathbb{R}^m$ is compact, there exists a bound $b < \infty$ such that for all $\nu \in U, |\nu^i| \leq b, i = 1,2,\ldots,m$. Hence, since our algorithms never violate the control constraint, we may assume, without loss of generality, that the operator F has the form $F(z,\nu) = \tilde{F}(z, SAT(\nu))$, where $SAT : \mathbb{R}^m \to \mathbb{R}^m$ is such $SAT(v) = (sat(\nu^1), sat(\nu^2), \ldots, sat(\nu^m))$, where for all $z \in \mathbb{R}$,

$$sat(z) = \begin{cases} z, & \text{if } |z| \leq 2b, \\ sgn(z)(2b + 1 - e^{(2b-|z|)}), & \text{if } |z| \geq 2b. \end{cases} \tag{2.2b}$$

This *growth condition* allows us in §3 to postulate local Lipschitz continuity conditions that are independent of bounds on the control.

We will assume that (2.2a) has a unique *mild solution*, which is defined as follows (see [Paz.1]).

DEFINITION 2.1. A function $\tilde{z}(\cdot,\tilde{u}) \in C([0,\tau],X)$ is said to be a *mild solution* to (2.2a) if for all $t \in [0,\tau]$,

$$\tilde{z}(t,\tilde{u}) = T(t)\tilde{z}_0 + \int_0^t T(t-s)F(\tilde{z}(s,\tilde{u}),\tilde{u}(s))ds. \tag{2.2c}$$

We can normalize[2] the final time in *fixed-time* optimal control problems (originally defined on $[0,\tau]$) to be 1 and reduce *free-time* optimal control problems to fixed-time

[2] Failure to normalize may lead to pathological computational results; see [Cul.1], [Cul.2].

optimal control problems on the interval $[0, 1]$ by replacing (2.2a) by *scaled* dynamics, with the scaling parameter denoted by τ. Thus, with each $\tilde{u} \in G(\tau)$, we associate a $u \in G(1)$ defined by $u(t) \triangleq \tilde{u}(\tau t)$ for $t \in [0, 1]$. With each $\tilde{z} \in C([0, \tau], X)$, we associate $z \in C([0, 1], X)$ defined by $z(t) \triangleq \tilde{z}(t\tau)$ for all $t \in [0, 1]$. Then, the function $z(t, u, \tau) \triangleq \tilde{z}(t\tau, \tilde{u})$ is a mild solution of the differential equation

$$\begin{aligned} \frac{d}{dt} z(t, u, \tau) &= \frac{d}{dt} \tilde{z}(t\tau, \tilde{u}) = \tau[A\tilde{z}(t\tau, \tilde{u}) + F(\tilde{z}(t\tau), \tilde{u}(t\tau))] \\ &= \tau[Az(t, u, \tau) + F(z(t, u, \tau), u(t))]. \end{aligned} \tag{2.2d}$$

Hence we abuse notation and let $G = G(1)$, and we replace the original dynamics (2.2a), with the scaled dynamics:

$$\begin{aligned} \frac{d}{dt} z(t, u, \tau) &= \tau[Az(t, u, \tau) + F(z(t, u, \tau), u(t))], \\ z(0, u) &= z_0 \in D(A), \quad t \in [0, 1]. \end{aligned} \tag{2.2e}$$

Note that for any final-time $\tau > 0$, the operator τA generates the semigroup $\{T(\tau t)\}_{t \geq 0}$ and hence $z(t, u, \tau)$ is a mild solution of (2.2e) if

$$z(t) = T(\tau t)z_0 + \tau \int_0^t T(\tau(t - s))F(z(s), u(s))\, ds. \tag{2.2f}$$

Next, for $j = 0, 1, 2, \ldots, q$, let $f^j : X \to \mathbb{R}$ be functions that are Lipschitz continuously differentiable on bounded sets. Then, for $j = 0, 1, 2, \ldots, q$, we define the functions $g^j : G \times (0, \infty) \to \mathbb{R}$ by $g^j(u, \tau) \triangleq f^j(z(1, u, \tau))$. The simplest problem that we will consider is

$$\textbf{MMP} : \inf\{\max_{j \in \mathbf{q}} g^j(u, \tau) | u \in G, \tau \in [\tau_{\min}, \tau_{\max}]\}, \tag{2.3a}$$

where $\mathbf{q} \triangleq \{1, 2, \ldots, q\}$, and $0 < \tau_{\min} \leq \tau_{\max} < \infty$. Note that when $\tau_{\max} = \tau_{\min}$, (2.3a) is a fixed-time problem; otherwise it is a free-time problem. In minimum time problems, $\tau_{\min}$ is chosen to be very small and $\tau_{\max}$ is chosen to be large, which ensures that the optimal value of the final-time, $\hat{\tau}$, is the minimum time.

We will also show that algorithms for solving **MMP** are trivially adapted to solving optimal control problems with control and end point inequality constraints, of the form

$$\textbf{CMP} : \inf\{g^0(u, \tau) | \, | \max_{j \in \mathbf{q}} g^j(u, \tau) \leq 0, u \in G, \tau \in [\tau_{\min}, \tau_{\max}]\}. \tag{2.3b}$$

Our next task is to establish optimality conditions for the problems **MMP** and **CMP**.

3. Optimality conditions. We begin with a few standard assumptions.

Assumption 3.1. (i) The operator $F(\cdot, \cdot)$ is Frechet differentiable. We will denote its partial Frechet derivatives, with respect to z and u, by $\partial F/\partial z(z, u)$ and $\partial F/\partial u(z, u)$, respectively.

(ii) For all $u \in G$, and $\tau \in [\tau_{\min}, \tau_{\max}]$, a solution to (2.2f) exists.

(iii) There exists $b_1 \in (0, \infty)$ such that for all $t \in [0, 1], u \in L_2^m[0, 1]$, and $\tau \in [\tau_{\min}, \tau_{\max}]$, $\|z(t, u, \tau)\|_X \leq b_1$.

(iv) For every bounded set $S \subset X$, there exists $K_S < \infty$ such that for all $z, z' \in S$ and all $u, u' \in \mathbb{R}^m$,

(a) $\|F(z',u') - F(z,u)\|_X \leq K_S[\|z'-z\|_X + \|u'-u\|]$,
(b) $\|\partial F/\partial z(z',u') - \partial F/\partial z(z,u)\| \leq K_S[\|z'-z\|_X + \|u'-u\|]$,
(c) $\|\partial F/\partial u(z',u') - \partial F/\partial u(z,u)\| \leq K_S[\|z'-z\|_X + \|u'-u\|]$.

(v) The functions $f^j(\cdot), j = 0,1,2,\ldots,q$, are Frechet differentiable; their Frechet differentials have the form $Df(z;\delta z) = \langle \nabla f^j(z), \delta z\rangle_X$, and their *gradients*, $\nabla f^j(\cdot)$, are Lipschitz continuous on the set $\{z \in X | \|z\|_X \leq b_1\}$.

The following assumption is needed only if the scaling parameter τ is allowed to vary.

Assumption 3.2. The semigroup generated by A, $\{T(t)\}_{t\geq 0}$, is an analytic semigroup.

The following two results can be gleaned from [Paz.1].

LEMMA 3.3. *The semigroup $\{T(t)\}_{t\geq 0}$ generated by the operator A is analytic if and only if there exists a constant $C < \infty$ such that* (i) *$T(t)$ is differentiable in $t > 0$;* (ii) *$d/dtT(t) = AT(t)$; and* (iii) *$\|AT(t)\|_X \leq C/t$, for all $t > 0$.*

Since Lemma 3.3 implies local Lipschitz continuity of $T(t)$ for $t > 0$, it follows from Assumption 3.1 and Lemma 3.3 that the following must be true.

LEMMA 3.4. *There exists a $b_2 \in (0,\infty)$, such that for all $z, z' \in S \triangleq \{z \in X | \|z\|_X \leq b_1\}$, all $u, u' \in \mathbb{R}^m$, all $\tau, \tau' \in [\tau_{\min}, \tau_{\max}]$, and all $t \in [0,1]$:*

(i) $\|\partial F/\partial z(z,u)\| \leq b_2$,
(ii) $\|\partial F/\partial u(z,u)\| \leq b_2$,
(iii) $\|\partial F/\partial z(z',u') - \partial F/\partial z(z,u)\| \leq b_2[\|z'-z\|_X + \|u'-u\|]$;
(iv) $\|\partial F/\partial u(z',u') - \partial F/\partial z(z,u)\| \leq b_2[\|z'-z\|_X + \|u'-u\|]$;
(v) $\|T(\tau' t) - T(\tau t)\| \leq b_2|\tau' - \tau|$.

In view of Assumption 3.1 and Lemma 3.4, it can be concluded from the Implicit Function Theorem in Banach spaces, as stated in [Lan.1], [Ale.1], that the solutions, $z(t,u,\tau)$, with $t \in [0,1]$, of (2.2f) are Lipschitz continuously Frechet differentiable with respect to (u,τ), with the Frechet differential, $Dz(t,u,\tau;\delta u,\delta\tau) = \delta z(t)$, where $\delta z(t)$ is the solution of the variational equation:

$$\begin{aligned}\delta z(t) = \int_0^t \Bigg\{ & T(\tau(t-s))\tau\left(\frac{\partial F}{\partial z}(z(s,u,\tau),u(s)\,\delta z(s) + \frac{\partial F}{\partial u}(z(s),u(s))\,\delta u(s)\right) \\ & + (T(\tau(t-s)) + \tau(t-s)AT(\tau(t-s)))F(z(s,u,\tau),u(s))\delta\tau \Bigg\} ds \\ & + tAT(\tau t)z_0\,\delta\tau.\end{aligned} \tag{3.1a}$$

We give an independent proof of this fact in the Appendix.

Since, by Assumption 3.1(v), the gradients of the functions $f^j(\cdot)$ are Lipschitz continuous on bounded sets, we immediately obtain the following result.

THEOREM 3.5. (i) *The functions $g^j : L_2^m[0,1] \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}, j = 0,1,2,\ldots,q$, defined in §2, are Frechet differentiable in (u,τ), i.e., for all $u \in G, \tau \in [\tau_{\min}, \tau_{\max}]$, there exists a continuous linear functional $Dg^j(u,\tau): L_2^m[0,1] \times \mathbb{R} \to \mathbb{R}$, such that for any $u, u' \in L_2^m[0,1], \tau, \tau' > \tau_{\min}$*

$$\lim_{\substack{\|u'-u\|_2 \to 0 \\ |\tau'-\tau|\to 0}} \frac{|g^j(u',\tau') - g^j(u,\tau) - Dg^j(u,\tau)(u'-u,\tau'-\tau)|}{(\|u'-u\|_2^2 + |\tau'-\tau|^2)^{1/2}} = 0. \tag{3.1b}$$

(ii) *There exist* gradients $\nabla g^j: L_2^m[0,1] \times [\tau_{\min}, \tau_{\max}] \to L_2^m[0,1] \times \mathbb{R}, j = 0,1,2,\ldots, q, \nabla g^j(u,\tau) = (\nabla_u g^j(u,\tau), \nabla_\tau g^j(u,\tau))$, *such that for all $u', u \in L_2^m[0,1], \tau', \tau \in [\tau_{\min}, \tau_{\max}]$,*

$$Dg^j(u,\tau)(u'-u,\tau'-\tau) = \langle \nabla_u g^j(u,\tau), u'-u\rangle_2 + \nabla_\tau g^j(u,\tau)(\tau'-\tau). \tag{3.1c}$$

(iii) *The gradients* $\nabla g^j(\cdot,\cdot)$ *are Lipschitz continuous on bounded sets.*

We are finally ready to address the question of optimality conditions for problems (2.3a), (2.3b). Because of algorithmic requirements, we chose a multiplier-free form for the optimality conditions. It is not difficult to show that these conditions are equivalent to standard optimality conditions involving multipliers. Thus, for problem (2.3a) we define the max function $\psi : G \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ and the corresponding *optimality function* $\theta_{\mathbf{MMP}} : G \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ by

$$\psi(u',\tau) \triangleq \max_{j\in\mathbf{q}} g^j(u',\tau'), \tag{3.2a}$$

$$\begin{aligned}\theta_{\mathbf{MMP}}(u',\tau') \triangleq \min_{(u,\tau)\in G\times[\tau_{\min},\tau_{\max}]} \Big\{ & \frac{1}{2}\|u-u'\|_2^2 + \frac{1}{2}|\tau-\tau'|^2 \\ & + \max_{j\in\mathbf{q}}\{g^j(u',\tau') - \psi(u',\tau') \\ & + \langle \nabla_u g^j(u',\tau'), u-u'\rangle_2 \\ & + \nabla_\tau g^j(u',\tau')(\tau-\tau')\}\Big\}.\end{aligned} \tag{3.2b}$$

Referring to Proposition 5.5 in [Pol.1], we see that $\theta_{\mathbf{MMP}}(u,\tau)$ is the obvious extension of an optimality function used in conjunction with first-order algorithms for the solution of minimax problems in $\mathbb{R}^n$. Hence, it is a correct optimality function to use in analyzing the convergence properties of *implementable* minimax algorithms for solving (2.3a), since such algorithms must construct finite-dimensional approximations to (2.3a).

THEOREM 3.6. (i) *The function* $\theta_{\mathbf{MMP}}(\cdot,\cdot)$ *is well defined and continuous.*

(ii) *If* $h_u(u',\tau') \in G - \{u'\}, h_\tau(u',\tau') \in [\tau_{\min},\tau_{\max}] - \{\tau'\}$ *are such that* $(u' + h_u(u',\tau'), \tau' + h_\tau(u',\tau'))$ *is a solution to the minimization problem* (3.2b)*, then* $h_u(\cdot,\cdot) : G \times [\tau_{\min},\tau_{\max}] \to L_2^m[0,1]$*, and* $h_\tau(\cdot,\cdot) : G \times [\tau_{\min},\tau_{\max}] \to \mathbb{R}$ *are unique and continuous.*

Proof. With $\sum_q \triangleq \{\mu \in \mathbb{R}^q | \sum_{j=1}^q \mu^j = 1, \mu \geq 0\}$, and making use of the Fan minimax theorem [Fan.1], we obtain that

$$\begin{aligned}\theta_{\mathbf{MMP}}(u,\tau) &= \min_{(u',\tau')\in G\times[\tau_{\min},\tau_{\max}]} \max_{\mu\in\Sigma_q} \Big\{ \frac{1}{2}\|u'-u\|_2^2 + \frac{1}{2}|\tau'-\tau|^2 \\ &\quad + \sum_{j\in\mathbf{q}} \mu^j\{g^j(u,\tau) - \psi(u,\tau) \\ &\quad + \langle \nabla_u g^j(u,\tau), u'-u\rangle_2 \\ &\quad + \nabla_\tau g^j(u,\tau)(\tau'-\tau)\}\Big\} \\ &= \max_{\mu\in\Sigma_q} \min_{(u',\tau')\in G\times[\tau_{\min},\tau_{\max}]} \Big\{ \frac{1}{2}\|u'-u\|_2^2 + \frac{1}{2}\|\tau'-\tau\|^2 \\ &\quad + \sum_{j\in\mathbf{q}} \mu^j\{g^j(u,\tau) - \psi(u,\tau) \\ &\quad + \langle \nabla_u g^j(u,\tau), u'-u\rangle_2 \\ &\quad + \nabla_\tau g^j(u,\tau)(\tau'-\tau)\}\Big\}\end{aligned} \tag{3.3a}$$

The minimization with respect to (u', τ') in (3.3a) is decoupled. The minimization with respect to τ' is a simple, one-dimensional quadratic problem. Because

$$\begin{aligned}&\frac{1}{2}\|u'-u\|_2^2+\sum_{j\in\mathbf{q}}\mu^j\langle\nabla_u g^j(u,\tau),u'-u\rangle_2\\&=\int_0^1\left(\frac{1}{2}\|u'(t)-u(t)\|^2+\sum_{j\in\mathbf{q}}\mu^j\langle\nabla_u g^j(u,\tau)(t),u'(t)-u(t)\rangle\right)dt,\end{aligned}\tag{3.3b}$$

the minimizing u' for (3.3a) can be constructed by minimizing the integrand pointwise in t in (3.3b). Consequently, $\theta_{\mathbf{MMP}}(u,\tau)$ is well defined. Continuity now follows from the Maximum Theorem in [Ber.1]. Similarly, since the solution $(h_u(\cdot,\cdot), h_\tau(\cdot,\cdot))$ of the minimization problem (3.2b) is unique, it again follows from the Maximum Theorem that it is continuous. □

THEOREM 3.7. *Suppose that* $(\hat{u},\hat{\tau})\in G\times[\tau_{\min},\tau_{\max}]$ *is an optimal solution to the problem* **MMP** (2.3a). *Then* $\theta_{\mathbf{MMP}}(\hat{u},\hat{\tau})=0$.

Proof. First, note that $\theta_{\mathbf{MMP}}(\hat{u},\hat{\tau})\le 0$ must hold. Hence, for the sake of contradiction, suppose that $\theta_{\mathbf{MMP}}(\hat{u},\hat{\tau})<0$ and that (u^*,τ^*) is the corresponding solution of the minimization problem (3.2b). Then, for $\lambda\in[0,1]$, we must have that $\hat{u}+\lambda(u^*-\hat{u})\in G, \hat{\tau}+\lambda(\tau^*-\hat{\tau})\in[\tau_{\min},\tau_{\max}]$, and

$$\begin{aligned}\psi(\hat{u}+&\lambda(u^*-\hat{u}),\hat{\tau}+\lambda(\tau^*-\hat{\tau}))-\psi(\hat{u},\hat{\tau})=\max_{j\in\mathbf{q}}g^j(\hat{u},\hat{\tau})-\psi(\hat{u},\hat{\tau})\\&+\lambda(\langle\nabla_u g^j(\hat{u},\hat{\tau}),u^*-\hat{u}\rangle_2+\nabla_\tau g^j(\hat{u},\hat{\tau})(\tau^*-\hat{\tau}))\}+o(\lambda)\\&\le\lambda\left\{\frac{1}{2}\|u^*-\hat{u}\|_2^2+\frac{1}{2}|\tau^*-\hat{\tau}|^2+\max_{j\in\mathbf{q}}\{g^j(\hat{u},\hat{\tau})-\psi(\hat{u},\hat{\tau})\right.\\&\left.+\langle\nabla_u g^j(\hat{u},\hat{\tau}),u^*-\hat{u}\rangle_2+\nabla_\tau g^j(\hat{u},\hat{\tau})(\tau^*-\hat{\tau})\}+\frac{o(\lambda)}{(\lambda)}\right\}\\&\le\lambda\{\theta_{\mathbf{MMP}}(\hat{u},\hat{\tau})+o(\lambda)/\lambda\},\end{aligned}\tag{3.4}$$

where $o(\lambda)/\lambda\to 0$ as $\lambda\to 0$. Hence there exists a $\hat{\lambda}\in[0,1]$ such that $\psi(\hat{u}+\hat{\lambda}(u^*-\hat{u}),\hat{\tau}+\hat{\lambda}(\tau^*-\hat{\tau}))<\psi(\hat{u},\hat{\tau})$, which is a contradiction. □

Under a convexity assumption, the above optimality condition becomes a necessary and sufficient condition. An examination of our definition of the functions $g^j(\cdot,\cdot)$ shows that they cannot be convex for free-time problems. However, in the case of linear dynamics and fixed end time, the problem can become convex.

We can easily obtain an optimailty condition for problem **CMP** (2.3b) from the one for **MMP** (2.3a) by making use of the following observation. Suppose that $(\hat{u},\hat{\tau})$ is an optimal pair for **CMP**. Let $\Psi: G\times[\tau_{\min},\tau_{\max}]\to\mathbb{R}$ be defined by

$$\Psi(u,\tau)=\max_{j\in\mathbf{q}}\max\{g^0(u,\tau)-g^0(\hat{u},\hat{\tau}),g^j(u,\tau)\}.\tag{3.5}$$

Then $\Psi(\hat{u},\hat{\tau})=0$ and, for any (u,τ) sufficiently close to $(\hat{u},\hat{\tau}), \Psi(u,\tau)\ge 0$. Hence $(\hat{u},\hat{\tau})$ is a local minimizer for the function $\Psi(\cdot,\cdot)$. Therefore, referring to (3.2a), (3.2b), we define the *optimality function* $\theta_{\mathbf{CMP}}: G\times[\tau_{\min},\tau_{\max}]\to\mathbb{R}$ by

$$\begin{aligned}\theta_{\mathbf{CMP}}(u',\tau')\triangleq\min_{(u,\tau)\in G\times[\tau_{\min},\tau_{\max}]}&\left\{\frac{1}{2}\|u-u'\|_2^2+\frac{1}{2}|\tau-\tau'|^2\right.\\&+\max_{j\in\mathbf{q}}\{-\psi(u',\tau')_+ +\langle\nabla_u g^0(u',\tau'),u-u'\rangle_2\\&+\nabla_\tau g^0(u',\tau')(\tau-\tau'),g^j(u,\tau)-\psi(u,\tau)_+\\&+\langle\nabla_u g^j(u,\tau),u'-u\rangle_2\\&\left.+\nabla_\tau g^j(u,\tau)(\tau'-\tau),j\in\mathbf{q}\}\right\},\end{aligned}\tag{3.6}$$

where $\psi(u,\tau)_+ \triangleq \max\{0,\psi(u,\tau)\}$. Although the term $\psi(u,\tau)_+$ has no effect at feasible points and hence also at optimal points, it is introduced into the optimality function for algorithmic reasons. The following result should be obvious.

THEOREM 3.8. (i) *The optimality function* $\theta_{\mathbf{CMP}}(\cdot\,,\cdot)$ *is well defined and continuous.*

(ii) *If* $h_u(u,\tau) \in G - \{u\}, h_\tau(u,\tau) \in [\tau_{\min},\tau_{\max}] - \{\tau\}$ *are such that* $(u + h_u(u,\tau), \tau + h_\tau(u,\tau))$ *is a solution to the minimization problem* (3.6), *then* $h_u(\cdot\,,\cdot), h_\tau(\cdot\,,\cdot)$ *are unique, continuous functions.*

(iii) *Suppose that* $(\hat{u},\hat{\tau}) \in G \times [\tau_{\min},\tau_{\max}]$ *is an optimal solution to the problem* **CMP** (2.3b). *Then* $\theta_{\mathbf{CMP}}(\hat{u},\hat{\tau}) = 0$.

It is customary to add a constraint qualification to optimization problems with inequality constraints. The analogue of the Slater constraint qualification [Sla.1] commonly used in nonlinear programming for problem **CMP** is as follows.

Assumption 3.9. We will assume that for all $(u,\tau) \in G \times [\tau_{\min},\tau_{\max}]$ such that $\psi(u,\tau) \geq 0, \theta_{\mathbf{MMP}}(u,\tau) < 0$.

Assumption 3.9 is standard in phase I–phase II methods of feasible directions. It implies that the constraint violation function $\psi(\cdot\,,\cdot)$ has no local minimizers outside of feasible set $\{(u,\tau) \in G \times [\tau_{\min},\tau_{\max}] | \psi(u,\tau) \leq 0\}$, nor on the set $\{(u,\tau) \in G \times [\tau_{\min},\tau_{\max}] | \psi(u,\tau) = 0\}$, a fact that prevents phase I–phase II feasible directions algorithms from converging to infeasible points. Finally, under Assumption 3.9 and a convexity assumption, $\theta_{\mathbf{CMP}}(u,\tau) = 0$ becomes both a necessary and sufficient condition of optimality.

4. Optimality conditions in the space of relaxed controls. Since the closed unit ball in $L_2^m[0,1]$ is not compact, there may be bounded sequences $\{(u_i,\tau_i)\}_{i=0}^{\infty}$, generated by an algorithm in solving the problem **MMP** or **CMP**, which have no accumulation points in $L_2^m[0,1]$, even when these problems do have solutions. However, as was established in [Ahm.1], [Pap.1], such sequences always have accumulation points in the space of relaxed controls. Hence, it is common to show that all the accumulation points generated by algorithms for solving optimal control problems, such as **MMP** and **CMP**, satisfy both a first-order optimality condition in $L_2^m[0,1]$ and the extension of these first-order conditions to first-order conditions for relaxed controls versions of **MMP** and **CMP**.

In order to define relaxed control versions of the problems **MMP** and **CMP**, we follow Warga [War.1], by defining $\overline{G}$, the relaxed controls closure of the set G, as follows:

$$\overline{G} = \{\sigma : [0,1] \to rpm(U) | \sigma \text{ is measurable}\}, \tag{4.1a}$$

where $rpm(U)$ denotes the set of Radon probability measures, topologized as in [War.1, Chap. 4]. In this topology, a sequence $\{\sigma_i\}_{i=0}^{\infty} \subset \overline{G}$ converges to a $\sigma \in \overline{G}$ if and only if

$$\lim_{i\to\infty} \int_0^1 \int_U \phi(t,u)\sigma_i(t)(du)dt = \int_0^1 \int_U \phi(t,u)\sigma(t)(du)dt, \forall \phi \in L_1([0,1], C(U)). \tag{4.1b}$$

The set $\overline{G}$ is sequentially compact. From our point of view, the most useful concept of continuity on $\overline{G}$ is that of sequential continuity. Hence all of our continuity statements, for functions defined on $\overline{G}$, are to be understood as sequential continuity statements, e.g., when we say that a function $\overline{g}$: $\overline{G} \to \mathbb{R}$ is continuous, we mean that for any sequence of relaxed controls $\{\sigma_i\}_{i=0}^{\infty} \subset \overline{G}$ that converges to a $\sigma \in \overline{G}, \overline{g}(\sigma_i) \to \overline{g}(\sigma)$, as $i \to \infty$.

Next, we extend the map $z : G \times [\tau_{\min},\tau_{\max}] \to C([0,1],X)$ to $\overline{z} : \overline{G} \times [\tau_{\min},\tau_{\max}] \to C([0,1],X)$ by defining for each $\sigma \in \overline{G}, \overline{z}(\cdot\,,\sigma,\tau) \in C([0,1],X)$ to be the solution to

$$z(t) = T(\tau t)z_0 + \tau \int_0^t \int_U T(\tau(t-s))F(z(s),u)\sigma(s)(du)ds. \tag{4.2}$$

Assuming that Assumption 3.1 holds, it can be shown that a mild solution to (4.2) exists, that it is unique, and that it is bounded by b_1, the bound introduced in Assumption 3.1(ii). The simplest relation between the solutions of (2.2f) and (4.2) is as follows.

PROPOSITION 4.1. *If $\sigma \in \overline{G}$ is an ordinary control, i.e., there exists $u \in G$ such that $\sigma(t)(S) = \delta_{u(t)}(S)$ for all measurable sets $S \subset U$ and almost all $t \in [0,1]$, then $z(t,\sigma,\tau) = z(t,u,\tau)$ for all $t \in [0,1]$, where $\overline{z}(\cdot,\sigma,\tau)$ is the solution to* (4.2) *and $z(\cdot,u,\tau)$ is the solution to* (2.2f).

The following result follows by simple extension of results in [Pap.1].

THEOREM 4.2 (Continuity of $\overline{z}(\cdot,\sigma,\tau)$ *in* (σ,τ)). *If the sequence $\{(\sigma_i,\tau_i)\}_{i=1}^{\infty} \subset \overline{G} \times [\tau_{\min},\tau_{\max}]$, as $i \to \infty$, is such that $\sigma_i \to \sigma \in \overline{G}, \tau_i \to \tau$, as $i \to \infty$, then $\overline{z}(\cdot,\sigma,\tau_i) \to \overline{z}(\cdot,\sigma,\tau)$, as $i \to \infty$.*

With these preliminaries out of the way, we are ready to define the relaxed control versions of the problems **MMP**, **CMP**, defined in (2.3a), (2.3b). Thus, for $j = 0,1,2,\ldots,q$, we define $\overline{g}^j : \overline{G} \times [\tau_{\min},\tau_{\max}] \to \mathbb{R}$ by $\overline{g}^j(\sigma,\tau) \triangleq f^j(\overline{z}(1,\sigma,\tau))$, and

$$\overline{\mathbf{MMP}}: \quad \min\{\max_{j\in\mathbf{q}} \overline{g}^j(\sigma,\tau) \mid \sigma \in \overline{G}, \tau \in [\tau_{\min},\tau_{\max}]\}, \tag{4.3a}$$

$$\overline{\mathbf{CMP}}: \quad \min\{\overline{g}^0(\sigma,\tau) \mid \max_{j\in\mathbf{q}} \overline{g}^j(\sigma,\tau) \le 0, \sigma \in \overline{G}, \tau \in [\tau_{\min},\tau_{\max}]\}. \tag{4.3b}$$

Next, we need to obtain extensions of the optimality functions $\theta_{\mathbf{MMP}}(\cdot,\cdot)$ and $\theta_{\mathbf{CMP}}(\cdot,\cdot)$ for the problems $\overline{\mathbf{MMP}}, \overline{\mathbf{CMP}}$, with the property that these extensions assume the same values on ordinary controls as the functions $\theta_{\mathbf{MMP}}(\cdot,\cdot)$ and $\theta_{\mathbf{CMP}}(\cdot,\cdot)$. On the surface, it is not at all clear how to obtain a relaxed control version of $\theta_{\mathbf{MMP}}(\cdot,\cdot)$ or of $\theta_{\mathbf{CMP}}(\cdot,\cdot)$. However, this task becomes much easier if we observe (see Theorem 3.4) that the solution $(u(u',\tau') - u', \tau(u',\tau') - \tau')$ of the search direction finding problem (3.2b) defines a pair of continuous functions $(h_u(\cdot,\cdot), h_\tau(\cdot,\cdot))$. Hence we set that (3.2b) is equivalent to

$$\theta_{\mathbf{MMP}}(\hat{u},\hat{\tau}) \triangleq \min_{\substack{(h_u,h_\tau)\in C(G\times[\tau_{\min},\tau_{\max}],\\ (G-\hat{u})\times([\tau_{\min},\tau_{\max}]-\hat{\tau}))}} \left\{ \begin{array}{l} \frac{1}{2}\|h_u(\hat{u},\hat{\tau})\|_2^2 + |h_\tau(\hat{u},\hat{\tau})|^2 \\ + \max_{j\in q}\{g_j(\hat{u},\hat{\tau}) - \psi(\hat{u},\hat{\tau}) \\ + \langle \nabla_u g^j(\hat{u},\hat{\tau}), h_u(\hat{u},\hat{\tau})\rangle_2 \\ + \nabla_\tau g^j(\hat{u},\hat{\tau}) h_\tau(\hat{u},\hat{\tau})\} \end{array} \right\} \tag{4.4a}$$

It is now clear that to obtain a relaxed control version of $\theta_{\mathbf{MMP}}(\cdot,\cdot)$ we must first obtain a relaxed control version of the directional derivatives $\langle \nabla_u g^j(u,\tau), h_u(u,\tau)\rangle_2 + \nabla_\tau g^j(u,\tau)(\tau' - \tau)$. Now, referring to (A.3) we see that $\delta z(t,u,\tau,\delta u,\delta\tau)$ is linear in $(\delta u,\delta\tau)$, and hence can be written as $\delta z(t,u,\tau,\delta u,\delta\tau) = \delta z_u(t,u,\tau,\delta u) + \delta z_\tau(t,u,\tau,\delta\tau) = \delta z(t,u,\tau,0,\delta\tau)$. Consequently,

$$\langle \nabla_u g^j(u,\tau), h_u(u,\tau)\rangle_2 = \langle \nabla f^j(z(1,u,\tau)), \delta z_u(1,u,\tau,h_u(1,u,\tau,h_u(u,\tau))\rangle_X, \tag{4.4b}$$

and

$$\nabla_\tau g^j(u,\tau)(\tau' - \tau) = \langle \nabla f^j(z(1,u,\tau)), \delta z_\tau(1,u,\tau,\tau' - \tau)\rangle_X. \tag{4.4c}$$

Hence, the relaxed control versions of (4.4b), (4.4c) appear to be

$$\langle \nabla f^j(\overline{z}(1,\sigma,\tau)), \overline{\delta z}_u(1,\sigma,\tau,h_u)\rangle_X, \tag{4.5a}$$

$$\langle \nabla f^j(\bar{z}(1,\sigma,\tau)), \overline{\delta z}_\tau(1,\sigma,\tau,\tau'-\tau)\rangle_X, \tag{4.5b}$$

where, with $h_u \in C([0,1]\times U\times[\tau_{\min},\tau_{\max}],\mathbb{R}^m)$ (i.e., its domain has been changed), $\overline{\delta z}_u(\cdot\,,\sigma,\tau,h_u)$ is the solution to

$$\begin{aligned}\overline{\delta z}(t) = \tau\int_0^t T(\tau(t-s))\int_U &\left\{\frac{\partial F}{\partial z}(\bar{z}(s,\sigma,\tau),u)\delta z(s)\right.\\ &\left.+\frac{\partial F}{\partial u}(\bar{z}(s,\sigma,\tau),u)h_u(s,u,\tau)\right\}\sigma(s)(du)ds,\end{aligned} \tag{4.5c}$$

and $\overline{\delta z}_\tau(\cdot\,,\sigma,\tau,\tau'-\tau)$ is the solution to

$$\begin{aligned}\overline{\delta z}(t) = \int_0^t\int_U &\left\{\tau T(\tau(t-s))\tfrac{\partial F}{\partial z}(\bar{z}(s,\sigma,\tau),u)\delta z(s)\right.\\ &\left.+(T(\tau(t-s))+\tau(t-s)AT(\tau(t-s)))F(\bar{z}(s,\sigma,\tau),u)\delta\tau\right\}\\ &\cdot\sigma(s)(du)ds+tAT(\tau t)z_0\delta\tau.\end{aligned} \tag{4.5d}$$

We will now show[3] that for any $\sigma\in\overline{G}$, $\lambda\in[-1,1]$, and a class of *search direction functions* $h_u\in C([0,1]\times U\times[\tau_{\min},\tau_{\max}],\mathbb{R}^m)$, $\lambda\overline{\delta z}_u(\cdot\,,\sigma,\tau,h_u)$ is a first-order approximation, in λ, to $\bar{z}(\cdot\,,\sigma,\tau,\lambda,h_u)-\bar{z}(\cdot\,,\sigma,\tau)$, where (with some abuse of notation) $\bar{z}(\cdot\,,\sigma,\tau,\lambda,h_u)$ is the solution to

$$z(t) = T(\tau t)z_0+\tau\int_0^t\int_U T(\tau(t-s))F(z(s),\nu+\lambda h_u(s,\nu,\tau))\sigma(s)(d\nu)ds. \tag{4.6}$$

We note that (4.5c) is the first variation of (4.6) along the curve in $\overline{G}$ defined by $\{\rho(\cdot\,;\lambda,h_u)|\ \lambda\in[0,1]\}$, where

$$\rho(t;\lambda,h_u)(S)\triangleq\{\sigma(t)(R), R\triangleq\{\nu\in U|\nu+\lambda h_u(t,\nu,\tau)\in S\}\} \tag{4.7}$$

if $\nu+h_u(t,\nu,\tau)\in U$ for all $\nu\in U$ and almost all $t\in[0,1]$; otherwise ρ is undefined. It is easily seen that if ρ is well defined, then $\rho\in\overline{G}$ and $\bar{z}(1,\rho(\cdot,\lambda,h_u),\tau)=\bar{z}(1,\sigma,\tau,\lambda,h_u)$. Hence we introduce the following definition.

DEFINITION 4.3. *The search direction function $h_u\in C([0,1]\times U\times[\tau_{\min},\tau_{\max}],\mathbb{R}^m)$ will be said to be admissible if $u'+h_u(t,u',\tau')\in U$ for all $u'\in U$ and almost all $t\in[0,1]$ and $\tau\in[\tau_{\min},\tau_{\max}]$. We will denote by Γ the set of admissible search direction functions.*

LEMMA 4.4. *There exists an $L<\infty$ such that for any $h_u\in\Gamma,\sigma\in\overline{G},\tau\in[\tau_{\min},\tau_{\max}]$, $t\in[0,1]$ and λ sufficiently small, $\|\bar{z}(t,\sigma,\tau,\lambda,h_u)-\bar{z}(t,\sigma,\tau)\|\le L|\lambda|$.*

Proof. Let

$$M_U\triangleq\max\{\|u'-u''\||u',u''\in U\}. \tag{4.8a}$$

Since U is compact, $M_U<\infty$. Clearly, for every $h_u\in\Gamma$, $\|h_u(t,u',\tau')\|\le M_U$ for all $t\in[0,1]$, $u'\in U$, and $\tau'\in[\tau_{\min},\tau_{\max}]$. Hence

$$\begin{aligned}&\|\bar{z}(t,\sigma,\tau,\lambda,h_u)-\bar{z}(t,\sigma,\tau)\|_X\\ &\quad=\left\|\int_0^t\int_U\tau T(\tau(t-s))[F(\bar{z}(s,\sigma,\tau,\lambda,h_u),u+\lambda h_u(s,u,\tau))\right.\\ &\qquad\left.-F(\bar{z}(s,\sigma,\tau),u)]\sigma(s)(du)ds\right\|_X\\ &\quad\le\tau_{\max}M\int_0^t K_S[\|\bar{z}(s,\sigma,\tau,\lambda,h_u)\\ &\qquad-\bar{z}(s,\sigma,\tau)\|+|\lambda|M_U]ds,\end{aligned} \tag{4.8b}$$

[3] A similar development for ODEs can be found in [Wil.1].

where M is a bound on $\|T(\tau(t-s))\|, s \in [0,t]$, as also used in the Appendix. Applying the Bellman–Gronwall inequality, we obtain that

$$\|\overline{z}(t,\sigma,\tau,\lambda,h_u) - \overline{z}(t,\sigma,\tau)\| \leq L|\lambda|, \tag{4.8c}$$

where $L \triangleq MK_SM_Ue^{\tau_{\max}MK_S}$ and K_S is defined as in Assumption 3.1(iv). □

LEMMA 4.5. *There exists $d_1 < \infty$ such that for all $t \in [0,1], \sigma \in \overline{G}, \tau \in [\tau_{\min}, \tau_{\max}], h_u \in \Gamma$ and $\lambda \in [-1,1]$,*

$$\|\overline{z}(t,\sigma,\tau,\lambda,h_u) - \overline{z}(t,\sigma,\tau) - \lambda\overline{\delta z}_u(t,\sigma,\tau,h_u)\|_X \leq d_1|\lambda|^2. \tag{4.9}$$

Proof. Let $\Delta\overline{z}(t,\sigma,\tau,\lambda,h_u) \triangleq \overline{z}(t,\sigma,\tau,\lambda,h_u) - \overline{z}(t,\sigma,\tau)$. Then, with M_U as in (4.8a),

$$\begin{aligned}
&\|\Delta\overline{z}(t,\sigma,\tau,\lambda,h_u) - \lambda\overline{\delta z}_u(t,\sigma,\tau,h_u)\|_X \\
&\quad\leq \left\| \int_0^t \int_U \tau T(\tau(t-s)) \Big[F(\overline{z}(s,\sigma,\tau,\lambda,h_u), u + \lambda h_u(s,u)) \right. \\
&\qquad - F(\overline{z}(s,\sigma,\tau),u) - \frac{\partial F}{\partial z}(\overline{z}(s,\sigma,\tau),u)\delta\overline{z}(s,\sigma,\tau,\lambda,h_u) \\
&\qquad \left. - \frac{\partial F}{\partial u}(\overline{z}(s,\sigma,\tau),u)\cdot\lambda h_u(s,u)\Big]\sigma(s)(du)ds \right\|_X \\
&\quad\leq \tau_{\max}M \int_0^t \int_U \left[\int_0^1 \left\| \frac{\partial F}{\partial z}(\overline{z}(s,\sigma,\tau,\lambda,h_u) + r\Delta\overline{z}(s,\sigma,\tau,\lambda,h^u), u + r\lambda h_u(s,u)) \right.\right. \\
&\qquad \left. - \frac{\partial F}{\partial z}(\overline{z}(s,\sigma,\tau),u)\right\| dr \Delta\overline{z}(s,\sigma,\tau,\lambda,h_u) \Big\|_X \\
&\qquad + \int_0^1 \left\| \frac{\partial F}{\partial u}(\overline{z}(s,\sigma,\tau,\lambda,h_u) + r\Delta\overline{z}(s,\sigma,\tau,\lambda,h_u), u + r\lambda h_u(s,u)) \right. \\
&\qquad\qquad \left. - \frac{\partial F}{\partial u}(\overline{z}(s,\sigma,\tau),u)\right\| dr|\lambda|M_U \\
&\qquad + \left\| \frac{\partial F}{\partial z}(\overline{z}(s,\sigma,\tau),u)\right\| \|\Delta\overline{z}(s,\sigma,\tau,\lambda,h_u) - \overline{\delta z}_u(s,\sigma,\tau,\lambda,h_u)\|_X \Big] \sigma(s)(du)ds \\
&\quad\leq \tau_{\max}M \int_0^t \int_U [K_S(\|\Delta\overline{z}(s,\sigma,\tau,\lambda,h_u)\|_X + |\lambda|M_U)\|\Delta\overline{z}(s,\sigma,\tau,\lambda,h_u)\|_X \\
&\qquad + K_S(\|\Delta\overline{z}(s,\sigma,\tau,\lambda,h_u)\|_X + |\lambda|M_U)|\lambda M_U \\
&\qquad + b_2|\lambda|M_U\|\Delta\overline{z}(s,\sigma,\tau,\lambda,h_u) - \delta\overline{z}(s,\sigma,\tau,\lambda,h_u)\|_X]\sigma(s)(du)ds,
\end{aligned} \tag{4.10a}$$

where b_2 is defined in Lemma 3.4. Since by Lemma 4.4 $\|\Delta\overline{z}(s,\sigma\tau,\lambda,h_u)\|_X \leq L|\lambda|$, it follows from the Bellman–Gronwall inequality that

$$\begin{aligned}
&\|\Delta\overline{z}(t,\sigma,\tau,\lambda,h_u) - \lambda\overline{\delta z}_u(t,\sigma,\tau,h_u)\|_X \\
&\quad\leq MK_Se^{Mb_2}(\|\Delta\overline{z}(s,\sigma,\tau,\lambda,h_u)\| + |\lambda|M_U)^2 \\
&\quad\leq d_1|\lambda|^2,
\end{aligned} \tag{4.10b}$$

where $d_1 = MK_Se^{Mb_2}(L + M_U)^2$. □

Proceeding in a similar manner, we can also prove the following, somewhat simpler, result.

LEMMA 4.6. *There exists a $d_2 < \infty$ such that for all $h_u \in \Gamma, t \in [0,1], \sigma \in \overline{G}, \lambda \in [-1,1], \tau', \tau \in [\tau_{\min}, \tau_{\max}]$,*

$$\|\overline{z}(t,\sigma,\tau,\lambda,h_u) - \overline{z}(t,\sigma,\tau) - \lambda\overline{\delta z}(t,\sigma,\tau,\tau'-\tau)\|_X \leq d_2|\lambda|^2. \tag{4.11}$$

In addition, it is fairly easy to establish the following result.

LEMMA 4.7. *For any* $t \in [0,1], \sigma \in \overline{G}, \tau \in [\tau_{\min}, \tau_{\max}]$, *admissible* h_u, *and* $\tau' \in [\tau_{\min}, \tau_{\max}]$, *let* $\overline{\delta z}(t, \sigma, \tau, h_u, \tau' - \tau)$ *denote the solution to*

$$
\begin{aligned}
\delta z(t) = \int_0^t \int_U \tau T(\tau(t-s)) \Big\{ & \frac{\partial F}{\partial z}(\overline{z}(s,\sigma,\tau),u)\delta z(s) \\
& + \frac{\partial F}{\partial u}(\overline{z}(s,\sigma,\tau),u)h_u(s,u,\tau) \Big\} \sigma(s)(du)ds \\
+ \int_0^t & \{(T(\tau(t-s)) + \tau(t-s)AT(\tau(t-s)))F(\overline{z}(s,\sigma,\tau),u)\delta\tau\}\, ds \\
+ & tAT(\tau t)z_0\delta\tau.
\end{aligned} \tag{4.12}
$$

Then (i) $\overline{\delta z}(t,\sigma,\tau,h_u,\tau'-\tau) = \overline{\delta z}_u(t,\sigma,\tau,h_u) + \overline{\delta z}_\tau(t,\sigma,\tau,\tau'-\tau)$, *and* (ii) $\overline{\delta z}(t,\sigma,\tau,h_u,\tau'-\tau)$ *is continuous in* $(t,\sigma,\tau,h_u,\tau')$.

We are now ready to extend the optimality conditions in §3 to the relaxed optimal control problems $\overline{\mathbf{MMP}}, \overline{\mathbf{CMP}}$. We define the max function $\overline{\psi} : \overline{G} \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ and the optimality function $\overline{\theta}_{\mathbf{MMP}} : \overline{G} \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ by

$$
\overline{\psi}(\sigma,\tau) \triangleq \max_{j \in \mathbf{q}} \overline{g}^j(\sigma,\tau), \tag{4.13a}
$$

$$
\begin{aligned}
\overline{\theta}_{\mathbf{MMP}}(\sigma,\tau) \triangleq \min_{(w,\tau') \in \Gamma \times [\tau_{\min},\tau_{\max}]} \Big\{ & \frac{1}{2}\int_0^1 \int_U \|w(t,u,\tau)\|^2 \sigma(t)(du)dt + \frac{1}{2}|\tau'-\tau|^2 \\
& + \max_{j \in \mathbf{q}} \{\overline{g}^j(\sigma,\tau) - \overline{\psi}(\sigma,\tau) \\
& + \langle \nabla f^j(\overline{z}(1,\sigma,\tau)), \overline{\delta z}_u(1,\sigma,\tau,w)\rangle_X \\
& + \langle \nabla f^j(\overline{z}(1,\sigma,\tau)), \overline{\delta z}_\tau(1,\sigma,\tau,\tau'-\tau)\rangle_X \} \Big\}.
\end{aligned} \tag{4.13b}
$$

Making use of Lemmas 4.6 and 4.7, we immediately get the following extension of Theorems 3.4 and 3.5.

THEOREM 4.8. (i) *The function* $\overline{\theta_{\mathbf{MMP}}}(\cdot,\cdot)$ *is well defined and continuous.*

(ii) *Suppose that* $(\hat{\sigma},\hat{\tau}) \in \overline{G} \times [\tau_{\min}, \tau_{\max}]$ *is an optimal solution to the problem* $\overline{\mathbf{MMP}}$ (4.3a). *Then* $\overline{\theta_{\mathbf{MMP}}}(\hat{\sigma},\hat{\tau}) = 0$.

Similarly, we can define an extension, $\overline{\theta}_{\mathbf{CMP}} : \overline{G} \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ of the optimality function $\theta_{\mathbf{CMP}}(\cdot,\cdot)$ as follows:

$$
\begin{aligned}
\overline{\theta}_{\mathbf{CMP}}(\sigma,\tau) \triangleq \min_{(w,\tau') \in \Gamma \times [\tau_{\min},\tau_{\max}]} \Big\{ & \frac{1}{2}\int_0^1 \int_U \|w(t,u,\tau)\|^2 \sigma(t)(du)dt + \frac{1}{2}|\tau'-\tau|^2 \\
& + \max\{-\overline{\psi}(\sigma,\tau)_+ \\
& + \langle \nabla f^0(\overline{z}(1,\sigma,\tau)), \overline{\delta z}_u(1,\sigma,\tau,w)\rangle_X \\
& + \langle \nabla f^0(\overline{z}(1,\sigma,\tau), \overline{\delta z}_\tau(1,\sigma,\tau,\tau'-\tau))\rangle_X, \overline{g}^j(\sigma,\tau) \\
& - \overline{\psi}(\sigma,\tau)_+ + \langle \nabla f^j(\overline{z}(1,\sigma,\tau)), \overline{\delta z}_u(1,\sigma,\tau,w)\rangle_X \\
& + \langle \nabla f^j(\overline{z}(1,\sigma,\tau), \overline{\delta z}_\tau(1,\sigma,\tau,\tau'-\tau))\rangle_X, j \in \mathbf{q}\} \Big\},
\end{aligned} \tag{4.14}
$$

where $\overline{\psi}(\sigma,\tau)_+ \triangleq \max\{0, \overline{\psi}(\sigma,\tau)\}$. We can now state the obvious extension of Theorem 3.6.

THEOREM 4.9. (i) *The function* $\overline{\theta}_{\mathbf{CMP}}(\cdot,\cdot)$ *is well defined and continuous.*

(ii) *Suppose that* $(\hat{\sigma}, \hat{\tau}) \in \overline{G} \times [\tau_{\min}, \tau_{\max}]$ *is an optimal solution to the problem* **CMP** (4.3b). *Then* $\overline{\theta}_{\mathbf{CMP}}(\hat{\sigma}, \hat{\tau}) = 0$.

We conclude this section with a rather obvious result that is essential in the analysis of optimal control algorithms.

THEOREM 4.10. *Suppose that* $\sigma^* \in \overline{G}$ *is an ordinary control, i.e., there exists a* $u^* \in G$ *such that* $\sigma^*(t)(S) = \delta_{u^*(t)}(S)$ *for all measurable sets* $S \subset U$ *and almost all* $t \in [0,1]$. *Then*

(i) *For any* $t \in [0,1], h_u \in \Gamma, \tau', \tau \in [\tau_{\min}, \tau_{\max}], \overline{\delta z}(t, \sigma^*, \tau, h_u, \tau' - \tau) = \delta z(t, u, \tau, \delta u, \delta\tau)$, *where* $\delta u(t) = h_u(t, u^*(t), \tau)$ *and* $\delta\tau = \tau' - \tau$.

(ii) $\overline{\theta}_{\mathbf{MMP}}(\sigma^*, \tau) = \theta_{\mathbf{MMP}}(u^*, \tau)$, *and* $\overline{\theta}_{\mathbf{CMP}}(\sigma^*, \tau) = \theta_{\mathbf{CMP}}(u^*, \tau)$.

Thus we see from Theorem 4.10 that when σ^* is an ordinary control, the stationary points of (3.2b) and (3.6) are also the stationary points of (4.13b) and (4.14), respectively.

5. Approximation theory. The numerical solution of optimal control problems such as **MMP** and **CMP** is impossible without some sort of discretization of the evolution equation (2.2f). We will now develop a theory for discretization of these problems. This theory depends on the convergence of the finite element method and on error bounds, such as those to be found in [Fuj.1], [Fuj.2], [Fuj.3].

The use of a numerical method in integrating the evolution system (2.2f) results in the replacement of the set of admissible controls G by G_n, a compact, convex, finite-dimensional subset of G, and of the original functions $g^j : G \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ by approximating functions $g_n^j : G_n \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$, where $n \in \mathbb{Z}_+$ is a precision control parameter. Thus, the use of numerical integration results in the replacement of the original problems **MMP** and **CMP** by approximations.

Hence, to establish an approximation theory, for each $n \in \mathbb{Z}_+$, we define the discretized problems $\mathbf{MMP}_n$ and $\mathbf{CMP}_n$ by

$$\mathbf{MMP}_n : \min\{\max_{j \in \mathbf{q}} g_n^j(u, \tau) | u \in G_n, \tau \in [\tau_{\min}, \tau_{\max}]\}, \tag{5.1a}$$

$$\mathbf{CMP}_n : \min\{g_n^0(u, \tau) | \max_{j \in \mathbf{q}} g_n^j(u, \tau) \leq 0, u \in G_n, \tau \in [\tau_{\min}, \tau_{\max}]\}. \tag{5.1b}$$

To ensure that the functions $g_n^j(\cdot, \cdot)$ inherit the continuity and differentiability properties of the functions $g^j(\cdot, \cdot)$, we make the following reasonable assumptions.

Assumption 5.1. (i) For all $n \in \mathbb{Z}_+$, the functions $g_n^j : G_n \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ are continuous.

(ii) For all $n \in \mathbb{Z}_+, j = 0, 1, 2, \ldots, q$, and each $(u, \tau) \in G_n \times [\tau_{\min}, \tau_{\max}]$, there exists a gradient $\nabla g_n^j(u, \tau) = (\nabla g_{u,n}^j(u, \tau), \nabla g_{\tau,n}^j(u, \tau)) \in L_2^m[0,1] \times \mathbb{R}$, such that for all $u' \in G_n, \tau' \in [\tau_{\min}, \tau_{\max}]$,

$$\lim_{\substack{\|u'-u\|_2 \to 0 \\ |\tau'-\tau| \to 0}} \frac{|g_n^j(u'\tau') - g_n^j(u, \tau) - (\langle \nabla_u g_n^j(u), \nu - u \rangle_2 + \nabla_\tau g_n^j(u, \tau)(\tau' - \tau))|}{(\|u' - u\|_2^2 + |\tau' - \tau|^2)^{1/2}} = 0. \tag{5.2a}$$

(iii) There exists a Lipschitz constant[4] $L \in (0, \infty)$, such that for all $n \in \mathbb{Z}_+, j = 0, 1, 2, \ldots, q, u', u \in G_n, \tau, \tau' \in [\tau_{\min}, \tau_{\max}]$,

$$\|\nabla g_n^j(u', \tau') - \nabla g_n^j(u, \tau)\|_2 \leq L(\|u' - u\|_2^2 + |\tau' - \tau|^2)^{1/2}. \tag{5.2b}$$

(iv) For all $n \in \mathbb{Z}_+, G_n \subset G_{n+1}$.

[4] The existence of such a Lipschitz constant is a consequence of Assumptions (3.1)(iv), (3.1)(v).

(v) The closure of $\bigcup_{n\in\mathbb{Z}_+} G_n$ is G.

(vi) (Uniform Approximation Property.) For all $\varepsilon > 0$, there exists n_ε such that for all $n \geq n_\varepsilon, j = 0, 1, 2, \ldots, q$, all $u \in G_n$, and all $\tau \in [\tau_{\min}, \tau_{\max}]$,

$$|g^j(u,\tau) - g_n^j(u,\tau)| \leq \varepsilon, \tag{5.2c}$$

$$\|\nabla g^j(u,\tau) - \nabla g_n^j(u,\tau)\|_2 \leq \varepsilon. \tag{5.2d}$$

Usually, when continuous dynamical equations are replaced by discrete dynamic equations, the resulting solutions inherit the continuity and differentiability properties of the original solutions, and hence satisfy Assumptions 5.1(i)–(iii). Assumption 5.1(vi) is satisfied at any particular u for any dynamics on which the finite element method converges. Thus the only thing we must verify is that the approximation is uniform on the finite-dimensional set, G_n, as assumed.

Referring to Proposition 5.5 in [Pol.1], we see that the following analogues of Theorems 3.11 and 3.12 must hold for the problems $\mathbf{MMP}_n, \mathbf{CMP}_n$.

THEOREM 5.2. *For $n \in \mathbb{Z}_+$, let $\psi_n : G_n \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ and the corresponding optimality function $\theta_{\mathbf{MMP}n} : G_n \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ be defined by*

$$\psi_n(u',\tau') \triangleq \max_{j\in\mathbf{q}} g_n^j(u',\tau'), \tag{5.3a}$$

$$\begin{aligned}\theta_{\mathbf{MMP}n}(u',\tau') \triangleq \min_{(u,\tau)\in G_n\times[\tau_{\min},\tau_{\max}]} \Big\{ & \frac{1}{2}\|u-u'\|_2^2 + \frac{1}{2}|\tau-\tau'|^2 \\ & + \max_{j\in\mathbf{q}}\{g_n^j(u',\tau') - \psi_n(u',\tau') \\ & + \langle \nabla_u g_n^j(u',\tau'), u-u'\rangle \\ & + \nabla_\tau g_n^j(u',\tau')(\tau-\tau')\}\Big\}.\end{aligned} \tag{5.3b}$$

Then

(i) *The optimality function $\theta_{\mathbf{CMP}}(\cdot,\cdot)$ is well defined and continuous.*

(ii) *If $h_u(u',\tau') \in G_n - \{u'\}$, $h_\tau(u',\tau') \in [\tau_{\min}, \tau_{\max}] - \{\tau'\}$ are such that $(u' + h_u(u',\tau'), \tau' + h_\tau(u',\tau'))$ is a solution to* (5.3b)*, then $h_u(\cdot,\cdot), h_\tau(\cdot,\cdot)$ are continuous functions.*

(iii) *Suppose that $(u_n', \tau_n') \in G_n \times [\tau_{\min}, \tau_{\max}]$ is an optimal solution to the problem* **CMP**. *Then $\theta_{\mathbf{CMP}}(u_n', \tau_n') = 0$.*

THEOREM 5.3. *For $n \in \mathbb{Z}_+$, let $\psi_n(u,\tau)_+ \triangleq \max\{0, \psi_n(u,\tau)\}$, and let the corresponding optimality function $\theta_{\mathbf{CMP}_n} : G_n \times [\tau_{\min}, \tau_{\max}] \to \mathbb{R}$ be defined by*

$$\begin{aligned}\theta_{\mathbf{CMP}_n}(u,\tau) \triangleq \min_{(u',\tau')\in G_n\times[\tau_{\min},\tau_{\max}]} \Big\{ & \frac{1}{2}\|u'-u\|_2^2 + \frac{1}{2}|\tau'-\tau|^2 \\ & + \max\{-\psi_n(u,\tau)_+ + \langle \nabla_u g_n^0(u,\tau), u'-u\rangle \\ & + \nabla_\tau g_n^0(u,\tau)(\tau'-\tau), g_n^j(u,\tau) - \psi(u,\tau)_+ \\ & + \langle \nabla_u g_n^j(u,\tau), u'-u\rangle \\ & + \nabla_\tau g_n^j(u,\tau)(\tau'-\tau), j \in \mathbf{q}\}\Big\}.\end{aligned} \tag{5.4}$$

Then

(i) *The optimality function $\theta_{\mathbf{CMP}}(\cdot,\cdot)$ is well defined and continuous.*

(ii) *If* $h_u(u,\tau), h_\tau(u,\tau)$ *are such that* $u + h_u(u,\tau) \in G_n, \tau + h_\tau(u,\tau) \in [\tau_{\min}, \tau_{\max}]$ *are a solution to* (5.4)*, then* $h_u(\cdot,\cdot), h_\tau(\cdot,\cdot)$ *are continuous functions.*

(iii) *Suppose that* $(\hat{u}_n, \hat{\tau}_n) \in G_n \times [\tau_{\min}, \tau_{\max}]$ *is an optimal solution to the problem* **CMP**. *Then* $\theta_{\mathbf{CMP}}(\hat{u}_n, \hat{\tau}_n) = 0$.

To simply notation, we define $H \stackrel{\Delta}{=} G \times [\tau_{\min}, \tau_{\max}]$, $H_n \stackrel{\Delta}{=} G_n \times [\tau_{\min}, \tau_{\max}]$, and $\eta = (u,\tau)$, and for any $\xi = (\xi_u, \xi_\tau) \in H$ and $\eta = (u,\tau) \in H$, we define $\langle \xi, \eta \rangle_H \stackrel{\Delta}{=} \langle \xi_u, u \rangle_2 + \xi_\tau \tau$, and $\|\eta\|_H \stackrel{\Delta}{=} (\|u\|_2^2 + |\tau|^2)^{1/2}$. Next, for any $\eta', \eta \in H$, we define

$$\hat{\psi}(\eta' - \eta|\eta) \stackrel{\Delta}{=} \max_{j \in \mathbf{q}} \{g^j(\eta) + \langle \nabla g^j(\eta), \eta' - \eta \rangle_H\} + \tfrac{1}{2}\|\eta' - \eta\|_H^2. \tag{5.5a}$$

Next, for any $n \in \mathbb{Z}_+, \eta', \eta \in H_n$, we define

$$\hat{\psi}_n(\eta' - \eta|\eta) \stackrel{\Delta}{=} \max_{j \in \mathbf{q}} \{g_n^j(\eta) + \langle \nabla g_n^j(\eta), \eta' - \eta \rangle_H\} + \tfrac{1}{2}\|\eta' - \eta\|_H^2. \tag{5.5b}$$

With this notation, we have that

$$\theta_{\mathbf{MMP}}(\eta) = \min_{\eta' \in H} \hat{\psi}(\eta' - \eta|\eta) - \psi(\eta), \tag{5.5c}$$

$$\theta_{\mathbf{MMP}_n}(\eta) = \min_{\eta' \in H_n} \hat{\psi}_n(\eta' - \eta|\eta) - \psi_n(\eta). \tag{5.5d}$$

LEMMA 5.4. *There exists a constant* $K_1 < \infty$ *such that for every* $\varepsilon > 0$*, there exists* n_ε *such that for all* $n \geq n_\varepsilon$*, and all* $\eta', \eta \in H_n$,

$$|\hat{\psi}_n(\eta' - \eta|\eta) - \hat{\psi}(\eta' - \eta|\eta)| \leq K_1 \varepsilon. \tag{5.6a}$$

Proof. It follows from Assumption 5.1 that there exists an $n_\varepsilon \in \mathbb{Z}_+$ such that for all $n \geq n_\varepsilon$,

$$\begin{aligned} \hat{\psi}_n(\eta' - \eta|\eta) - \hat{\psi}(\eta' - \eta|\eta)| &\leq \max_{j \in \mathbf{q}} \{g_n^j(\eta) - g^j(\eta) \\ &\qquad + \langle \nabla g_n^j(\eta) - \nabla g^j(\eta), \eta' - \eta \rangle_H\} \\ &\leq [1 + K_H]\varepsilon, \end{aligned} \tag{5.6b}$$

where $K_H = \max\{\|\eta' - \eta\|_H | \eta', \eta \in H\}$. Reversing the roles of $\hat{\psi}_n(\eta' - \eta|\eta)$ and $\hat{\psi}(n' - \eta|\eta)$ we get the desired result. □

THEOREM 5.5. *There exists a constant* $K_2 < \infty$ *such that for every* $\varepsilon \in (0, 1]$*, there exists* n_ε *such that for all* $n \geq n_\varepsilon$*, and all* $\eta \in H_n$,

$$|\theta_{\mathbf{MMP}_n}(\eta) - \theta_{\mathbf{MMP}}(\eta)| \leq K_2 \varepsilon. \tag{5.7}$$

Proof. For any $\eta \in H_n$, let

$$\xi(\eta) \stackrel{\Delta}{=} \arg\min_{\eta' \in H} \hat{\psi}(\eta' - \eta|\eta), \tag{5.8a}$$

$$\xi_n(\eta) \stackrel{\Delta}{=} \arg\min_{\eta' \in H_n} \hat{\psi}_n(\eta' - \eta|\eta), \tag{5.8b}$$

$$\tilde{\xi}_n(\eta) \triangleq \arg\min_{\eta' \in H_n} \|\eta' - \xi(\eta)\|_H. \tag{5.8c}$$

Let $\varepsilon > 0$ be given and let $n_\varepsilon \in \mathbb{Z}_+$ be defined as in Assumption 5.1(vi). In view of Assumption 5.1(v), there exists an $n'_\varepsilon \in \mathbb{Z}_+$, with $n'_\varepsilon \geq n_\varepsilon$, such that for any $n \geq n'_\varepsilon$ and any $\eta \in H$ there exists an $\eta_n \in H_n$ such that $\|\eta - \eta_n\|_H \leq \varepsilon$. Hence we obtain

$$\begin{aligned} \theta_{\mathbf{MMP}}(\eta) &\leq \hat{\psi}(\xi_n(\eta) - \eta|\eta) - \psi(\eta) \\ &\leq \hat{\psi}_n(\tilde{\xi}_n(\eta) - \eta|\eta) - \psi_n(\eta)] + [\psi_n(\eta) - \psi(\eta)] + K_1\varepsilon \\ &\leq \theta_{\mathbf{MMP}_n}(\eta) + [K_1 + 1]\varepsilon; \end{aligned} \tag{5.9a}$$

$$\begin{aligned} \theta_{\mathbf{MMP}_n}(\eta) &\leq \hat{\psi}_n(\tilde{\xi}_n(\eta) - \eta|\eta) - \Psi_n(\eta) \\ &\leq \hat{\psi}(\tilde{\xi}_n(\eta) - \eta|\eta) - \psi(\eta) + [\psi(\eta) - \psi_n(\eta)] + K_1\varepsilon \\ &\leq \hat{\psi}(\xi(\eta) - \eta|\eta) - \psi(\eta) + K'\|\xi(\eta) - \tilde{\xi}_n(\eta)\|_H \\ &\qquad + \frac{1}{2}|\|\xi(\eta) - \eta\|_H - \|\tilde{\xi}_n(\eta) - \eta\|_H|\|\xi(\eta) - \tilde{\xi}_n(\eta)\|_H \\ &\qquad + (K_1 + 1)\varepsilon \\ &\leq \theta_{\mathrm{MMP}}(\eta) + K_2\varepsilon, \end{aligned} \tag{5.9b}$$

where $K_2 = 1 + K_1 + K' + K''$, with $K' = \sup_{\eta \in H} \max_{j \in \mathbf{q}} \|\nabla g^j(\eta)\|_H$ and $K'' = \frac{1}{2}\sup_{\eta, \eta' \in H} \|\eta' - \eta\|_H$. The desired result now follows. □

The proof of the following result for problem **CMP** is quite similar to the one above and hence is omitted.

THEOREM 5.6. *There exists a constant $K_3 < \infty$ such that for every $\varepsilon > 0$, there exists n_ε such that for all $n \geq n_\varepsilon$, and all $\eta \in H_n$,*

$$|\theta_{\mathbf{CMP}_n}(\eta) - \theta_{\mathbf{CMP}}(\eta)| \leq K_3\varepsilon. \tag{5.10}$$

The problems **MMP** and **CMP** are finite-dimensional and hence can be solved with arbitrary precision using a finite-dimensional minimax or nonlinear programming algorithm, respectively, such as any one of the following [Pol.1], [Pol.3]. The first question we must answer is whether doing that is useful, i.e., we must establish whether our discretizations are consistent in an appropriate sense. The following pair of theorems gives an affirmative answer.

THEOREM 5.7. (i) *Suppose that $\{(\hat{u}_n, \hat{\tau}_n)\}_{n=1}^{\infty}$ is a sequence of optimal solutions to the sequence of problems $\mathbf{MMP}_n$. Let $I \subset \mathbb{Z}_+$ be such that $\hat{u}_n \xrightarrow{I} \hat{\sigma} \in \overline{G}$ (in the sense of control measures* (i.s.c.m.)*) and $\hat{\tau}_n \xrightarrow{I} \hat{\tau} \in [\tau_{\min}, \tau_{\max}]$, as $i \to \infty$, then $(\hat{\sigma}, \hat{\tau})$ is an optimal solution of $\overline{\mathbf{MMP}}$.*

(ii) *Suppose that $\{(U_n^*, \tau_n^*)\}_{n=1}^{\infty}$, with $u_n^* \in G_n$ and $\tau_n^* \in [\tau_{\min}, \tau_{\max}]$, is such that*

$$\theta_{\mathbf{MMP}_n}(u_n^*, \tau_n^*) \geq -\frac{1}{n}. \tag{5.11}$$

Let $I \subset \mathbb{Z}_+$ be such that $u_n^* \xrightarrow{I} \sigma^* \in \overline{G}$ (i.s.c.m.) *and $\tau_n^* \xrightarrow{I} \tau^* \in [\tau_{\min}, \tau_{\max}]$, as $n \to \infty$, then $\overline{\theta}_{\mathbf{MMP}}(\sigma^*, \tau^*) = 0$.*

Proof. (i) For the sake of contradiction, suppose that $(\hat{\sigma}, \hat{\tau})$ is not an optimal solution of $\overline{\mathbf{MMP}}$. Then there exists a pair (σ^{**}, τ^{**}), with $\sigma^{**} \in (\overline{G})$ and $\tau^{**} \in [\tau_{\min}, \tau_{\max}]$ such that $\overline{\psi}(\sigma^{**}, \tau^{**}) < \overline{\psi}(\hat{\sigma}, \hat{\tau})$. Since $\overline{\psi}(\cdot, \cdot)$ is continuous, and $\hat{u}_n \in G_n$ is an ordinary

control, we must have that $\overline{\psi}(\hat{u}_n, \hat{\tau}_n) \xrightarrow{I} \overline{\psi}(\hat{\sigma}, \hat{\tau})$. Hence, because of Assumption 5.1(vi), we must also have that $\psi_n(\hat{u}_n, \hat{\tau}_n) \xrightarrow{I} \overline{\psi}(\hat{\sigma}, \hat{\tau})$. Now, by Assumption 5.1(v), there exists a sequence $\{u'_n\}_{n \in I}$ such that $u'_n \xrightarrow{I} \sigma^{**}$ (i.s.c.m.), as $i \to \infty$. Hence because $\overline{\psi}(\cdot, \cdot)$ is continuous and because of Assumption 5.1(vi), $\psi_n(u'_n, \tau^{**}) \xrightarrow{I} \overline{\psi}(\sigma^{**}, \tau^{**})$ which, for n sufficiently high, contradicts the optimality of the pairs $(\hat{u}_n, \hat{\tau}_n)$.

(ii) This part follows directly from the continuity of the function $\overline{\theta}(\cdot, \cdot)$ and Theorem 5.5. □

We get a similar result for problem **CMP**, which we state without proof.

THEOREM 5.8. (i) *Suppose that* $\{(\hat{u}_n, \hat{\tau}_n)\}_{n=1}^{\infty}$ *is a sequence of optimal solutions to the sequence of problems* $\mathbf{CMP}_n$. *Let* $I \subset \mathbb{Z}_+$ *be such that* $\hat{u}_n \xrightarrow{I} \hat{\sigma} \in \overline{G}$ (i.s.c.m.) *and* $\hat{\tau}_n \xrightarrow{I} \hat{\tau} \in [\tau_{\min}, \tau_{\max}]$, *as* $n \to \infty$, *then* $(\hat{\sigma}, \hat{\tau})$ *is an optimal solution of* $\overline{\mathbf{CMP}}$.

(ii) *Suppose that* $\{(u_n^*, \tau_n^*)\}_{n=1}^{\infty}$, *with* $u_n^* \in G_n$ *and* $\tau_n^* \in [\tau_{\min}, \tau_{\max}]$ *is such that*

$$\theta_{\mathbf{MMP}_n}(u_n^*, \tau_n^*) \geq -\frac{1}{n}. \tag{5.12a}$$

$$\psi_n(u_n^*, \tau_n^*) \leq \frac{1}{n}. \tag{5.12b}$$

Let $I \subset \mathbb{Z}_+$ *be such that* $u_n^* \xrightarrow{I} \sigma^* \in \overline{G}$ (i.s.c.m.) *and* $\tau_n^* \xrightarrow{I} \tau^* \in [\tau_{\min}, \tau_{\max}]$, *as* $i \to \infty$, *then* $\overline{\theta}_{\mathbf{MMP}}(\sigma^*, \tau^*) = 0$ *and* $\overline{\psi}(\sigma^*, \tau^*) \leq 0$.

The computational scheme represented by Theorems 5.7 and 5.8 can be implemented as follows. An algorithm is applied to problem $\mathbf{MMP}_n$ (or $\mathbf{CMP}_n$), producing a sequence of iterates $(u_{n,i}, \tau_{n,i})$, $i = 0, 1, 2, \ldots, i_n$ which is arrested when (5.11) (or (5.12a) and (5.12b)) is satisfied. Then a new sequence, $(u_{n+1,i}, \tau_{n+1,i})$, $i = 0, 1, 2, \ldots$, is started for problem $\mathbf{MMP}_{n+1}$ (or $\mathbf{CMP}_{n+1}$), with $(u_{n+1,0}, \tau_{n+1,0}) = (u_{n,i_n}, \tau_{n,i_n})$. The main disadvantage of this scheme is that Theorems 5.7 and 5.8 deal only with a special subsequence of all the iterates computed, rather than with the whole sequence.

We will now show that it is possible to generalize the *algorithm implementation scheme* in [Kle.1] so as to obtain algorithms for solving **MMP** and **CMP**, with the property that *any* accumulation point of the computed sequence of iterates satisfies our first order optimality conditions. However, this requires that we strengthen Assumption 5.1(vi), as follows.

Assumption 5.9. There exists a constant $\hat{K} < \infty$ such that for all $n \in \mathbb{Z}_+, j = 0, 1, 2, \ldots, q$, all $u \in G_n$, and $\tau \in [\tau_{\min}, \tau_{\max}]$,

$$|g^j(u, \tau) - g_n^j(u, \tau)| \leq \frac{\hat{K}}{2^n}. \tag{5.13}$$

Referring to [Kle.1], we see that Assumption 5.9 is satisfied when *ordinary* differential equations are integrated numerically by a method of order at least one. It is shown in [Bak.1, §6.5], making use of the results in [Fuj.1], [Fuj.2], [Fuj.3], [Ode.1], that, when the finite element method is implemented using linear elements and Newmark's β method is used with $\beta = 0$, Assumption 5.9 is satisfied by the example treated in §6. We believe that it will also hold for many other cases as well.

For problem **MMP** we will extend a variant of the Pironneau–Polak–Pshenichnyi minimax algorithm (see [Pir.1], [Pol.1], [Psh.1]), which can be used for solving $\mathbf{MMP}_n$. To simplify proofs, we will use an exact line search step size rule; however, the results to follow remain valid also with an Armijo type step size rule (see [Pol.1], [Psh.1] for step size rule). To simplify exposition, we resume the notation $\eta = (u, \tau), H_n = G_n \times [\tau_{\min}, \tau_{\max}]$.

MINIMAX ALGORITHM 5.10 (solves MMP).

Parameter. $\gamma \in (0,1)$.

Data. $\eta_0, \eta_0 \in H_{n_0}$.

Step 0. Set $i = 0, n(0) = n_0$.

Step 1. Compute the search direction,

$$h_i = h_{n(i)}(\eta_i) \triangleq \arg \min_{\eta \in H_{n(i)}} \max_{j \in \mathbf{q}} \left\{ g^j_{n(i)}(\eta_i) + \langle \nabla g^j_{n(i)}(\eta_i), \eta - \eta_i \rangle_H + \frac{1}{2}\|\eta - \eta_i\|^2_H \right\} - \eta_i. \tag{5.14a}$$

Step 2. Compute the step size

$$\lambda_i \in \lambda_{n(i)}(\eta_i) \triangleq \arg \min_{\gamma \in [0,1]} \psi_{n(i)}(\eta_i + \lambda h_i). \tag{5.14b}$$

Step 3. Set $\eta^* = \eta_i + \lambda_i h_i$.

If

$$\psi_{n(i)}(\eta^*) - \psi_{n(i)}(\eta_i) > -\frac{1}{2^{\gamma n(i)}}, \tag{5.14c}$$

replace $n(i)$ by $n(i)+1$ and go to Step 1.

Else set $n(i+1) = n(i), \eta_{i+1} = \eta_i + \lambda_i h_i$.

Step 4. Replace i by $i+1$ and go to Step 1.

Note that (5.14c) causes the algorithm to increase precision when the decrease in cost becomes "unacceptably" small.

THEOREM 5.11. *Suppose that Algorithm* 5.10 *constructs a sequence* $\{\eta_i\}_{i=0}^{\infty}$. *Then this sequence has accumulation points in* $\overline{H} \triangleq \overline{G} \times [\tau_{\min}, \tau_{\max}]$, *and every such accumulation point,* $\hat{\eta}$, *satisfies* $\bar{\theta}_{\mathbf{MMP}}(\hat{\eta}) = 0$.

Proof. First we note that since $\overline{H}$ is sequentially compact, the sequence $\{\eta_i\}_{i=0}^{\infty}$ must have accumulation points in the relaxed controls topology. The rest of our proof is in three parts: (a) we will show that $n(i) \to \infty$ as $i \to \infty$, then (b) we will show that for any $\eta^* = (\sigma^*, \tau^*) \in \overline{H}$ such that $\bar{\theta}_{\mathbf{MMP}}(\eta^*) < 0$, there exists an integer n^* and a $\delta^* > 0$, such that for all $n(i) \geq n^*$, if $\eta_i \in H_{n(i)}$ is sufficiently close to η^*, then $\psi_{n(i)}(\eta_{i+1}) - \psi_{n(i)}(\eta_i) \leq -\delta^*$, and (c) we will obtain a contradiction by showing that if the theorem is not true, then $\psi(\eta_i) \to -\infty$.

(a) Suppose that there exists integers i_0 and n_0 such that for all $i \geq i_0, n(i) = n_0$. Then we must have that $\psi_{n_0}(\eta_{i+1}) - \psi_{n_0}(\eta_i) \leq -1/2^{\gamma n_0}$ for all $i \geq i_0$, which implies that $\Psi_{n_0}(\eta_i) \to -\infty$ as $i \to \infty$. Since H_{n_0} is compact, this is impossible, and hence we conclude that $n(i) \to \infty$ as $i \to \infty$.

(b) Suppose that for $n \in \mathbb{Z}_+, \eta_n \in H_n$ and that $\eta_n \xrightarrow{\text{i.s.c.m.}} \eta^* \in \overline{H}$, as $n \to \infty$. Furthermore, suppose that $\bar{\theta}_{\mathbf{MMP}}(\eta^*) \triangleq -8\delta^* < 0$. Since $\bar{\theta}_{\mathbf{MMP}}(\cdot)$ is continuous, and since $\theta_{\mathbf{MMP}}(\eta_n) = \bar{\theta}_{\mathbf{MMP}}(\eta_n)$ for all $n \in \mathbb{Z}_+$, it follows that there is an integer n_0 such that for all $n \in \mathbb{Z}_+, n \geq n_0, \theta_{\mathbf{MMP}}(\eta_n) \leq -4\delta^*$. It now follows from Theorem 5.5 that there exists an integer $n_1 \geq n_0$ such that for all $n \in \mathbb{Z}_+, n \geq n_1, \theta_{\mathbf{MMP}_n}(\eta_n) \leq -2\delta^*$. Hence, for all $n \geq n_1$ and $\lambda \in [0,1]$,

$$\begin{aligned}
&\psi_n(\eta_n + \lambda h_n(\eta_n)) - \psi_n(\eta_n) \\
&\quad = \max_{j \in \mathbf{q}} \{ g^j_n(\eta_n) - \psi_n(\eta_n) + \lambda \langle \nabla g^j_n(\eta_n), h_n(\eta_n) \rangle_H + \tfrac{1}{2}\|h_n(\eta_n)\|^2_H \\
&\qquad + \lambda \textstyle\int_0^1 \langle \nabla g^j_n(\eta_n + s\lambda h_n(\eta_n)) - \nabla g^j_n(\eta_n), h_n(\eta_n) \rangle_H \\
&\qquad - \tfrac{1}{2}\|h_n(\eta_n)\|^2_H \} \\
&\quad \leq \lambda [\theta_{\mathbf{MMP}_n}(\eta_n) + \lambda L \|h_n(\eta_n)\|^2_H],
\end{aligned} \tag{5.15a}$$

where L is as in (5.2b). Since the sets G_n are uniformly bounded, there exists a $b < \infty$ such that $\|h_n(\eta_i)\|_H \leq b$ for all $n \in \mathbb{Z}_+$. Hence it follows from (5.15a) that there exists a $\hat{\lambda} \in (0, 1]$, such that for all $n \geq n_1$,

$$\psi_n(\eta_n + \lambda_n(\eta_n)h_n(\eta_n)) - \psi_n(\eta_n) \leq \psi_n(\eta_n + \hat{\lambda}h_n(\eta_n)) - \psi_n(\eta_n) \leq -\hat{\lambda}\delta^*, \tag{5.15b}$$

which completes the second part of our proof.

(c) Now, by construction, we have that $\psi_{n(i)}(\eta_{i+1}) - \psi_{n(i)}(\eta_i) \leq -1/2^{\gamma n(i)}$, and hence, making use of Assumption 5.9,

$$\psi(\eta_{i+1}) - \psi(\eta_i) \leq -\frac{1}{2^{n(i)}}(2^{(1-\gamma)n(i)} - \hat{K}). \tag{5.15c}$$

Hence, since $n(i) \to \infty$ as $i \to \infty$, there exists an i_o such that for all $i \geq i_o, \psi(\eta_{i+1}) - \psi(\eta_i) \leq 0$.

Now, for the sake of contradiction, suppose that the sequence $\{\eta_i\}_{i=0}^{\infty}$ has an accumulation point $\hat{\eta} \in \overline{H}$ such that $\overline{\theta}_{\mathbf{MMP}}(\hat{\eta}) < 0$. Then there exists an infinite subset I of the positive integers such that $\eta_i \overset{I}{\to} \hat{\eta}$ (i.s.c.m.) as $i \to \infty$, and hence because $\overline{\psi}(\cdot)$ is continuous and $\psi(\eta_i) = \overline{\psi}(\eta_i), \psi(\eta_i) \overset{I}{\to} \hat{\eta}$ as $i \to \infty$. Now, $\{\psi(\eta_i)\}_{i=i_o}^{\infty}$ is monotone decreasing, and hence we conclude that $\psi(\eta_i) \to \overline{\psi}(\hat{\eta})$ as $i \to \infty$. Since $n(i) \to \infty$, it follows from (b) that there exist a $\hat{\delta} > 0$ and an integer i_1, such that for all $i \geq i_1, i \in I, \psi_{n(i)}(\eta_{i+1}) - \psi_{n(i)}(\eta_i) \leq -\hat{\delta} < 0$. Hence, for all $i \in I$,

$$\psi(\eta_{i+1}) - \psi_{n(i)}(\eta_i) \leq -\hat{\delta} + \frac{\hat{K}}{2^{n(i)}}. \tag{5.15d}$$

Since $n(i) \to \infty$ as $i \to \infty$, (5.15d) contradicts the convergence of the sequence $\{\psi(\eta_i)\}_{i=0}^{\infty}$. This completes our proof. □

Two observations are in order at this point. First, it follows from (5.15c) that the cost sequence is eventually monotone decreasing. Since it is bounded, it must converge. Second, it can be deduced from the above proof that $\theta_{\mathbf{MMP}n(i)}(\eta_i) \to 0$, which implies in turn that $h_{n(i)} \to 0$. Hence, referring to Theorem 1.3.66 in [Pol.3], we conclude that if $\psi(\cdot)$ has only a finite number of stationary points, then the sequence of trajectories $\{x^{\eta_i}\}$ must converge. Furthermore, if $\{\eta_i\}$ has an accumulation point in the H topology, then the entire sequence $\{\eta_i\}$ must converge to that point.

For problem **CMP**, we will extend the unified phase I–phase II method of feasible directions, using an Armijo step size rule, described in [Pol.2].

ALGORITHM 5.12

Parameters. $\gamma > 0, \alpha, \beta \in (0, 1)$.

Data. $n_0, \eta_0 \in H_{n_0}$.

Step 0. Set $i = 0, n(0) = n_0$.

Step 1. Compute the value of the *optimality function* $\theta_i = \theta_{\mathbf{CMP}_{n(i)}}(\eta_i)$, and the corresponding *search direction* $h_i = h_{n(i)}(\eta_i)$, where

$$\begin{aligned}\theta_{\mathbf{CMP}_{n(i)}}(\eta_i) \triangleq \min_{\eta \in H_{n(i)}} \Big\{ &\tfrac{1}{2}\|\eta - \eta_i\|_H^2 + \max\{-\psi_{n(i)}(\eta_i)_+ + \langle \nabla g^0_{n(i)}(\eta_i), \eta - \eta_i\rangle_H, \\ &g^j_{n(i)}(\eta_i) - \psi(\eta_i)_+ + \langle \nabla g^j_n(\eta_i), \eta - \eta_i\rangle_H, j \in \mathbf{q}\}\Big\}.\end{aligned} \tag{5.16a}$$

$$h_{n(i)}(\eta_i) \triangleq \arg \min_{\eta \in H_{n(i)}} \Big\{ \tfrac{1}{2}\|\eta - \eta_i\|_H^2 + \max\{-\psi_{n(i)}(\eta_i)_+ + \langle \nabla g^0_{n(i)}(\eta_i), \eta - \eta_i\rangle_H, g^j_{n(i)}(\eta_i) - \psi(\eta_i)_+ + \langle \nabla g^j_n(\eta_i), \eta - \eta_i\rangle_H, j \in \mathbf{q}\} \Big\} - \eta_i. \tag{5.16b}$$

Step 2. Compute the *step size* λ_i:

$$\lambda_i = \max\{\beta^k | k \in \mathbb{N}, F_{n(i)}(\eta_i + \beta^k h_i | \eta_i) \leq \beta^k \alpha \theta_i\}, \tag{5.16c}$$

where, for $n \in \mathbb{Z}_+, \eta, \eta* \in H_n$,

$$F_n(\eta|\eta^*) \triangleq \max\{g^0_n(\eta) - g^0_n(\eta^*) - \psi_n(\eta^*)_+, \psi_n(\eta) - \psi_n(\eta^*)_+\}. \tag{5.16d}$$

Step 3. Set $\eta^* = \eta_i + \lambda_i h_i$.
If

$$F_{n(i)}(\eta^*|\eta_i) > -\frac{1}{2^{\gamma n(i)}}. \tag{5.16e}$$

Replace $n(i)$ by $n(i)+1$, and go to Step 1. Else set $n(i+1) = n(i), \eta_{i+1} = \eta_i + \lambda_i h_i$.

Step 4. Replace i by $i+1$ and go to Step 1.

THEOREM 5.13. *Suppose that* (i) *for every* $n \in \overline{H}$ *such that* $\overline{\psi}(\eta) \geq 0, \overline{\theta}_{\mathbf{MMP}}(\eta) < 0$; *and* (ii) *for every* $n \in \mathbb{Z}_+$ *and every* $\eta \in H_n$ *such that* $\psi_n(\eta) \geq 0, \theta_{\mathbf{MMP}_n}(\eta) < 0$. *If Algorithm* 5.12 *constructs a sequence* $\{\eta_i\}_{i=0}^{\infty}$, *then this sequence has accumulation points in* $\overline{H} \triangleq \overline{G} \times [\tau_{\min}, \tau_{\max}]$, *and every such accumulation point,* η, *satisfies* $\overline{\psi}(\hat{\eta}) \leq 0, \overline{\theta}_{\mathbf{CMP}}(\hat{\eta}) = 0$.

Proof. First we note that since $\overline{H}$ is sequentially compact, the sequence $\{\eta_i\}_{i=0}^{\infty}$ must have accumulation points in the relaxed controls topology. The rest of our proof is in three parts: (a) we will show that $n(i) \to \infty$ as $i \to \infty$, then (b) we will show that for any $\eta^* = (\sigma^*, \tau^*) \in \overline{H}$ such that $\overline{\theta}_{\mathbf{CMP}}(\eta^*) < 0$, there exists an integer n^* and a $\delta^* > 0$, such that for all $n(i) \geq n^*$, and any η_i sufficiently close to $\eta^*, F_{n(i)}(\eta_{i+1}|\eta_i) \leq -\delta^*$, and (c) we will obtain a contradiction by showing that if the theorem is not true, then either $\psi(\eta_i) \to -\infty$ as $i \to \infty$ or $g^0(\eta_i) \to -\infty$ as $i \to \infty$.

(a) Suppose that there is a finite integer n^* such that $n(i) = n^*$ for all $i \geq i^*$, with $i^* < \infty$. Then the test (5.16e) fails to be satisfied for all $i \geq i^*$, and hence $F_{n^*}(\eta_{i+1}|\eta_i) \leq -(1/2^{n^*})^\gamma$ for all $i \geq i^*$. Without loss of generality, suppose that $\psi_{n^*}(\eta_{i^*}) \geq 0$ and that $\psi_{n^*}(\eta_i) > 1/n^*$ for all $i \geq i^*$. Then $\psi_{n^*}(\eta_{i+1}) - \psi_{n^*}(\eta_i)_+ \leq -(1/2^{n^*})^\gamma$ for all $i \geq i^*$, and hence there must exist an i_0 such that $\psi_{n^*}(\eta_i) \leq 0$ for all $i \geq i_0$. Furthermore, for $i \geq i_0$, we must also have that $g^0_{n^*}(\eta_{i+1}) - g^0_{n^*}(\eta_i) \leq -(1/2^{n^*})^\gamma$, which implies that $g^0_{n^*}(\eta_i) \to -\infty$ as $i \to \infty$. However, since H_n is compact and $g^0_{n^*}(\cdot)$ is continuous, this is clearly impossible, and we have a contradiction. Hence we must have that $n(i) \to \infty$ as $i \to \infty$.

(b) The proof of this part is quite similar to that of part (b) in the proof of Theorem 5.13, and is therefore omitted.

(c) For any $\eta, \eta^* \in H$, let $F_n(\eta|\eta^*)$ be defined by

$$F(\eta|\eta^*) \triangleq \max\{g^0(\eta) - g^0(\eta^*) - \psi(\eta^*)_+, \psi(\eta) - \psi(\eta^*)_+\}. \tag{5.17a}$$

Then, because of the test (5.16e) and Assumption 5.9, we have that for all $i \in \mathbb{Z}_+$,

$$F(\eta_{i+1}|\eta_i) \leq -\frac{1}{2^{n(i)}}(2^{(1-\gamma)n(i)} - 2\hat{K}). \tag{5.17b}$$

Since $\gamma \in (0,1)$ and $n(i) \to \infty$ as $i \to \infty$, it follows that there is an $i_1 \in \mathbb{Z}_+$ such that

$$F(\eta_{i+1}|\eta_i) \leq 0, \qquad \forall i \geq i_1, \tag{5.17c}$$

and hence, for all $i \geq i_1$,

$$\psi(\eta_{i+1}) - \psi(\eta_i)_+ \leq 0, \tag{5.17d}$$

and

$$g^0(\eta_{i+1}) - g^0(\eta_i) - \psi(\eta_i)_+ \leq 0. \tag{5.17e}$$

Now suppose that $\eta_i \overset{I}{\to} \hat{\eta} \in \overline{H}$ (i.s.c.m.) as $i \to \infty$ and that $\overline{\theta}(\hat{\eta}) < 0$. We distinguish between two possibilities:

(i) $\psi(\eta_i) > 0$ for all $i \geq i_1$. Then, by (5.17d), $\{\psi(\eta_i)\}_{i=i_1}^{\infty}$ is a monotone decreasing sequence, and, since by continuity $\psi(\eta_i) \overset{I}{\to} \overline{\psi}(\hat{\eta})$ as $i \to \infty$, it follows that $\psi(\eta_i) \to \overline{\psi}(\hat{\eta})$ as $i \to \infty$. It now follows from (b) and Assumption 5.9 that there exist a $\hat{\delta}$ and an $i_2 \geq i_1$ such that for all $i \in I, i \geq i_2$,

$$\psi(\eta_{i+1}) - \psi(\eta_i) \leq F_{n(i)}(\eta_{i+1}|\eta_i) + \frac{2\hat{K}}{2^{n(i)}} \leq -\hat{\delta} + \frac{2\hat{K}}{2^{n(i)}}, \tag{5.17f}$$

which contradicts the fact that $\psi(\eta_i) \to \overline{\psi}(\hat{\eta})$ as $i \to \infty$. Hence we must have that $\overline{\theta}(\hat{\eta}) = 0$, and hence, by assumption, that $\overline{\psi}(\hat{\eta}) \leq 0$ also holds.

(ii) There exists an $i_3 \geq i_1$ such that $\psi(\eta_{i_3}) \leq 0$. Then it follows from (5.17d) that $\psi(\eta_i) \leq 0$ for all $i \geq i_3$. Next, by (5.17e), $\{g^0(\eta_i)\}_{i=i_3}^{\infty}$ is a monotone decreasing sequence, and, since by continuity $g^0(\eta_i) \overset{I}{\to} \overline{g}^0(\hat{\eta})$ as $i \to \infty$, it follows that $g^0(\eta_i) \to \overline{g}^0(\eta^*)$ as $i \to \infty$, (i.s.c.m.). It now follows again from (b) and Assumption 5.9 that there exists an $i_4 \geq i_1$ such that for all $i \in I, i \geq i_4$,

$$g^0(\eta_{i+1}) - g^0(\eta_i) \leq F_{n(i)}(\eta_{i+1}|\eta_i) \leq -\hat{\delta} + \frac{2\hat{K}}{2^{n(i)}} \leq -\hat{\delta}/2, \tag{5.17g}$$

which contradicts the fact that $g^0(\eta_i) \to \overline{g}^0(\eta^*)$ as $i \to \infty$. Hence we must have that $\overline{\theta}(\eta^*) = 0$, which completes our proof. □

Again we can make some observations. First, it follows from (5.17f) that if the tail of the sequence $\{\eta_i\}$ is infeasible, then the constraint violation function $\psi(\cdot)$ eventually decreases monotonically to zero. In this case, making use of (5.17b), we can conclude that either the cost sequence $\{g^0(\eta_i)\}$ converges, or it has infinitely many accumulation points, a rather unlikely event. If the tail of the sequence $\{\eta_i\}$ is feasible, then the the tail of the cost sequence is monotone decreasing, and hence, since it is bounded, it converges. Second, it can be deduced from the above proof that $\theta_{\mathbf{CMP}n(i)}(\eta_i) \to 0$, which implies in turn that $h_{n(i)} \to 0$. Hence, referring to Theorem 1.3.66 in [Pol.3], we conclude that if $\overline{\theta}_{\mathbf{CMP}}(\cdot)$ has only a finite number of zeros, then the trajectory sequence $\{x^{\eta_i}\}$ must converge. Furthermore, if $\{\eta_i\}$ has an accumulation point in the H topology, then the entire sequence $\{\eta_i\}$ must converge to that point.

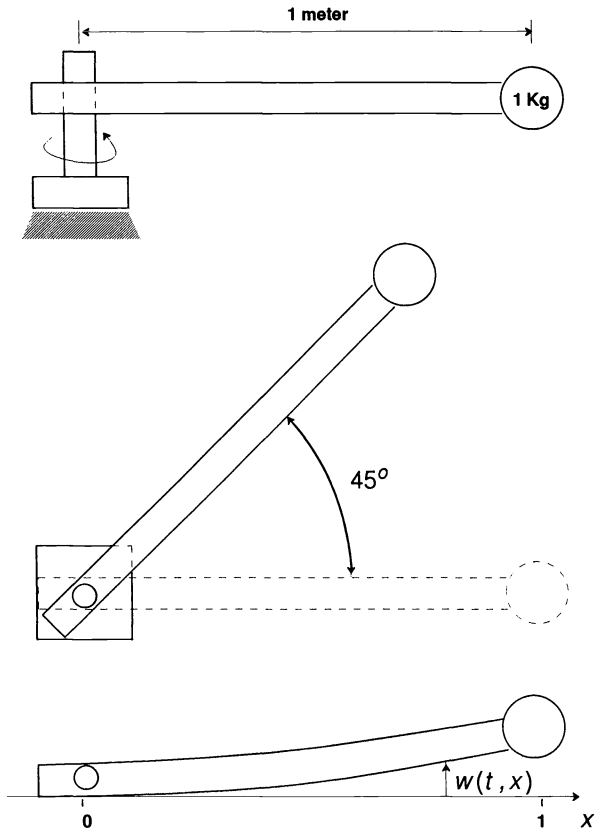


FIG. 1. *Configuration of slewing experiment.*

6. Computational results. We carried out three computational experiments involving the slewing motion of the hollow aluminum tube depicted in Fig. 1. The tube is one meter long, has a cross sectional radius of 1.0 cm, and a thickness of 1.6 mm. Attached to one end of the tube is a mass of 1 kg, and attached to the other end is a shaft connected to a motor. To reduce the computational burden, we neglected small nonlinear terms, the coupling between the flexural and extensional vibrations, and assumed that the acceleration can be controlled, instead of assigning a mass to the shaft and assuming that the torque is controlled. These simplifications lead to a model in the form of the standard Euler–Bernoulli tube with Kelvin–Voigt viscoelastic damping:

$$\begin{aligned} &mw_{tt}(t,x) + CIw_{txxxx}(t,x) + EIw_{xxxx}(t,x) - m\Omega^2(t)w(t,x) \\ &\quad = -\frac{m}{1+m/3}u(t)x, \qquad t \in [0,\tau], \quad x \in [0,1], \end{aligned} \tag{6.1a}$$

with boundary conditions:

$$w(t,0) = 0, \quad w_x(t,0) = 0, \quad CIw_{txx}(t,1) + EIw_{xx}(t,1) = 0, \quad t \in [0,\tau], \tag{6.1b}$$

$$\Omega^2(t)w(t,1) - w_{tt}(t,1) - u(t) - CIw_{txxx}(t,1) - EIw_{xxx}(t,1) = 0, \quad t \in [0,\tau], \tag{6.1c}$$

(6.1d) $$\Theta_t(t) = \Omega(t), \quad t \in [0, \tau], \quad \Omega_t(t) = u(t), \quad t \in [0, \tau],$$

where $w(t, x)$ is the displacement of the tube from the *shadow tube* (which remains undeformed during the motion) due to bending as a function of time and distance along the tube; $u(t)$ is the acceleration produced by the motor, and $\Omega(t)$ is the resulting angular velocity (in radians per second), and $\Theta(t)$ is the angular displacement of the rigid body (in radians). The values for the parameters in (6.1a)–(6.1c) were chosen to be $m = .2815$ kg/m, $I = 1.005 \times 10^{-8} m^4, C = 6.89 \times 10^7$ pascals/sec.; $E = 6.89 \times 10^9$ pascals, as given in the CRC Handbook of Material Science. The tube is very lightly damped (0.1 percent).

When time is normalized to the interval [0, 1], the dynamics become:

(6.2a) $$\begin{aligned} &mw_{tt}(t,x) + \tau CIw_{txxxx}(t,x) + \tau^2 EIw_{xxxx}(t,w) - \tau^2 m\Omega^2(t)w(t,x) \\ &\quad = -\tau^2 \frac{m}{1+m/3} u(t)x, \qquad t \in [0,1], \quad x \in [0,1], \end{aligned}$$

with boundary conditions:

(6.2b) $$w(t,0) = 0, \quad w_x(t,0) = 0, \quad CIw_{txx}(t,1) + \tau EIw_{xx}(t,1) = 0, \quad t \in [0,1],$$

(6.2c) $$\begin{aligned} &\tau^2\Omega^2(t)w(t,1) - w_{tt}(t,1) - \tau^2 u(t) - \tau CIw_{txxx}(t,1) \\ &\quad -\tau^2 EIw_{xxx}(t,1)ad = 0, \quad t \in [0,1], \end{aligned}$$

(6.2d) $$\Theta_t(t) = \tau\Omega(t), \quad t \in [0,1], \quad \Omega_t(t) = \tau u(t), \quad t \in [0,1].$$

To transcribe these dynamics into the standard form (2.2a), we proceed as follows. First we define $\zeta(t) \in X \triangleq L_2([0,1]) \times \mathbb{R}$, and $\Phi : X \times \mathbb{R}^2 \to X$ by

(6.2e) $$\begin{aligned} \zeta(t) &\triangleq \begin{bmatrix} w(t,x) \\ w(t,1) \end{bmatrix}, \\ \Phi(\zeta(t), u(t), \Omega(t)) &\triangleq \tau^2 \begin{bmatrix} \Omega^2(t)w(t,x) - u(t)x/(1+m/3) \\ \Omega^2(t)w(t,1) - u(t) \end{bmatrix}. \end{aligned}$$

Next we define the operators A_1 and Q, and their respective domains $D(A_1)$ and $D(Q)$ as follows:

(6.2f) $$\begin{aligned} D(A_1) \triangleq \Big\{ \zeta = \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \in X \,|\, &\zeta_{1xxxx} \in L_2([0,1]), \zeta_1(0) \\ &= \zeta_{1x}(0) = \zeta_{1xx}(1) = 0, \zeta_1(1) = \zeta_2 \Big\}, \end{aligned}$$

$A_1 : D(A_1) \to X$ is defined by

(6.2g) $$A_1 \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} = \tau^2 \begin{bmatrix} \frac{\mathrm{EI}}{\mathrm{m}} \zeta_{1xxxx}(\cdot) \\ \mathrm{EI}\zeta_{1xxx}(1) \end{bmatrix},$$

and with $D(Q) \triangleq D(A_1), Q : D(Q) \to X$ is defined by $Q \triangleq C/\tau E A_1$. Then (6.2a)–(6.2c) can be written in the form

(6.2h) $$\zeta_{tt} + Q\zeta_t + A_1\zeta = \Phi(\zeta, u, \Omega).$$

It is shown in §6.4 and Appendix II in [Bak.1], that Φ is a an operator that is Lipschitz continuous on bounded sets, and that A_1 and Q satisfy the assumptions in [Gib.1] needed

to derive the infinitesimal generator of a contraction semigroup. We give a brief outline of this derivation; see [Gib.1] for the details. First, we define the space $V \triangleq D(A_1^{1/2}) \times X$, so that if $y = (y_1, y_2) \in V$, then

$$\|y\|^2 = \langle y_1, A_1 y_1 \rangle + \langle y_2, y_2 \rangle, \tag{6.2i}$$

where $\langle \cdot, \cdot \rangle$ is the L_2 inner product. For any given $t \in [0, 1]$, let $\nu(t) \in V$ be defined by $v(t) \triangleq (w(t,x), w(t,1), w_t(t,x), w_t(t,1))$, and let the operator $A_2 : D(A_2) \to V$, where $D(A_2) = D(A_1) \times D(A_1) \subset V$, be defined by

$$A_2\nu(t) \triangleq \begin{bmatrix} 0 & I \\ -A_1 & -Q \end{bmatrix} \nu(t) = \begin{bmatrix} w_t(t,x) \\ w_t(t,1) \\ -\tau \frac{CI}{w} w_{txxxx}(t,x) - \tau^2 \frac{EI}{m} w_{xxxx}(t,x) \\ -\tau CI w_{txxx}(t,1) - \tau^2 EI w_{xxx}(t,1) \end{bmatrix}. \tag{6.2j}$$

It is shown in §2 in [Gib.1], that there exists a unique maximal dissipative extension of A_2 to A_3, where A_3 is the generator of a contraction semigroup that represents the free response of the system (6.2h). It is shown in [Sho.1] that A_3 generates an analytic semigroup. The standard form (2.2a) is then obtained by defining the state by $z(t) \triangleq (\nu(t), \Theta, \Omega) \in V \times \mathbb{R}^2$, and

$$A \triangleq \begin{bmatrix} A_3 & 0 & 0 \\ 0 & 0 & \tau \\ 0 & 0 & 0 \end{bmatrix}, \quad F(z(t), u(t)) \triangleq \begin{bmatrix} 0 \\ 0 \\ \Phi(\zeta(t), u(t), \Omega(t)) \\ 0 \\ \tau u(t) \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ \tau^2 \Omega^2(t) w(t,x) - \tau^2 u(t) x/(1 + m/3) \\ \tau^2 \Omega^2(t) w(t,1) - \tau^2 u(t) \\ 0 \\ \tau u(t) \end{bmatrix}. \tag{6.2k}$$

It follows that A satisfies Assumption 3.2 and that F satisfies Assumption 3.1.

We considered three slewing problems that shared two requirements: (a) The tube had to be rotated 45 degrees, from rest[5] to rest, and (b) the acceleration produced by the motor was limited to five rads/sec^2. The first problem, $\mathbf{P_1}$, was a minimum time problem, subject to the above constraints; the second problem, $\mathbf{P_2}$, was a minimum energy problem, subject to the above constraints and an upper bound on the time allowed; and the last problem, $\mathbf{P_3}$, was a minimum time problem, subject to the above constraints and an upper bound on the potential energy due to deformation of the tube throughout the entire maneuver (i.e., a worst case deformation constraint).

The transcription of the problems $\mathbf{P_1}$, $\mathbf{P_2}$, and $\mathbf{P_3}$ into the form (2.3b) required the introduction of the following functions. With τ denoting the final time, let

$$g^1(u, \tau) \triangleq \tau. \tag{6.3}$$

The energy consumed by the maneuver is given by

$$g^2(u, \tau) \triangleq \int_0^1 u(t)^2 dt. \tag{6.4}$$

[5] We say that *the tube is at rest* when the total energy of the tube is zero. This energy is composed of the energy due to rigid body motion and energy due to vibration and deformation.

The angular error at the final time is measured by

$$g^3(u,\tau) \triangleq (\Theta(1) - \pi/4)^2. \tag{6.5}$$

The rigid body energy at final time is given by

$$g^4(u,\tau) \triangleq \Omega(1)^2. \tag{6.6}$$

The kinetic energy due to vibration of the tube at time τ is given by

$$g^5(u,\tau) \triangleq \frac{m}{2}\int_0^1 w_t(t,x)^2 dx, \tag{6.7}$$

and the potential energy due to deformation of the tube at time τ is given by

$$g^6(u,\tau) \triangleq P(\tau,u) = \frac{EI}{2}\int_0^1 w_{xx}(\tau,x)^2 dx. \tag{6.8}$$

We see that the tube is at rest when $g^4(u,\tau) = g^5(u,\tau) = g^6(u,\tau) = 0$.

The deformation constraint for problem $\mathbf{P_3}$ has the form $P(t,u) \leq f(t)$ for all $t \in [0,1]$, where $f(\cdot)$ is a given positive bound function. This is a *state-space constraint.* To reduce the computational burden, we replaced it by the equivalent requirement $g^7(u,\tau) \leq 0$, where

$$g^7(u,\tau) \triangleq \int_0^1 [\max\{P(t,u) - f(t), 0\}]^2 dt. \tag{6.9}$$

Since $P(t,u)$ is continuous, $g^7(u,\tau) = 0$ if and only if $P(t,u) \leq f(t)$ for all $t \in [0,\tau]$. Transformations such as (6.9) must be used with great care because for any feasible pair $(u,\tau), g^7(u,\tau) = 0$ and $\nabla g^7(u,\tau) = 0$, and hence $\theta(u,\tau) = 0$, which causes our algorithm to stop up at such a pair. However, the problems caused by this violation can be circumvented by initializing the algorithm with an infeasible point, keeping the parameter γ, in Algorithm 5.10, small, and introducing an ε into the problem statement, as shown below.

It can be shown that all the above functions $g^j : G \times [0,\tau] \to \mathbb{R}$ are continuously differentiable (in the $L_2[0,1] \times \mathbb{R}$ topology) in u and t for all $j \in \{1,2,\ldots,7\}$. To conform with the format of problem (2.3b), we relax each of the equality constraints by a small amount. The three problems now acquire the following mathematical form,[6] where $G \triangleq \{u \in L_2[0,1] \mid |u(t)| \leq 1 \text{ for all } t \in [0,1]\}$ and $\mathbf{T} = [\tau_0, \tau_f]$, with $\tau_0 > 0$ very small and $\tau_f < \infty$ very large.

$$\begin{aligned}\mathbf{P_1} : \min\{g^1(u,\tau) \mid & g^3(u,\tau) - \varepsilon \leq 0, g^4(u,\tau) - \varepsilon \leq 0, g^5(u,\tau) - \varepsilon \leq 0,\\ & g^6(u,\tau) - \varepsilon \leq 0, (u,\tau) \in G \times \mathbf{T}\}.\end{aligned} \tag{6.10a}$$

$$\begin{aligned}\mathbf{P_2} : \min\{g^2(u,\tau) \mid & g^1(u,\tau) - \tau_f \leq 0, g^3(u,\tau) - \varepsilon \leq 0, g^4(u,\tau) - \varepsilon \leq 0,\\ & g^5(u,\tau) - \varepsilon \leq 0, g^6(u,\tau) - \varepsilon \leq 0, (u,\tau) \in G \times \mathbf{T}\}.\end{aligned} \tag{6.10b}$$

$$\begin{aligned}\mathbf{P_3} : \min\{g^1(u,\tau) \mid & g^3(u,\tau) - \varepsilon \leq 0, g^4(u,\tau) - \varepsilon \leq 0, g^5(u,\tau) - \varepsilon \leq 0,\\ & g^6(u,\tau) - \varepsilon \leq 0, g^7(u,\tau) - \varepsilon \leq 0, (u,\tau) \in G \times \mathbf{T}\}.\end{aligned} \tag{6.10c}$$

[6] Note that we find it convenient at this point to abandon the convention that the cost function is $g^0(\cdot,\cdot)$ as well as the linear numbering of the constraints.

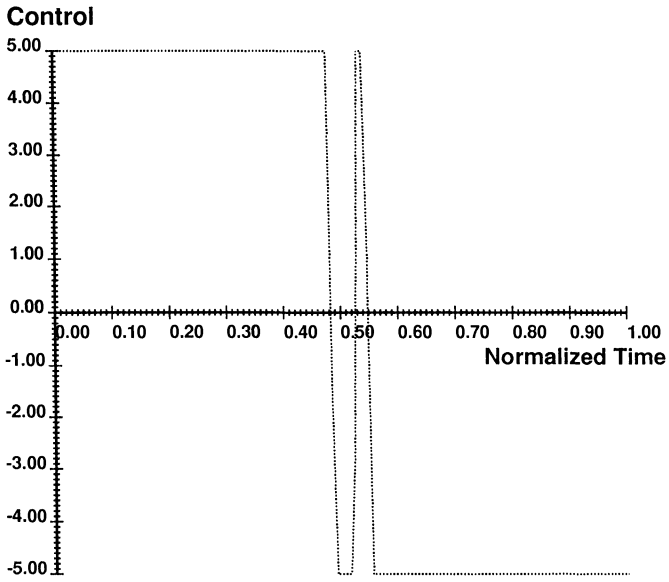


FIG. 2. *Final control for Problem* 1.

In our experiments, we set $\varepsilon = 10^{-4}$. Thus, with this relaxation, we are requiring that the final value of the angle Θ be in the interval $[45° - 0.5°, 45° + 0.5°]$. We assume that because of model simplifications and other inevitable modeling errors, a linear feedback system would be used to assure final pointing accuracy.

In the computational experiments reported in this paper, the term $\Omega^2(t)$ was neglected in (6.1a)–(6.1c). Similar results were obtained in computational experiments in which the term $\Omega^2(t)$ was kept. We used a cubic Hermit spline implementation of the Finite Element Method for spatial discretization and Newmark's β-method, with $\beta = 0$, for temporal discretization of both responses and sensitivities.[7] This approach is quite stable and gives accurate simulations. The results of our computational experiments are shown in Figs. 2–11.

PROBLEM $\mathbf{P_1}$. For simplicity, we chose the zero function as the initial control and 2 for an initial value for the maneuver time. The initial discretization consisted of 32 time steps and six finite elements. The discretization was refined at iterations 67, 99, and 123. Figure 2 is a graph of the control after 150 iterations. At this point, the number of time steps was 256 and the number of finite elements 48. Figure 3(a) is a graph of $\psi_{q_s,q_t}(u,\tau)$ as a function of the iteration number. Figure 3(b) shows $\psi_{q_s,q_t}(u,\tau)$ for the first 15 iterations. After precision refinement, the algorithm finds a control $u \in G_{q_t}$ and final time $\tau \in \mathbf{T}$ such that $\psi_{q_s,q_t}(u,\tau) < 0$ in only a few additional iterations. Note that each time precision of discretization was increased, the value of $\psi_{q_s,q_t}(u_i,\tau_i)$ increases. This is due to improvement in the accuracy of the evaluation of the partial differential equation. This increase in constraint violation $\psi_{q_s,q_t(u_i,\tau_i)}$ decreases each time the discretization is increased and we see that in the limit the increase is zero. Figure 4 is the graph of the cost as a function of iteration number. Figure 5 is the graph of $w(t,1)$, the displacement of the tip of the tube, from the *shadow tube*, as a function of time. The maximum displacement

[7] See [Bak.1, Chap 8] for implementation details, that are based on the results in [Fuj.1], [Fuj.2], [Fuj.3], [Ode.1].

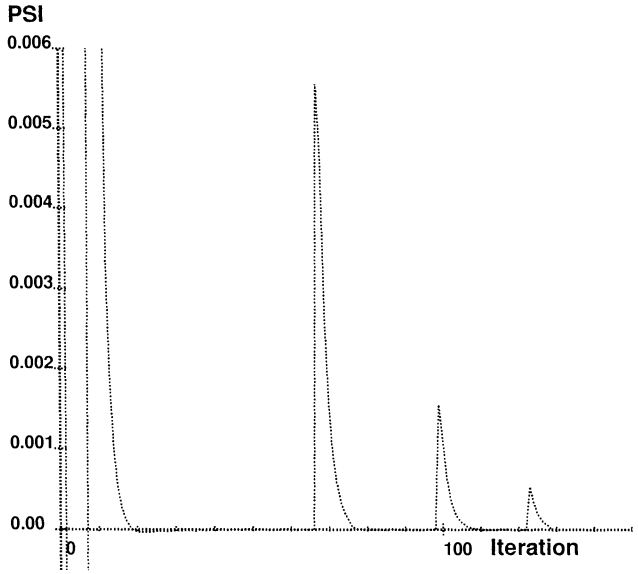


FIG. 3a. *Constraint violation in Problem* 1: 150 *iterations.*

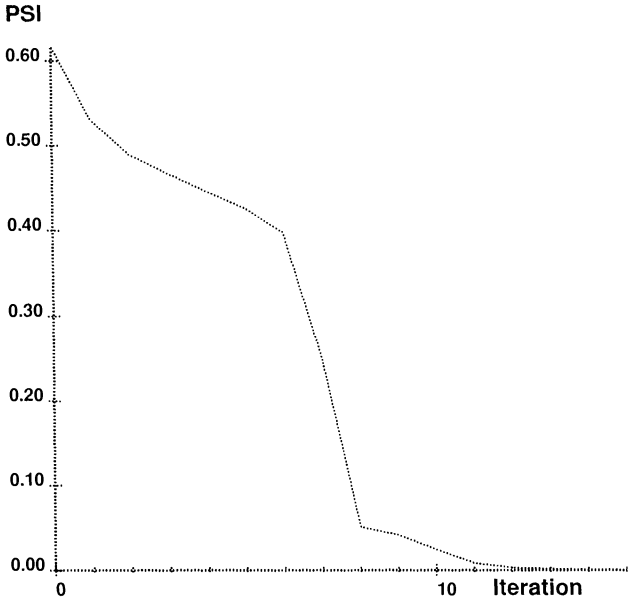


FIG. 3b. *Constraint violation in Problem* 1: *First* 15 *iterations.*

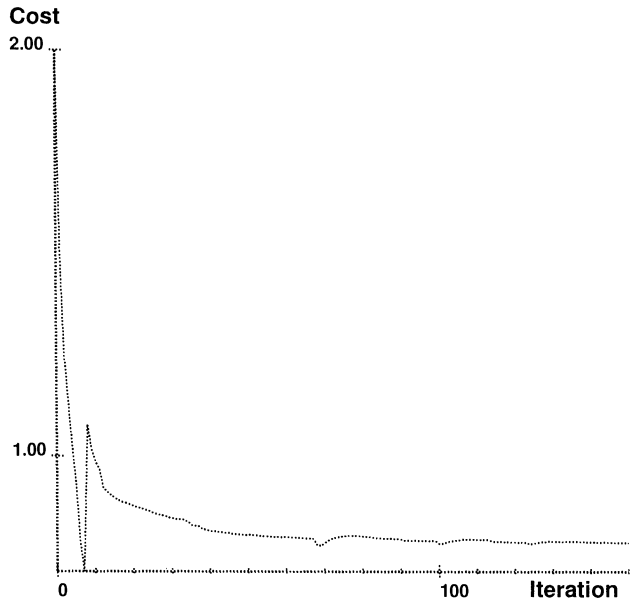


FIG. 4. *Cost v/s iteration number for Problem* 1.

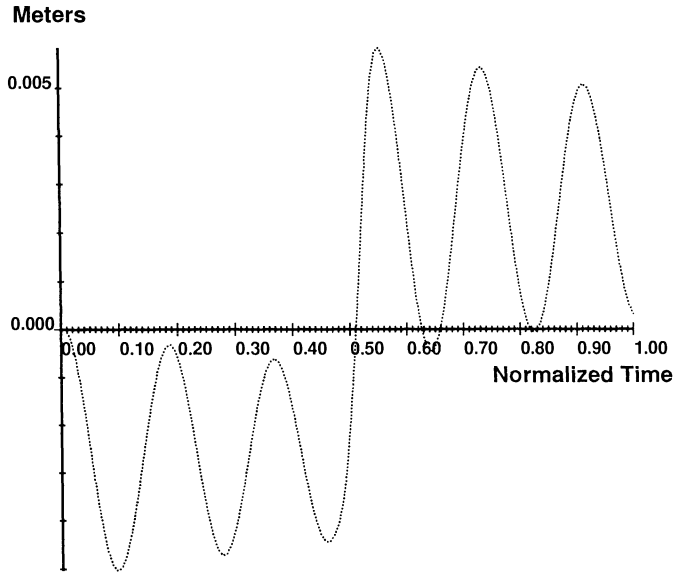


FIG. 5. *Displacement of tip of tube, Problem* 1.

of the tip is about 5 mm and is within the range of validity of the Euler–Bernoulli model. The tip displacement is large between 0.36 seconds and 0.437 seconds. Figure 6 is a profile of the tube deformation, $w(t, x)$ (see Fig. 1), during this interval. The total time for the entire maneuver is 0.7886 seconds.

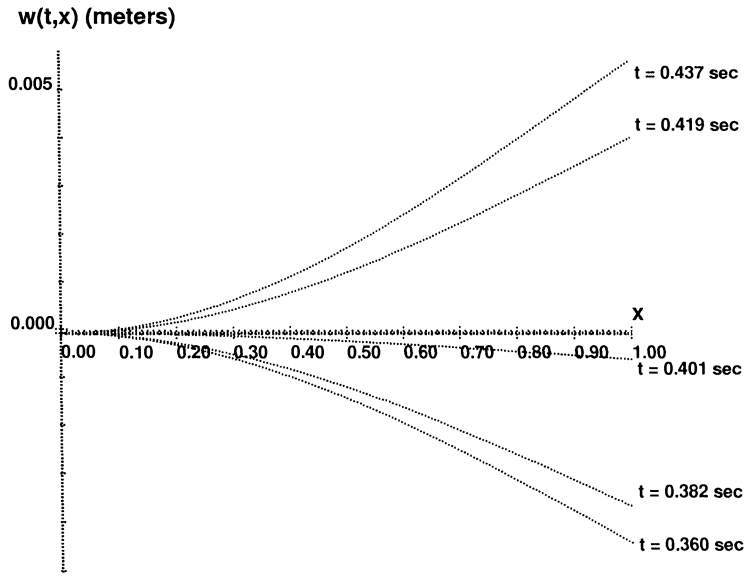


FIG. 6. *Beam profile.*

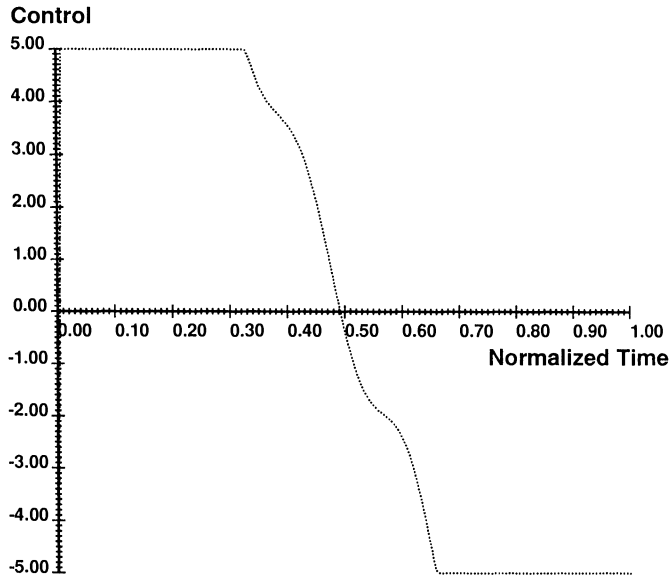


FIG. 7. *Final control for Problem* 2: *Time of maneuver* = 0.800 *seconds.*

PROBLEM $\mathbf{P_2}$. Figure 7 is the graph of the control produced by minimizing the total input energy while constraining the final time to be less than 0.800 seconds, i.e., only 1.4 percent longer than the minimum time computed for $\mathbf{P_1}$. The resulting final time is 0.800. The control is much smoother than the minimum time control, and the total energy consumption is reduced by 18 percent, from 19.15 to 15.72. Figure 8 is the graph of the control when the bound on the final time is extended to 1.00 second, 27 percent over the

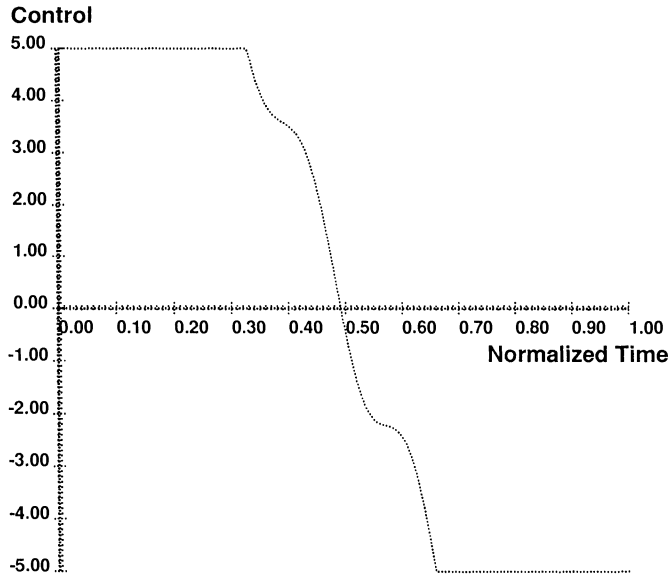


FIG. 8a. *Final control for Problem* 2: *Time of maneuver* = 0.900 *seconds.*

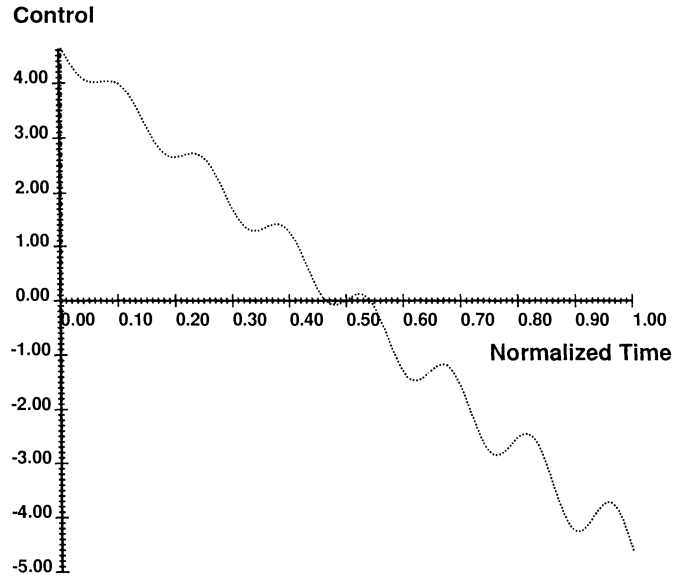


FIG. 8b. *Final control for Problem* 2: *Time of maneuver* = 1.000 *seconds.*

minimum time for the maneuver. The result is a total energy is reduction by 62 percent, to 7.27.

PROBLEM $\mathbf{P_3}$. In problem $\mathbf{P_3}$, we have the additional requirement to keep the potential energy, which is a measure of the total tube deformation, below the parabola (B) for all time. Figure 9 shows the optimal control for problem $\mathbf{P_3}$. The optimal final time for this case is 0.8177 seconds, an increase of 3.7 percent over the solution of problem $\mathbf{P_1}$. Figure 10

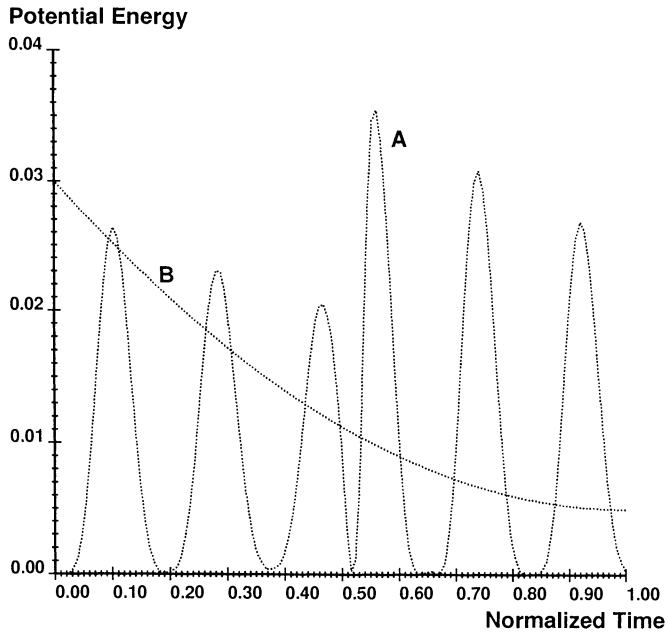


FIG. 9. *Problem* 1: *Potential energy.*

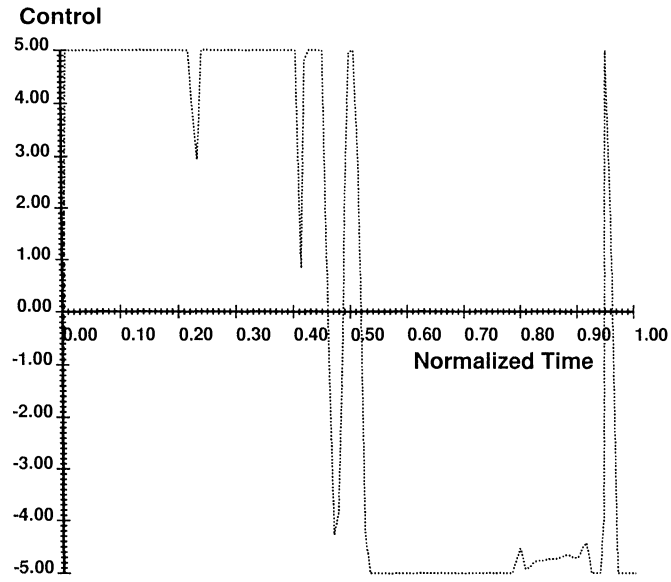


FIG. 10. *Final control for Problem* 3.

shows the potential energy curve for this case, which was constrained to lie below a parabola (B). For comparison in Fig. 11, curve A is the graph of the potential energy of the tube as a function of time for the control generated in solving the minimum time problem $\mathbf{P_1}$.

7. Conclusion. We have presented an approximation theory for the numerical solution of optimal control problems with dynamics in evolution equation form, with control and

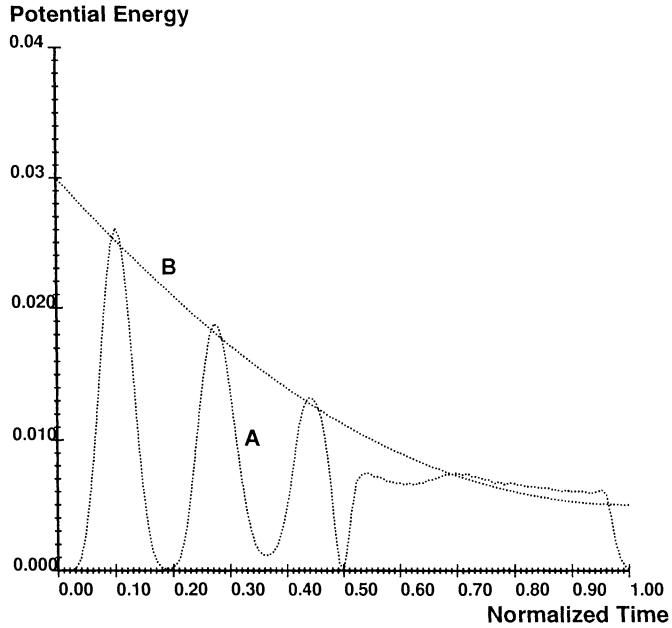


FIG. 11. *Problem* 3: *Potential energy.*

state space constraints. It should be obvious that the theory can be trivially adapted to deal with problems with constraints on the initial state, as well as with unconstrained problems. Although not included in this paper, we have results (reported in [Bak.1], [Bak.2]) which show that our theory can be used in conjunction with finite element techniques to produce reasonably efficient numerical procedures, which have the property that all the accumulation points of the control sequences that they produce satisfy the problem constraints, as well as an optimality condition either for the original or the relaxed problem, depending on whether the accumulation point is in the $L_2^m[0,1]$ topology or in the relaxed controls topology.

Appendix: Differentiability of mild solutions. We will now establish the Frechet differentiability of solutions of (2.2f) with respect to the control $u \in L_2^m[0,1]$ and the scaling parameter τ.

Let $\tilde{M}, \omega \in (0,\infty)$ be such that $\|T(t)\| \leq \tilde{M}e^{\omega t}$ for all $t \in [0,1]$, and let $M \triangleq \tilde{M}e^{\omega \tau_{\max}}$.

LEMMA A.1 (Lipschitz continuity of $z(t,u,\tau)$ in (u,τ).) *There exists* $b_3 \in (0,\infty)$ *such that for all* $u', u \in L_2^m[0,1], t \in [0,1], \tau \in [\tau_{\min}, \tau_{\max}]$,

$$\|z(t,u',\tau') - z(t,u,\tau)\|_X \leq b_3(\|u'-u\|_2^2 + |\tau'-\tau|^2)^{1/2}. \tag{A.1}$$

Proof. For any $u, u' \in L_2^m[0,1]$ and $t \in [0,1]$,

$$\begin{aligned} z(t,u',\tau') - z(t,u,\tau) &= T(\tau' t)z_0 + \int_0^t \tau' T(\tau'(t-s))F(z(s,u',\tau'),u'(s))ds \\ &\quad -T(\tau t)z_0 - \int_0^t \tau T(\tau(t-s))F(z(s,u,\tau),u(s))ds \\ &= [T(\tau' t) - T(\tau t)]z_0 \qquad \text{(A.2a)} \\ &\quad + \int_0^t \{\tau' T(\tau'(t-s))[F(z(s,u',\tau'),u'(s)) \\ &\quad -F(z(s,u,\tau),u(s))] - [\tau T(\tau(t-s)) \\ &\quad -\tau' T(\tau'(t-s))]F(z(s,u,\tau),u(s))\}ds. \end{aligned}$$

Since $\{z(t,u,\tau) \in S \triangleq \{z \in X | \|z\|_X \leq b_1\}$, by Assumption 3.1(ii), we conclude from Assumption 3.1(iii) and Lemma 3.4 that there exists constants $K_S, L \in (0,\infty)$, such that, with $y(t) \triangleq \|z(t,u',\tau) - z(t,u,\tau)\|_X$, for $t \in [0,1]$,

$$y(t) \leq \tau_{\max} M K_S \int_0^t [y(s) + \|u'(s) - u(s)\|_2]\, ds + L|\tau' - \tau|. \tag{A.2b}$$

Applying the Bellman–Gronwall Inequality, and making use of the fact that by the Schwartz Inequality, $\|u\|_1 \leq \|u\|_2$, we obtain that

$$y(t) \leq e^{t_{\max} M K_S}\{\tau_{\max} M K_S \|u' - u\|_1 + L|\tau' - \tau|\} \leq b_3(\|u' - u\|_2^2 + |\tau' - \tau|^2)^{1/2}, \tag{A.2c}$$

where $b_3 \triangleq \sqrt{2}\max\{\tau_{\max} M K_S, L\} e^{\tau_{\max} M K_S}$. □

Next, for $u', u \in L_2^m[0,1]$ and $\tau', \tau \in [\tau_{\min}, \tau_{\max}]$, we define $\delta u = u' - u, \delta\tau = \tau' - \tau$, and $\delta z(\cdot, u, \tau, \delta u, \delta\tau) \in C([0,1], X)$ to be the solution to the equation

$$\begin{aligned}\delta z(t) = \int_0^t \Big\{ & T(\tau(t-s))\tau\left(\frac{\partial F}{\partial z}(z(s,u,\tau),u(s))\delta z(s) + \frac{\partial F}{\partial u}(z(s),u(s))\delta u(s)\right) \\ & + (T(\tau(t-s)) + \tau(t-s)AT(\tau(t-s)))F(z(s,u,\tau),u(s))\delta t\Big\} ds \\ & + tAT(\tau t) z_0 \delta\tau.\end{aligned} \tag{A.3}$$

Note that (A.3) is the first variation with respect to (u,τ) of (2.2f).

THEOREM A.2 (Frechet differentiability of $z(t,u,\tau)$ with respect to (u,τ).) *For all* $u', u \in L_2^m[0,1], \tau', \tau \in [\tau_{\min}, \tau_{\max}]$

$$\|z(t,u',\tau') - z(t,u,\tau) - \delta z(t,u,\tau,u'-u,\tau'-\tau)\|_X \leq o(u'-u,\tau'-\tau), \tag{A.4}$$

where $o(\delta u,\delta\tau)/(\|\delta u\|_2^2 + |\delta\tau|^2)^{1/2} \to 0$ *as* $(\delta u,\delta\tau) \to 0$.

Proof. To simplify notation, we define $\Delta z(t) \triangleq z(t,u',\tau') - z(t,u,\tau)$, $\delta z(t) \triangleq \delta z(t,u,\tau,\delta u,\delta\tau), \delta u \triangleq u' - u, \delta\tau \triangleq \tau' - \tau$, and we remove obvious arguments by setting $F(t) \triangleq F(z(t,u,\tau),u(t)), F'(t) \triangleq F(z(t,u',\tau'),u'(t)), F_z(t) \triangleq \partial F/\partial z(z(t,u,\tau),u(t))$, $F_u(t) \triangleq \partial F/\partial u(z(t,u,\tau),u(t))$.

First, in terms of this simplified notation, we have that

$$\begin{aligned}\delta z(t) = \int_0^t & \{\tau T(\tau(t-s))[F_z(s)\delta z(s) + F_u(s)\delta u(s)] \\ & + [T(\tau(t-s)) + \tau(t-s)AT(\tau(t-s))]F(s)\delta\tau\} ds \\ & + tAT(\tau t) z_0 \delta\tau,\end{aligned} \tag{A.5}$$

$$\begin{aligned}\Delta z(t) &= [T(\tau' t) - T(\tau t)]z_0 + \int_0^t [\tau' T(\tau'(t-s))F'(s) - \tau T(\tau(t-s))F(s)] ds \\ &= [T(\tau' t) - T(\tau t)]z_0 + \tau \int_0^t T(\tau(t-s))[F'(s) - F(s)] ds \\ &\quad + \int_0^t [\tau' T(\tau'(t-s)) - \tau T(\tau(t-s))]F'(s) ds.\end{aligned} \tag{A.6}$$

Hence,

$$
\begin{aligned}
[\Delta z(t) - \delta z(t)] = {} & [T(\tau' t) - T(\tau t) - \delta\tau t A T(\tau t)] z_0 \\
& + \int_0^t t\{[\tau' T(\tau'(t-s)) - \tau T(\tau(t-s))] F'(s) \\
& - \delta\tau[\tau(t-s) A T(\tau(t-s)) + T(\tau(t-s))] F(s) t\} ds \\
& + \tau \int_0^t T(\tau(t-s))\{F_z(s)[\Delta z(s) - \delta z(s)] \\
& + [F'(s) - F(s) - F_z(s)\Delta z(s) - F_u(s)\delta u(s) t\} ds.
\end{aligned} \tag{A.7a}
$$

We will deal with the three groups of terms in the right-hand side of (A.7a) one at a time. We will give full details for the last group only, since the calculations are quite laborious. First, since by Lemma 3.3, $(d/dt)T(t) = AT(t)$,

$$\|T(s+\delta s) - T(s) - AT(t)\delta s\| = o_1(\delta s),$$

where $o_1(\delta s)/\delta s \to 0$ as $\delta s \to 0$. Now let $s = t\tau$ and $s + \delta s = t\tau'$. Hence $\delta s = t(\tau' - \tau) = t\delta\tau$ where $\delta\tau = \tau' - \tau$. Therefore,

$$\|T(t\tau') - T(t\tau) - AT(t)\delta\tau t\| \le o_1(\delta\tau t),$$

and hence

$$\|[T(\tau' t) - T(\tau t) - \delta\tau t A T(\tau t)] z_0\| \le \|z_0\|_X o_1(\delta\tau). \tag{A.7b}$$

Next, making use of Lemmas 3.3 and 3.4, we can show that

$$
\begin{aligned}
& \left\| \int_0^t [\tau' T(\tau'(t-s)) - \tau T(\tau(t-s))] F'(s) - \delta\tau[\tau(t-s) A T(\tau(t-s)) \right. \\
& \left. + T(\tau(t-s))] F(s) ds \right\|_X = o_2(\delta u, \delta\tau),
\end{aligned} \tag{A.7c}
$$

where $o_2((\delta u, \delta\tau))/(\|\delta u\|_2^2 + |\delta\tau|^2)^{1/2} \to 0$ as $(\delta u, \delta\tau) \to 0$.

Finally, making use of Assumption 3.1 and Lemmas 3.4 and A.1, we obtain that

$$
\begin{aligned}
& \tau \left\| \int_0^t (\tau(t-s))\{F_z(s)[\Delta z(s) - \delta z(s)] + F'(s) - F(s) \right. \\
& \left. - F_z(s)\Delta z(s) - F_u(s)\delta u(s)\} ds \right\|_X \\
& \le \tau_{\max} M b_2 \int_0^t \|\Delta z(s) - \delta z(s)\|_X ds + \tau_{\max} M \int_0^t \|\{F'(s) \\
& - F(s) - F_z(s)\Delta z(s) - F_u(s)\delta u(s)\}\|_X ds \\
& \le \tau_{\max} M \int_0^t \left\{ b_2 \|\Delta z(s) - \delta z(s)\|_X + \int_0^1 \left\| \frac{\partial F}{\partial z}(z(s) \right. \right. \\
& \left. + r\Delta z(s), u(s) + r\delta u(s)) - F_z(s) \right\| dr \|\Delta z(s)\|_X \\
& + \left. \int_0^1 \left\| \frac{\partial F}{\partial u}(z(s) + r\Delta z(s), u(s) + r\delta u(s)) - F_u(s) \right\| dr \delta u(s) \right\| \Big\} ds \\
& \le \tau_{\max} M \int_0^t \left\{ b_2 \|\Delta z(s) - \delta z(s)\|_X \right. \\
& + \int_0^1 K_S r(\|\Delta z(s)\|_X + \|\delta u(s)\|) dr \|\Delta z(s)\|_X \\
& \left. + \int_0^1 K_S r(\|\Delta z(s)\|_X + \|\delta u(s)\|) dr \|\delta u(s)\| \right\} ds \\
& \le \tau_{\max} M \int_0^t \{b_2 \|\Delta z(s) - \delta z(s)\|_X + K_S[\|\Delta z(s)\|_X + \|\delta u(s)\|]^2\} ds.
\end{aligned} \tag{A.7d}
$$

Since by Lemma A.1, $\|\Delta z(s)\|_X \leq b_3(\|\delta u\|_2^2 + |\delta\tau|^2)^{1/2}$, we obtain, combining (A.7b)–(A.7d) that

$$\text{(A.7e)} \qquad \|\Delta z(t) - \delta z(t)\|_X \leq \tau_{\max} M \int_0^t \{b_2 \|\Delta z(s) - \delta z(s)\|_X\} ds + \tau_{\max} M K_S [b_3 \|\delta u\|_2 + o_3((\delta u, \delta\tau))],$$

where $o_3((\delta u, \delta\tau))/(\|\delta u\|_2^2 + |\delta\tau|^2)^{1/2} \to 0$ as $(\delta u, \delta\tau) \to 0$. Applying the Bellman–Gronwall Lemma, we obtain that

$$\text{(A.7f)} \qquad \|\Delta z(t) - \delta z(t)\|_X \leq o((\delta u, \delta\tau)),$$

where

$$o((\delta u, \delta\tau))/(\|\delta u\|_2^2 + |\delta\tau|^2)^{1/2} \to 0 \quad \text{as } (\delta u, \delta\tau) \to 0,$$

which completes our proof. □

Proceeding by analogy with the proof of Lemma A.1, it is easy to establish the following result.

LEMMA A.3. *The solution* $\delta z(t, u, \tau, \delta u, \delta\tau)$, *of* (A.3), *is linear in* $(\delta u, \delta\tau)$ *for each* $t \in [0, 1], u \in L_2^m[0, 1]$, *and* $\geq \in [\tau_{\min}, \tau_{\max}]$, *and it is Lipschitz continuous in* $(u, \tau) \in G \times [\tau_{\min}, \tau_{\max}]$, *i.e., there exists* $b_4 < \infty$ *such that for all* $u', u \in L_2^m[0, 1], t \in [0, 1], \tau \in [\tau_{\min}, \tau_{\max}]$,

$$\text{(A.8)} \qquad \|\delta z(t, u', \tau', \delta u, \delta\tau) - \delta z(t, u, \tau, \delta u, \delta\tau)\|_X \leq b_4(\|u' - u\|_2^2 + |\tau' - \tau|^2)^{1/2}.$$

If we denote by $z_{u,\tau}(t, u, \tau)$ the linear map $\delta u \to \delta z(t, u, \tau, \delta u, \delta\tau)$ and make use of Assumption 3.1(v) and Theorem A.4, we obtain the following theorem.

THEOREM A.4. *For all* $u \in L_2^m[0, 1], \tau \in [\tau_{\min}, \tau_{\max}]$, *and* $t \in [0, 1], z(t, u, \tau)$ *admits a Lipschitz continuous Frechet derivative. That is, there exists a Lipschitz continuous linear operator* $Dz(t, u, \tau) = (D_u z(t, u, \tau), D_\tau z(t, u, \tau)) \in \mathbf{B}(L_2^m[0, 1], X)$ *such that for all* $\delta u \in L_2^m[0, 1]$ *and* $\delta\tau \in \mathbb{R}$,

$$\text{(A.9)} \qquad \lim_{\substack{\|\delta u\|_2 \to 0 \\ |\delta\tau| \to 0}} \frac{z(t, u + \delta u, \tau + \delta\tau) - z(t, u, \tau) - D_u z(t, u, \tau)\delta u - D_u z(t, u, \tau)\delta\tau}{(\|\delta u\|_2^2 + |\delta\tau|^2)^{1/2}} = 0.$$

REFERENCES

[Ahm.1] N. H. AHMED, *Properties of relaxed trajectories for a class of nonlinear evolution equations on a Banach space*, SIAM J. Control Optim., 21 (1983), pp 953–967.

[Ale.1] B. M. ALEKSEEV, V M. TIKHOMIROV, AND S. V. FOMIN, *Optimalnoye Upravleniye (Optimal Control)*, Nauka, Moscow, 1979.

[Arm.1] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp 1–3.

[Bak.1] T. E. BAKER, *Algorithms for optimal control of systems described by partial and ordinary differential equations,* Ph.D. dissertation, University of California, Berkeley, June, 1988; Memo No. UCB/ERL M88/45, University of California, Berkeley, Electronics Research Laboratory, June 21, 1988.

[Bak.2] T. E. BAKER AND E. POLAK, *An Algorithm for optimal slewing of flexible structures*, Memo No. UCB/ERL M89/37 University of California, Berkeley, Electronics Research Laboratory, April 11, 1989.

[Ben.1] J. BEN-ASHER, J. A. BURNS, AND E. M. CLIFF, *Time optimal slewing of flexible spacecraft,* in Proc. 26th IEEE Conference on Decision and Control, 1987, pp. 524–528.

[Bur.1] J. A. BURNS, R. E. MILLER, AND E. M. CLIFF, *Control of a viscoelastic shaft with attached tip mass,* in Proc. 26th Conference on Decision and Control, 1987, pp. 997–999.

[Ber.1] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.

[Chu.1] H. M. Chun, *Large-angle slewing maneuvers for flexible spacecraft*, Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[Cul.1] J. Cullum, *Discrete approximations to continuous optimal control problems,* SIAM J. Control, 7 (1969) pp.

[Cul.2] ———, *An explicit procedure for discretizing continuous optimal control problems*, J. Optim. Theory Appl., 8 (1971), pp. 15–34.

[Dun.1] J. C. Dunn, *Diagonally modified conditional gradient methods for input constrained optimal control problems,* SIAM J. Control Optim. 24 (1986), pp. 1177–1191.

[Dun.2] J. C. Dunn and E. Sachs, *The effect of perturbations on the convergence rates of optimization algorithms,* Appl. Math. Optim., 10 (1983), pp. 143–157.

[Flo.1] M. A. Floyd, M. E. Brown, J. D. Turner, and W. E. Vanderwelde, *Implementation of a minimum time and fuel on/off thruster control system for flexible spacecraft,* J. Astronau. Sci., to appear.

[Fuj.1] H. Fujii, *Finite Element schemes: Stability and convergence*, in Advances in Computational Methods in Structural Mechanics and Design, J. T. Oden, R. W. Clough, and Y. Yamamoto, eds., pp. 201–218; Papers at Second U.S. Japan Seminar, University of Huntsville, Alabama Press, 1972.

[Fuj.2] ———, *A note on finite element approximation for evolution equations,* in Kokyuroku, RIMS, No. 202, Kyoto University, 1974, pp 96–117.

[Fuj.3] H. Fujita and T. Suzuki, *Evolution Problems,* in Handbook of Numerical Analysis Vol. II, Finite Element Methods (Part 1), P. G. Ciarlet and J. L. Lions, eds., Elsevier, New York, 1991.

[Gib.1] J. S. Gibson, *An Analysis of optimal modal regulation: Convergence and stability,* SIAM J. Control Optim. 19 (1981), pp.

[Gib.2] J. S. Gibson, D. L. Mingori, A. Adamian, and F. Jabbari, *Approximation of optimal infinite dimensional compensators for flexible structures,* in Proc. Workshop on Identification and Control of Flexible Space Structures, Vol. II, April 1985, pp. 201–218.

[Gib.3] J. S. Gibson, *The Riccati integral equations for optimal control problems on Hilbert spaces,* SIAM J. Control Optim. 17 (1979), pp.

[Jun.1] J. L. Junkins and J. D. Turner, *Optimal Spacecraft Rotational Maneuvers,* Studies in Astronautics 3, Elsevier, Amsterdam, 1986.

[Hua.1] P. Huard, *Programmation Mathematique Convexe*, Rev. Francaise Inf. Rech. Oper., 7 (1968), pp. 43–59.

[Kle.1] R. Klessig and E. Polak, *An adaptive algorithm for unconstrained optimization with applications to optimal control,* SIAM J. Control, 11 (1973), pp. 80–94.

[Lan.1] S. Lang, *Real Analysis*, 2nd ed., Addison Wesley, Reading, MA, 1983.

[May.1] D. Q. Mayne and E. Polak, *First order, strong variations algorithms for optimal control,* J. Optim. Theory Appl., 16 (1975), pp. 277–301.

[May.3] ———, *A feasible directions algorithm for optimal control problems with terminal inequality constraints*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 741–751.

[May.4] ———. *An exact penalty function algorithm for optimal control problems with control and terminal equality constraints, Part* 1, J. Optim. Theory Appl., 32 (1980), pp. 211–246.

[May.5] ———, *An exact penalty function algorithm for optimal control problems with control and terminal equality constraints, Part* 2, J. Optim. Theory Appl., 32 (1980), pp. 345–363.

[Ode.1] J. T. Oden and R. B. Fost, *Convergence, accuracy and stability of finite element approximations of a class of non-linear hyperbolic equations,* Internat. J. Numer. Methods Engrg. 6 (1973), pp. 357–365.

[Pap.1] N. S. Papageorgiou, *Properties of relaxed trajectories of evolution equations and optimal control*, Siam J. Control Optim., 27 (1989), pp. 267–288.

[Paz.1] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Springer-Verlag, Berlin, New York, 1983.

[Pir.1] O. Pironneau and E. Polak, *On the rate of convergence of certain methods of centers*, Math. Programming, 2 (1972), pp. 230–258.

[Pir.2] ———, *A dual method for optimal control problems with initial and final boundary constraints,* SIAM J. Control, 11 (1973), pp. 534–549.

[May.2] E. Polak and D. Q. Mayne, *First order, strong variations algorithms for optimal control problems with terminal inequality constraints,* J. Optim. Theory Appl. 16 (1975), pp. 303–325.

[Pol.1] E. Polak, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., (1987), pp. 21–91.

[Pol.2] E. Polak and L. He, *A unified phase* I *phase* II *method of feasible directions for semi-infinite optimization*, Memo UCB/ERL M89/7, University of California, Berkeley, Electronics Research Laboratory, Feb. 3, 1989; J. Optim. Theory Appl., 69 (1991), pp. 83–107.

[Pol.3] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1972.

[Psh.1] B. N. Pshenichnyi and Yu. M. Danilin, *Numerical Methods in Extremal Problems,* Nauka, Moscow,

1975.

[Sho.1] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations,* Pitman, London, 1977.

[Sin.1] G. Sing, P. T. Kabamba, and N. H. Mclamroch, *Planar, time-optimal rest to rest slewing maneuvers of flexible spacecraft,* J. Guidance Control Dyn., 12 (1989), pp. 71–81.

[Sla.1] M. Slater, *Lagrange multipliers revisited: A contribution to nonlinear programming.* Cowles Commission Discussion Paper, Mathematics 403, November, 1950.

[Teo.1] K. L. Teo and Z. S. Wu, *Computational Methods for Optimizing Distributed Systems,* Academic Press, New York, 1984.

[Teo.2] K. L. Teo, K. H. Wong, and D. J. Clements, *A Feasible directions algorithm for time-lag optimal control problems with control and terminal inequality constraints,* J. Optim. Theory Appl., 46 (1985), pp. 295–318.

[War.1] J. Warga, *Optimal Control of Differential Equations and Functional Equations*, Academic Press, New York, 1972.

[War.2] ———, *Steepest descent with relaxed controls*, SIAM J. Control, 15 (1977), pp. 674–682.

[War.3] ———, *Iterative procedures for constrained and unilateral optimization problems*, SIAM J. Control, 20 (1982), pp. 360–367.

[War.4] ———, *Iterative optimization with equality constraints*, Math. Oper. Res. 9 (1984), pp. 592–605.

[Wil.1] L. J. Williamson and E. Polak, *Relaxed controls and the convergence of optimal control algorithms*, SIAM J. Control, 14 (1976), pp. 737–757.

[Won.1] K. H. Wong and K. L. Teo, *A conditional gradient method for a class of time-lag optimal control problems*, J. Australian Math. Soc., Ser. B, 25 (1984), pp. 518–537.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 261–278, January 1994

015

REDHEFFER'S LEMMA AND H_∞-CONTROL FOR INFINITE-DIMENSIONAL SYSTEMS*

BERT VAN KEULEN†

Abstract. An infinite-dimensional version of a lemma that has been crucial in the theory of H_∞-control with measurement-feedback for finite-dimensional systems is proved. This extension is used to parametrize all controllers that solve the suboptimal H_∞-control problem for a large class of infinite-dimensional systems.

Key words. Redheffer's lemma, H_∞-control, infinite-dimensional state-space systems

AMS subject classifications. 93B36, 93C25

1. Introduction.

In this paper we prove an infinite-dimensional version of a lemma that has been crucial in the theory of H_∞-control with measurement-feedback for finite-dimensional systems (see [2, Lemma 15]). This lemma is sometimes referred to as Redheffer's lemma since it resembles some results published by Redheffer in [9]. Redheffer considered a very general Hilbert space setting, but the lemma that we present here is far from immediate from the original results in [9].

In [2] the lemma is stated in the frequency domain and proved using finite-dimensional frequency domain techniques, including a Nyquist contour argument. We note that these techniques cannot be applied to the infinite-dimensional time-domain version that we consider here. The proof in this paper depends solely on time-domain techniques and is therefore applicable to a very large class of systems. We believe that the same kind of reasoning can be used to solve the measurement-feedback H_∞-control problem for nonlinear or time-varying systems.

Using the infinite-dimensional extension, we give a parametrization of all suboptimal controllers that solve the H_∞-control problem with measurement-feedback for a large class of infinite-dimensional systems (this type of result was derived in [2] and [12] for the finite-dimensional case). This completes the results that were given in [5]. In doing so, we use procedures and formulations that have been published in [2], [12], and [11] for the finite-dimensional case.

2. Preliminary results.

In this section we introduce our class of infinite-dimensional systems, quote some known results for this class, and derive some new results that are interesting in their own right. We consider infinite-dimensional linear systems of the following form (see also [6], [5]). Suppose that A is the infinitesimal generator of a C_0-semigroup $T(\cdot)$ on the real separable Hilbert space X, $B \in \mathcal{L}(U,X), C \in \mathcal{L}(X,Y)$, and $D \in \mathcal{L}(U,Y)$, where U and Y are also real separable Hilbert spaces. For $u(\cdot) \in L_2^{\text{loc}}(0,\infty;U)$, $x_0 \in X$, and $t \geq 0$

$$\begin{aligned} x(t) &= T(t)x_0 + \textstyle\int_0^t T(t-s)Bu(s)ds \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{2.1}$$

* Received by the editors August 21, 1991; accepted for publication (in revised form) August 25, 1992.

† Mathematics Institute, P. O. Box 800, 9700 AV Groningen, the Netherlands (bertvk@math.rug.nl).

is a well-defined system with state $x(t)$, input $u(t)$, and output $y(t)$. To simplify notation we denote such a system by

$$\begin{aligned} \dot{x} &= Ax + Bu, \qquad x(0) = x_0 \\ y &= Cx + Du, \qquad t \geq 0. \end{aligned} \tag{2.2}$$

We recall some basic facts about this class of systems (see, e.g, [1]): If $x_0 = 0$, (2.2) defines a linear map G from $L_2^{\mathrm{loc}}(0,\infty;U)$ to $L_2^{\mathrm{loc}}(0,\infty;Y)$ and in this case we call (2.2) a *realization* of G. G is *causal*, i.e., for all $T > 0$ we have

$$u_1(t) = u_2(t) \quad \text{a.e. } t \in [0,T] \Rightarrow (Gu_1)(t) = (Gu_2)(t) \quad \text{a.e. } t \in [0,T].$$

If $U = Y$ and $D \in \mathcal{L}(U)$ we can define the inverse system G^{-1} as the system given by

$$\begin{aligned} \dot{x} &= (A - BD^{-1}C)x + BD^{-1}u, \qquad x(0) = 0 \\ y &= -D^{-1}Cx + D^{-1}u, \qquad t \geq 0, \end{aligned} \tag{2.3}$$

and it is easy to see that for all $u \in L_2^{\mathrm{loc}}(0,\infty;U)$ we have $G^{-1}(Gu) = G(G^{-1}u) = u$ (use some well-known perturbation results, see, e.g., [7, §3.1]).

If A is the infinitesimal generator of an exponentially stable C_0-semigroup, it is well known that $G \in \mathcal{L}(L_2(0,\infty;U), L_2(0,\infty;Y))$. In general, G is unbounded, so we consider G as a map from $D(G) \subseteq L_2(0,\infty;U)$ to $L_2(0,\infty;Y)$, where $D(G)$ is given by

$$D(G) := \{u \in L_2(0,\infty;U) |\ \ (Gu)(\cdot) \in L_2(0,\infty;Y)\}.$$

Since for every $T > 0$, G is a bounded linear map from $L_2(0,T;U)$ to $L_2(0,T;Y)$, it is not difficult to see that $(G, D(G))$ defines a closed linear map. Hence, $G \in \mathcal{L}(L_2(0,\infty;U), L_2(0,\infty;Y))$ if and only if $D(G) = L_2(0,\infty;U)$ (apply the closed-graph theorem). If $D(G) = L_2(0,\infty;U)$, we call G *i/o-stable* (input/output stable) and we denote its operator norm by $\|G\|$. Finally, we note that if G is i/o-stable, the transfer function $G(\cdot)$ of system (2.2) satisfies $G(\cdot) \in H_\infty(\mathrm{C}^+, \mathcal{L}(U,Y))$ and $\|G\| = \|G(\cdot)\|_\infty$.

In this paper, we adopt the usual definitions of stabilizability and detectability: the pair (A,B) is called *exponentially stabilizable* if there exists an $F \in \mathcal{L}(X,U)$ such that the C_0-semigroup generated by $A + BF$ is exponentially stable, and the pair (C,A) is called *exponentially detectable* if there exists a $K \in \mathcal{L}(Y,X)$ such that the C_0-semigroup generated by $A + KC$ is exponentially stable.

In the following two lemmas, we relate i/o-stability with exponential stability and internal stability. These lemmas extend some results in [4], where finite-dimensionality of U and Y is essential. The first lemma is a result from [5, Lemma 3.2].

LEMMA 2.1. *Suppose we have a system G given by* (2.2) *with $x_0 = 0$ and suppose that (A,B) is exponentially stabilizable and (C,A) is exponentially detectable. Then the C_0-semigroup $T(\cdot)$ generated by A is exponentially stable if and only if G is i/o-stable.*

Now suppose that we have two systems of the form (2.2) given by

$$G: \begin{cases} \dot{x}_1 = A_1x_1 + B_1u_1, \qquad x_1(0) = 0 \\ \\ y_1 = C_1x_1 + D_1u_1, \end{cases} \tag{2.4}$$

$$K: \begin{cases} \dot{x}_2 = A_2 x_2 + B_2 u_2, & x_2(0) = 0 \\ y_2 = C_2 x_2 + D_2 u_2, \end{cases} \tag{2.5}$$

where A_1 and A_2 are infinitesimal generators of the C_0-semigroups $T_1(\cdot)$ and $T_2(\cdot)$ on the real separable Hilbert spaces X_1 and X_2, $u_1(t), y_2(t) \in U$, $u_2(t)$, $y_1(t) \in Y$, $B_1 \in \mathcal{L}(U, X_1)$, $C_1 \in \mathcal{L}(X_1, Y)$, $D_1 \in \mathcal{L}(U, Y)$, $B_2 \in \mathcal{L}(Y, X_2)$, $C_2 \in \mathcal{L}(X_2, U)$, and $D_2 \in \mathcal{L}(Y, U)$ with U and Y also real separable Hilbert spaces. In [5, Lemma 3.3] it is shown that if $(I - D_1 D_2)^{-1} \in \mathcal{L}(Y)$ and $(I - D_2 D_1)^{-1} \in \mathcal{L}(U)$, the closed-loop system on $X_1 \times X_2$ determined by $u_1 = y_2 + v_1$, $u_2 = y_1 + v_2$, from (v_1, v_2) to (u_1, u_2) is given by

$$\begin{pmatrix} \dot{x}_1 \\ x_2 \end{pmatrix} = \mathcal{A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \mathcal{B} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}; \qquad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \mathcal{C} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \mathcal{D} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix},$$

where

$$\mathcal{A} = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} + \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \begin{pmatrix} I & -D_2 \\ -D_1 & I \end{pmatrix}^{-1} \begin{pmatrix} 0 & C_2 \\ C_1 & 0 \end{pmatrix},$$

$$\mathcal{B} = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \begin{pmatrix} I & -D_2 \\ -D_1 & I \end{pmatrix}^{-1}, \tag{2.6}$$

$$\mathcal{C} = \begin{pmatrix} I & -D_2 \\ -D_1 & I \end{pmatrix}^{-1} \begin{pmatrix} 0 & C_2 \\ C_1 & 0 \end{pmatrix}, \quad \mathcal{D} = \begin{pmatrix} I & -D_2 \\ -D_1 & I \end{pmatrix}^{-1},$$

and $\mathcal{A}$ is the infinitesimal generator of a C_0-semigroup $\mathcal{T}(\cdot)$ on the Hilbert space $X_1 \times X_2$ (see also Fig. 2.1).

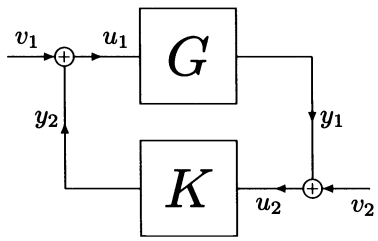


FIG. 2.1. *Internal stability.*

Using the linear maps defined by G and K, we can also formulate this as

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = G_{cl} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} (I - KG)^{-1} & (I - KG)^{-1} K \\ G(I - KG)^{-1} & (I - GK)^{-1} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \tag{2.7}$$

(Note that since $(I - D_1 D_2)^{-1} \in \mathcal{L}(Y)$ and $(I - D_2 D_1)^{-1} \in \mathcal{L}(U)$, $(I - GK)^{-1}$ and $(I - KG)^{-1}$ are both well defined.) If G_{cl} is i/o-stable K is usually called an *internally stabilizing* controller for G.

The following result holds.

LEMMA 2.2. *Consider the systems G and K given by* (2.4) *and* (2.5). *We have the following equivalence: $\mathcal{A}$ given by* (2.6) *is the generator of an exponentially stable C_0-semigroup if and only if (A_1, B_1) and (A_2, B_2) are exponentially stabilizable, (C_1, A_1) and (C_2, A_2) are exponentially detectable and G_{cl} defined by* (2.7) *is i/o-stable.*

Proof. Necessity. Suppose that $\mathcal{A}$ generates an exponentially stable C_0-semigroup.

The fact that G_{cl} is i/o-stable follows from well-known results (see, e.g., [6, App. A.1]). The fact that (A_1, B_1) and (A_2, B_2) are exponentially stabilizable and (C_1, A_1) and (C_2, A_2) are exponentially detectable follows from [6, Rem. 5.2].

Sufficiency. Suppose that G_{cl} is i/o-stable and that (A_1, B_1) and (A_2, B_2) are exponentially stabilizable and (C_1, A_1) and (C_2, A_2) are exponentially detectable.

Since (A_1, B_1) and (A_2, B_2) are exponentially stabilizable it is easy to see from (2.6) that the pair $(\mathcal{A}, \mathcal{B})$ is exponentially stabilizable. Similarly, since (C_1, A_1) and (C_2, A_2) are exponentially detectable, we see that the pair $(\mathcal{C}, \mathcal{A})$ is exponentially detectable. The result now follows from Lemma 2.1. □

Next we quote another result from [5], which is a kind of small gain theorem with exponential stability (see again [5, Lemma 3.3]).

LEMMA 2.3. *Consider again G and K given by* (2.4) *and* (2.5). *Suppose that $T_1(\cdot)$ and $T_2(\cdot)$ are both exponentially stable, that $\|G\| \leq 1$, and that $\|K\| < 1$. Then $\mathcal{A}$ given by* (2.6) *is the generator of an exponentially stable C_0-semigroup.*

We conclude this section with some useful definitions and diagrams regarding feedback interconnections and linear fractional transformations.

Suppose that we have a system G of the form

$$G: \begin{cases} \dot{x} = Ax + B_1 w + B_2 u \\ z = C_1 x + D_{12} u, \qquad x(0) = 0 \\ y = C_2 x + D_{21} w \end{cases} \tag{2.8}$$

(interpreting (2.8) as (2.2); now $x(t) \in X$, $u(t) \in U$, $w(t) \in W$, $z(t) \in Z$, and $y(t) \in Y$, where X, U, W, Z, and Y are all real separable Hilbert spaces, A is the infinitesimal generator of a C_0-semigroup $T(\cdot)$, and $B_1, B_2, C_1, D_{12}, C_2$, and D_{21} are linear and bounded maps with the appropriate spaces).

Let us express (2.8) in the following way:

$$\begin{pmatrix} z \\ y \end{pmatrix} = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} \begin{pmatrix} w \\ u \end{pmatrix}, \tag{2.9}$$

where G_{ij} represent the corresponding linear maps denoted below:

$$\begin{aligned} G_{11} w &= \textstyle\int_0^t C_1 T(t-s) B_1 w(s) ds, \\ G_{12} u &= \textstyle\int_0^t C_1 T(t-s) B_2 u(s) ds + D_{12} u, \\ G_{21} w &= \textstyle\int_0^t C_2 T(t-s) B_1 w(s) ds + D_{21} w, \\ G_{22} u &= \textstyle\int_0^t C_2 T(t-s) B_2 u(s) ds. \end{aligned} \tag{2.10}$$

Furthermore, let G_2 be a system of the form

$$G_2: \begin{cases} \dot{p} = Mp + N y_1, \qquad p(0) = 0 \\ \\ u_1 = Lp + R y_1, \end{cases} \tag{2.11}$$

where M is the infinitesimal generator of the C_0-semigroup $V(\cdot)$ on the real separable Hilbert space P, $y_1(t) \in Y$, $u_1(t) \in U$, $N \in \mathcal{L}(U, P)$, $L \in \mathcal{L}(P, U)$, and $R \in \mathcal{L}(Y, U)$.

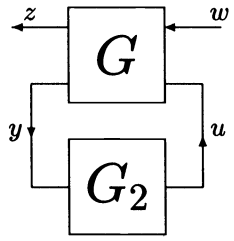


FIG. 2.2. $G_{zw} = \mathcal{F}(G, G_2)$.

We define the closed-loop system G_{zw} as the interconnection of (2.8) and (2.11) with $y_1 = y$ and $u_1 = u$ as in Fig. 2.2. Hence G_{zw} is the map from $L_2^{\mathrm{loc}}(0,\infty;W)$ to $L_2^{\mathrm{loc}}(0,\infty;Z)$ given by

$$G_{zw}: \begin{cases} \begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \mathcal{A}\begin{pmatrix} x \\ p \end{pmatrix} + \mathcal{B}w, \qquad \begin{pmatrix} x \\ p \end{pmatrix}(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \\ z = \mathcal{D}_1\begin{pmatrix} x \\ p \end{pmatrix} + \mathcal{D}_2 w \end{cases} \tag{2.12}$$

with

$$\mathcal{A} = \begin{pmatrix} A + B_2RC_2 & B_2L \\ NC_2 & M \end{pmatrix}, \qquad \mathcal{B} = \begin{pmatrix} B_2RD_{21} + B_1 \\ ND_{21} \end{pmatrix},$$

$$\mathcal{D}_1 = (C_1 + D_{12}RC_2 \quad D_{12}L), \qquad \mathcal{D}_2 = D_{12}RD_{21}, \tag{2.13}$$

where $\mathcal{A}$ is the infinitesimal generator of the C_0-semigroup $\mathcal{T}(\cdot)$ on the Hilbert space $X \times P$. As far as $\mathcal{A}$ is concerned, the relation with (2.4)–(2.7) is clear: G_{22} plays the role of G in (2.4) and G_2 is as in (2.5).

We can formulate (2.12) differently, using (2.9) with $u = G_2y$: From (2.10) we see that the feedthrough operator of G_{22} is zero. Therefore, $I - G_{22}G_2$ is invertible as a map from $L_2^{\mathrm{loc}}(0,\infty;Y)$ to $L_2^{\mathrm{loc}}(0,\infty;Y)$ and so

$$z = G_{zw}w = (G_{11} + G_{12}G_2(I - G_{22}G_2)^{-1}G_{21})w.$$

For any G of the form (2.8)–(2.9) and G_2 of the form (2.11) we define $\mathcal{F}$ (a linear fractional transformation) as

$$\mathcal{F}(G, G_2) := G_{11} + G_{12}G_2(I - G_{22}G_2)^{-1}G_{21}. \tag{2.14}$$

We have seen above that $\mathcal{F}$ is well defined and that $\mathcal{F}(G, G_2)$ has a realization of the form (2.12), (2.13). Unless stated otherwise, we use this realization of $\mathcal{F}(G, G_2)$ on $X \times P$ throughout.

3. Redheffer's lemma. In this section we present our infinite-dimensional time-domain version of Redheffer's lemma, which is closely related to the finite-dimensional result in [2, Lemma 15]. Our version is stated in Theorem 3.2.

First we prove a crucial lemma that is related to the finite-dimensional result that if $G(\cdot) \in \mathrm{R}L_\infty$, $\|G(\cdot)\|_\infty < 1$ and $(I - G(\cdot))^{-1} \in \mathrm{R}H_\infty$ it follows that $G(\cdot) \in \mathrm{R}H_\infty$ (this finite-dimensional result is usually proved using a Nyquist contour argument; see [11, Lemma 2.11]).

LEMMA 3.1. *Suppose that G is a system of the form (2.2) with $x_0 = 0$. If $U = Y$ and D is such that $(I - D)^{-1} \in \mathcal{L}(U)$, then the inverse system $(I - G)^{-1}$ exists. If in addition $(I - G)^{-1}$ is i/o-stable and $\|Gu\|_2 \leq \|u\|_2$ for all $u \in D(G)$, then it follows that G is i/o-stable.*

Proof. Since $(I - D)^{-1} \in \mathcal{L}(U)$, $(I - G)^{-1}$ exists (see §2) and we know that $(I - G)^{-1}$ is i/o-stable and $\|Gu\|_2 \leq \|u\|_2$ for all $u \in D(G)$.

Note first that for arbitrary $u_1 \in L_2(0, \infty; U)$, we have $Gu_1 \in L_2^{\text{loc}}(0, \infty; U)$. We must prove that $D(G) = L_2(0, \infty; U)$, so we show that Gu_1 is an element of $L_2(0, \infty; U)$.

Let $T > 0$ be arbitrary and define $y_{2T} \in L_2(0, \infty; U)$ as follows:

$$y_{2T}(t) := \begin{cases} ((I - G)u_1)(t) & \text{for a.e. } t \in [0, T] \\ 0 & \text{for } t > T. \end{cases} \tag{3.1}$$

Since $(I - G)^{-1}$ is i/o stable we can define $u_{2T} \in L_2(0, \infty; U)$ by

$$u_{2T} := (I - G)^{-1} y_{2T}. \tag{3.2}$$

It follows that

$$((I - G)u_{2T})(t) = ((I - G)u_1)(t) \quad \text{for a.e. } t \in [0, T]. \tag{3.3}$$

Since $(I - G)^{-1}$ is a causal system this implies that

$$u_{2T}(t) = u_1(t) \quad \text{for a.e. } t \in [0, T]. \tag{3.4}$$

Furthermore, since $u_{2T}, y_{2T} \in L_2(0, \infty; U)$, (3.2) implies that

$$Gu_{2T} = u_{2T} - y_{2T} \in L_2(0, \infty; U) \tag{3.5}$$

and so $u_{2T} \in D(G)$.

Hence, using the fact that $\|Gu\|_2 \leq \|u\|_2$ for all $u \in D(G)$ we have

$$\|Gu_{2T}\|_2^2 = \|u_{2T} - y_{2T}\|_2^2 \leq \|u_{2T}\|_2^2. \tag{3.6}$$

Since $y_{2T}(t) = 0$ for $t > T$ (see (3.1)), we can express (3.6) as

$$\int_0^T \|u_{2T}(t) - y_{2T}(t)\|_U^2 \, dt + \int_T^\infty \|u_{2T}(t)\|_U^2 \, dt \leq \int_0^\infty \|u_{2T}(t)\|_U^2 \, dt. \tag{3.7}$$

It follows from (3.3) and (3.4) that $(Gu_{2T})(t) = (Gu_1)(t)$ for a.e. $t \in [0, T]$ and using this with (3.5) in (3.7) gives

$$\begin{aligned} \int_0^T \|Gu_1(t)\|_U^2 \, dt &\leq \int_0^\infty \|u_{2T}(t)\|_U^2 \, dt - \int_T^\infty \|u_{2T}(t)\|_U^2 \, dt \\ &= \int_0^T \|u_{2T}(t)\|_U^2 \, dt = \int_0^T \|u_1(t)\|_U^2 \, dt, \end{aligned} \tag{3.8}$$

where the last equality follows from (3.4).

T was arbitrary so (3.8) implies that

$$\int_0^T \|Gu_1(t)\|_U^2 \, dt \leq \int_0^T \|u_1(t)\|_U^2 \, dt \leq \|u_1\|_2^2 \text{ for all } T > 0$$

and this completes the proof. □

Now suppose that we have a system G of the form (2.8), (2.9) and a system G_2 of the form (2.11). Note that there is no feedthrough from w to z nor from u to y; this is done only for simplicity of presentation. The result that we give in Theorem 3.2 is also valid if these terms are included, under some mild wellposedness conditions.

We shall make the following assumptions:

(3.9) The C_0-semigroup $T(\cdot)$ generated by A is exponentially stable.

(3.10) G is inner, i.e. for all $w \in L_2(0,\infty;W)$ and $u \in L_2(0,\infty;U)$ we have $\|z\|_2^2 + \|y\|_2^2 = \|w\|_2^2 + \|u\|_2^2$.

(3.11) G_{21}^{-1} exists and is i/o-stable.

Note that assumption (3.9) implies that G is i/o-stable and that assumption (3.11) implies that $Y = W$ and $D_{21}^{-1} \in \mathcal{L}(Y)$.

The following result holds.

THEOREM 3.2. *Suppose that we have a system G of the form* (2.8), (2.9) *and a system G_2 of the form* (2.11). *Furthermore, suppose that the assumptions* (3.9)–(3.11) *are satisfied. Then the C_0-semigroup $\mathcal{T}(\cdot)$ that corresponds to the closed-loop system* (2.12) *is exponentially stable and $\|G_{zw}\| = \|\mathcal{F}(G,G_2)\| < 1$ if and only if the C_0-semigroup $V(\cdot)$ generated by M in* (2.11) *is exponentially stable and $\|G_2\| < 1$.*

Proof. First, we recall from §3 that $(I - G_{22}G_2)^{-1}$ exists so that $G_{zw} = \mathcal{F}(G,G_2)$ is well defined.

Sufficiency. Suppose that $V(\cdot)$ is exponentially stable and $\|G_2\| < 1$. Since G is i/o-stable and inner we have $\|G_{22}\| \le 1$. It follows from Lemma 2.3 that $\mathcal{T}(\cdot)$ is exponentially stable.

Now let $w \in L_2(0,\infty;W)$ be an input for the closed-loop system (2.12) and let $u \in L_2(0,\infty;U)$, $y \in L_2(0,\infty;Y)$ and $z \in L_2(0,\infty;Z)$ have the corresponding values so that (2.9) holds and $u = G_2 y$. Now since $\|G_2\| < 1$, there exists some $\epsilon > 0$ such that $\|u\|_2^2 \le (1-\epsilon)\|y\|_2^2$. We know that G is inner so

$$\|z\|_2^2 = \|w\|_2^2 + \|u\|_2^2 - \|y\|_2^2 \le \|w\|_2^2 - \epsilon\|y\|_2^2. \tag{3.12}$$

Since (2.9) holds with $u = G_2 y$, we have

$$y = (I - G_{22}G_2)^{-1}G_{21}w.$$

Since G_{21}^{-1} exists and is i/o-stable and G_2 and G_{22} are also i/o-stable, it follows that

$$\|w\|_2^2 = \left\|G_{21}^{-1}(I - G_{22}G_2)y\right\|_2^2 \le const\,\|y\|_2^2. \tag{3.13}$$

Combining (3.12) and (3.13) shows that there exists a $\delta > 0$ such that $\|z\|_2^2 = \|G_{zw}w\|_2^2 \le (1-\delta)\|w\|_2^2$ for all $w \in L_2(0,\infty;W)$, so $\|\mathcal{F}(G,G_2)\| = \|G_{zw}\| < 1$.

Necessity. Suppose that $\mathcal{T}(\cdot)$ is exponentially stable and $\|\mathcal{F}(G,G_2)\| = \|G_{zw}\| < 1$. We proceed in seven steps.

1. Note that $G_2(I - G_{22}G_2)^{-1}$ and $(I - G_{22}G_2)^{-1}$ are i/o-stable. This follows from the fact that $\mathcal{T}(\cdot)$ is exponentially stable, formula (2.7) and Lemma 2.2.

2. Prove that $\|G_2 y\|_2 \le \|y\|_2$ for all $y \in D(G_2)$. Let $y \in D(G_2)$ and define $u := G_2 y \in L_2(0,\infty;U)$. Define $w := G_{21}^{-1}(I - G_{22}G_2)y = G_{21}^{-1}y - G_{21}^{-1}G_{22}u$, so that

$w \in L_2(0,\infty;W)$. Define $z := G_{zw}w \in L_2(0,\infty;Z)$. Now it is easy to see that (2.8), (2.9) is satisfied so $\|z\|_2^2 + \|y\|_2^2 = \|w\|_2^2 + \|u\|_2^2$ (G is inner). $\|G_{zw}\| < 1$ implies that $\|z\|_2^2 \leq (1-\epsilon)\|w\|_2^2$ for some $\epsilon > 0$ and so $\|u\|_2^2 = \|G_2 y\|_2^2 = \|z\|_2^2 + \|y\|_2^2 - \|w\|_2^2 \leq \|y\|_2^2 - \epsilon\|w\|_2^2 \leq \|y\|_2^2$.

3. Note that $D(G_{22}G_2) = D(G_2)$. Since G_{22} is i/o-stable, we have $D(G_{22}G_2) \supseteq D(G_2)$. Now suppose that $y \in D(G_{22}G_2)$. It follows that $(I - G_{22}G_2)y \in L_2(0,\infty;Y)$. Since $G_2(I - G_{22}G_2)^{-1}$ is i/o-stable it follows that $G_2 y = G_2(I - G_{22}G_2)^{-1}(I - G_{22}G_2)y \in L_2(0,\infty;U)$, and so $y \in D(G_2)$.

4. Use Lemma 3.1 to prove that $G_{22}G_2$ is i/o-stable. Using steps 2 and 3 and the fact that $\|G_{22}\| \leq 1$ we have $\|G_{22}G_2 y\|_2 \leq \|G_2 y\|_2 \leq \|y\|_2$ for all $y \in D(G_{22}G_2)$. We know that $(I - G_{22}G_2)^{-1}$ is i/o-stable. Lemma 3.1 implies that $G_{22}G_2$ is i/o-stable.

5. Prove that G_2 is i/o-stable. We know that $G_2(I - G_{22}G_2)^{-1}$ and $I - G_{22}G_2$ are both i/o-stable. Now we use $G_2 = G_2(I - G_{22}G_2)^{-1}(I - G_{22}G_2)$.

6. Prove that $V(\cdot)$ is exponentially stable. Since $\mathcal{T}(\cdot)$ is exponentially stable, it follows from Lemma 2.2 that (M,N) is exponentially stabilizable and (L,M) is exponentially detectable. Since G_2 is i/o-stable, it follows from Lemma 2.1 that $V(\cdot)$ is exponentially stable.

7. Prove that $\|G_2\| < 1$. We conclude from steps 2 and 5 that $\|G_2\| \leq 1$. Now let $y \in L_2(0,\infty;Y)$, define $u := G_2 y$ and $w := G_{21}^{-1}(I - G_{22}G_2)y$. As in step 2 we have $\|G_2 y\|_2^2 \leq \|y\|_2^2 - \epsilon\|w\|_2^2$. Since $y = (I - G_{22}G_2)^{-1}G_{21}w$, we see that $\|y\|_2^2 \leq \text{const}\,\|w\|_2^2$, so there exists some $\delta > 0$ such that $\|G_2 y\|_2^2 \leq (1-\delta)\|y\|_2^2$, and so $\|G_2\| < 1$. □

Remark 3.3. Apart from the exponential stability, the sufficiency part of Theorem 3.2 follows from Redheffer's results in [9]. The necessity part might be derived by applying some results in [9], but this would give a proof that is much longer than the one we have given here. The clue would be [9, (17)] and the relation between matrix and $*$-product inverses for isometric operators. Since Redheffer considers only bounded operators and G_2 is not a priori bounded, the result should first be obtained for $L_2(0,T)$ and then somehow extended to $L_2(0,\infty)$.

In the next section we use Theorem 3.2 to parametrize all controllers that solve the regular suboptimal H_∞-control problem for a class of systems of the form (2.8).

4. Controller parametrization.

Consider again the systems given by (2.8) and (2.11). We are looking for controllers K of the form (2.11) that make the closed-loop system given by (2.12) exponentially stable, i.e., $\mathcal{T}(\cdot)$ is exponentially stable, and satisfy $\|\mathcal{F}(G,K)\| = \|G_{zw}\| < \gamma$. A controller with these properties will be called *admissible.*

Under some regularity conditions, in [5] necessary and sufficient conditions are derived for the existence of an admissible controller. As in the finite-dimensional case, these necessary and sufficient conditions are expressed by the solvability of two coupled Riccati equations. Using Theorem 3.2, in this section we simplify some of the proofs in [5] and give a parametrization of all admissible controllers. To do this, we need some a priori assumptions (see also [5]):

(4.1) there exists an $\epsilon > 0$ such that for all $(\omega,x,u) \in \mathrm{R} \times D(A) \times U$ with $i\omega x = Ax + B_2 u$, there holds $\|C_1 x + D_{12}u\|_Z^2 \geq \epsilon\|x\|_X^2$,

(4.2) $$D_{12}^*[C_1 \;\; D_{12}] = [0 \;\; I],$$

(4.3) there exists an $\epsilon > 0$ such that for all $(\omega, x, y) \in \mathrm{R} \times D(A^*) \times Y$ with $i\omega x = A^*x + C_2^*y$, there holds $\|B_1^*x + D_{21}^*y\|_W^2 \geq \epsilon \|x\|_X^2$,

$$D_{21}[B_1^* \quad D_{21}^*] = [0 \quad I]. \tag{4.4}$$

Assumptions (4.1) and (4.3) are the infinite-dimensional analogues of the weakest assumptions under which the regular version of the finite-dimensional H_∞-problem has been solved (see, e.g., [3]). Just as in the finite-dimensional case [3], (4.2) and (4.4) can be replaced by the assumption that $D_{12}^*D_{12}$ and $D_{21}D_{21}^*$ are coercive. Also, feedthrough terms from disturbance w to the to-be-controlled output z and from the control u to the measured output y can be included, but all this leads only to more complicated formulas. Furthermore, without loss of generality, we restrict ourselves to the case $\gamma = 1$ (as usual the general case can be obtained by scaling).

Before we derive the parametrization of all admissible controllers we must present two preliminary results, which follow from [5] and [6]. The first result follows from [6, Lemma 5.1] and [5, Lemma 3.10].

LEMMA 4.1. *Suppose that the assumptions* (4.1), (4.2) *hold. If there exists an exponentially stabilizing dynamic output-feedback controller K of the form* (2.11) *with* $\|G_{zw}\| = \|\mathcal{F}(G, K)\| < 1$, *then there exists a nonnegative definite operator* $P_1 \in \mathcal{L}(X)$ *satisfying*

$$\begin{aligned} &\text{for all } x \in D(A),\ P_1x \in D(A^*), \\ &(A^*P_1 + P_1A + P_1(B_1B_1^* - B_2B_2^*)P_1 + C_1^*C_1)x = 0 \\ &\text{and } A_1 := A + (B_1B_1^* - B_2B_2^*)P_1 \text{ is exponentially stable.} \end{aligned} \tag{4.5}$$

Furthermore, the C_0-semigroup generated by $A - B_2B_2^*P_1$ is exponentially stable and the system G_I given by

$$G_I : \begin{cases} \dot{x}_I = (A - B_2B_2^*P_1)x_I + B_1w + B_2u_0, \\ z = (C_1 - D_{12}B_2^*P_1)x_I + D_{12}u_0, \quad x_I(0) = 0, \\ w_0 = -B_1^*P_1x_I + w \end{cases} \tag{4.6}$$

satisfies $\|z\|_2^2 + \|w_0\|_2^2 = \|w\|_2^2 + \|u_0\|_2^2$, *for all* $w \in L_2(0, \infty; W)$ *and* $u_0 \in L_2(0, \infty; U)$.

The second result concerns system transposition.

Suppose we have a system G of the form (2.8) and a controller K of the form (2.11). Define the transposed versions of these systems as

$$G^\natural : \begin{cases} \dot{x} = A^*x + C_1^*\tilde{w} + C_2^*\tilde{u}, \\ \tilde{z} = B_1^*x + D_{21}^*\tilde{u}, \quad x(0) = 0, \\ \tilde{y} = B_2^*x + D_{12}^*\tilde{w} \end{cases} \tag{4.7}$$

and

$$K^\natural : \begin{cases} \dot{p} = M^*p + L^*\tilde{y}, \quad p(0) = 0, \\ \\ \tilde{u} = N^*p + R^*\tilde{y}. \end{cases} \tag{4.8}$$

Note that $(K^\natural)^\natural = K$ and $(G^\natural)^\natural = G$.

LEMMA 4.2. *The following are equivalent. The controller K (as in* (2.11)*) is admissible for G (as in* (2.8)*) if and only if the controller $K^\natural$ (as in* (4.8)*) is admissible for $G^\natural$ (as in* (4.7)*).*

Proof. It is straightforward to show that a realization of $\mathcal{F}(G^\natural, K^\natural)$ is given by

$$\mathcal{F}(G^\natural, K^\natural): \begin{cases} \begin{pmatrix} \dot{x} \\ p \end{pmatrix} = \mathcal{A}^* \begin{pmatrix} x \\ p \end{pmatrix} + \mathcal{D}_1^* \tilde{w}, \\ \tilde{z} = \mathcal{B}^* \begin{pmatrix} x \\ p \end{pmatrix} + \mathcal{D}_2^* \tilde{w}, \end{cases} \tag{4.9}$$

where $\mathcal{A}, \mathcal{B}, \mathcal{D}_1$ and $\mathcal{D}_2$ are as in (2.13).

We conclude that K exponentially stabilizes G if and only if $K^\natural$ exponentially stabilizes $G^\natural$. It follows from the proof of [5, Lemma 3.9] that if K is exponentially stabilizing we have $\|\mathcal{F}(G,K)\| = \|\mathcal{F}(G,K)^\natural\|$. The result now follows from the fact that $\mathcal{F}(G,K) = \mathcal{F}(G^\natural, K^\natural)^\natural$. □

To parametrize all admissible controllers for G (given by (2.8)) we proceed as follows. Using Lemma 4.1 and Theorem 3.2, we define a transformed system denoted by G_{P_1} with the following property: a controller of the form (2.11) is admissible for G if and only if it is admissible for G_{P_1} (this corresponds to [2, Lemma 9]). If there exists an admissible controller for G_{P_1}, we can apply Lemma 4.1 once again to obtain a solution P_2 to a second Riccati equation. Then we define another system $G_{P_1P_2}$ such that a controller of the form (2.11) is admissible for G if and only if it is admissible for $G_{P_1P_2}$ (this step is not taken in [2]; it is an idea from [11]). The reason for these transformations is that $G_{P_1P_2}$ has a very nice structure that enables us to find all its admissible controllers.

We now give a lemma that characterizes all admissible controllers for a system with a structure that $G_{P_1P_2}$ is going to have.

Suppose that we have a system $\bar{G}$ given by

$$\bar{G}: \begin{cases} \dot{x} = Ax + B_1 w + B_2 u, \\ z = C_1 x + u, \qquad x(0) = 0 \\ y = C_2 x + w, \end{cases} \tag{4.10}$$

where A is the infinitesimal generator of the C_0-semigroup $T(\cdot)$ on the real separable Hilbert space X, $u(t), z(t) \in U$, $y(t), w(t) \in Y$, where U and Y are also real separable Hilbert spaces, etc. (A special case of (2.8), now $U = Z$ and $Y = W$). We also consider the description of $\bar{G}$ as in (2.9), (2.10):

$$\bar{G}: \begin{pmatrix} z \\ y \end{pmatrix} = \begin{pmatrix} \bar{G}_{11} & \bar{G}_{12} \\ \bar{G}_{21} & \bar{G}_{22} \end{pmatrix} \begin{pmatrix} w \\ u \end{pmatrix}, \tag{4.11}$$

where $\bar{G}_{ij}$ represent the corresponding linear maps. Comparing with (2.9), (2.10), we see that now the feedthrough operators of $\bar{G}_{12}$ and $\bar{G}_{21}$ are both equal to the identity so that $\bar{G}_{12}$ and $\bar{G}_{21}$ are both invertible. As before, $\bar{G}_{11}$ and $\bar{G}_{22}$ have no feedthrough operator.

Define the system $\bar{G}_2$ by

$$\bar{G}_2: \begin{cases} \dot{p}_1 = (A - B_2C_1 - B_1C_2)p_1 + B_1 y_2 + B_2 v, \\ u_2 = -C_1 p_1 + v, \qquad p_1(0) = 0, \\ r = -C_2 p_1 + y_2, \end{cases} \tag{4.12}$$

or

$$\bar{G}_2: \begin{pmatrix} u_2 \\ r \end{pmatrix} = \begin{pmatrix} \bar{G}_2^{11} & \bar{G}_2^{12} \\ \bar{G}_2^{21} & \bar{G}_2^{22} \end{pmatrix} \begin{pmatrix} y_2 \\ v \end{pmatrix}, \tag{4.13}$$

where $\bar{G}_2^{ij}$ represent the corresponding linear maps. It is straightforward to show that we have

$$\begin{pmatrix} \bar{G}_{11} & \bar{G}_{12} \\ \bar{G}_{21} & \bar{G}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \begin{pmatrix} \bar{G}_2^{11} & \bar{G}_2^{12} \\ \bar{G}_2^{21} & \bar{G}_2^{22} \end{pmatrix} \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}. \tag{4.14}$$

Let K be any controller for (4.10) of the form (4.15)

$$K : \begin{cases} \dot{p} = Mp + Ny_1, & p(0) = 0, \\ u_1 = Lp + Ry_1. \end{cases} \tag{4.15}$$

We show that there exists a system Λ of the form

$$\Lambda : \begin{cases} \dot{\lambda} = A_\Lambda \lambda + B_\Lambda u_\Lambda, & \lambda(0) = 0, \\ y_\Lambda = C_\Lambda \lambda + D_\Lambda u_\Lambda, \end{cases} \tag{4.16}$$

where A_Λ generates a C_0-semigroup $T_\Lambda(\cdot)$ on a Hilbert space Λ_1, $u_\Lambda(t) \in Y$, $y_\Lambda(t) \in U$ etc., such that the linear map K satisfies

$$K = \bar{G}_2^{11} + \bar{G}_2^{12}\Lambda(I - \bar{G}_2^{22}\Lambda)^{-1}\bar{G}_2^{21} = \mathcal{F}(\bar{G}_2, \Lambda), \tag{4.17}$$

i.e., K can be seen as the interconnection of (4.12) and (4.16) with $u_\Lambda = r$ and $y_\Lambda = v$ as in Fig. 4.1 (note that $(I - \bar{G}_2^{22}\Lambda)^{-1}$ exists because there is no feedthrough term in $\bar{G}_2^{22}$).

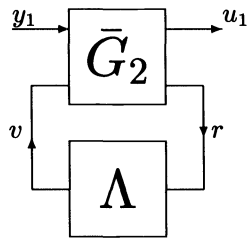


FIG. 4.1. $K = \mathcal{F}(\bar{G}_2, \Lambda)$.

Indeed, since $\bar{G}_2^{21}$ and $\bar{G}_2^{12}$ are both invertible, we can define $\tilde{\Lambda}$ as

$$\tilde{\Lambda} := (\bar{G}_2^{12})^{-1}(K - \bar{G}_2^{11})(\bar{G}_2^{21})^{-1}$$

and define Λ as

$$\Lambda := (I + \tilde{\Lambda}\bar{G}_2^{22})^{-1}\tilde{\Lambda}$$

(note that $(I + \tilde{\Lambda}\bar{G}_2^{22})^{-1}$ exists because there is no feedthrough term in $\bar{G}_2^{22}$).

Then $\tilde{\Lambda} = \Lambda(I - \bar{G}_2^{22}\Lambda)^{-1}$ and it is straightforward to show that (4.17) is satisfied. Furthermore, Λ can be realized as in (4.16).

Now let $\bar{G}_{zw}$ denote the closed-loop system determined by (4.10) and (4.15) with $y_1 = y$ and $u_1 = u$, i.e.,

$$\bar{G}_{zw} = \bar{G}_{11} + \bar{G}_{12}K(I - \bar{G}_{22}K)^{-1}\bar{G}_{21} = \mathcal{F}(\bar{G}, K).$$

The following result may seem surprising, but in fact it follows from (4.14).

LEMMA 4.3. *For all* $w \in L_2^{\mathrm{loc}}(0,\infty;W)$ *we have* $\bar{G}_{zw}w = \Lambda w$, *i.e.,*

$$\bar{G}_{zw} = \mathcal{F}(\bar{G},K) = \mathcal{F}(\bar{G},\mathcal{F}(\bar{G}_2,\Lambda)) = \Lambda \tag{4.18}$$

and

$$K = \mathcal{F}(\bar{G}_2,\Lambda) = \mathcal{F}(\bar{G}_2,\mathcal{F}(\bar{G},K)). \tag{4.19}$$

Proof. The closed-loop system $\bar{G}_{zw}$ can be described as in Fig. 4.2.

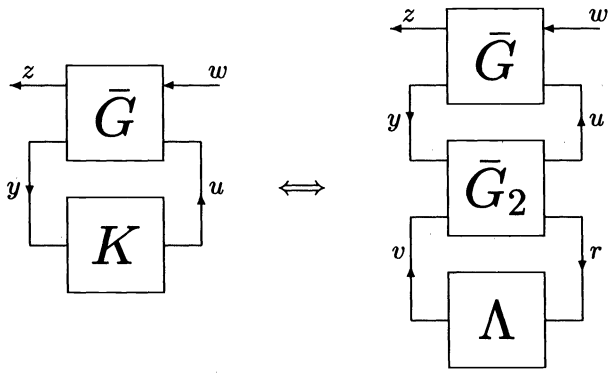


FIG. 4.2. $\bar{G}_{zw}$.

Now define $\tilde{G}$ as the interconnection of (4.10) and (4.12) with $y_2 = y$ and $u_2 = u$, as in Fig. 4.3, i.e.,

$$\tilde{G}: \begin{cases} \begin{pmatrix} \dot{x} \\ \dot{p}_1 \end{pmatrix} = \begin{pmatrix} A & -B_2C_1 \\ B_1C_2 & A - B_2C_1 - B_1C_2 \end{pmatrix} \begin{pmatrix} x \\ p_1 \end{pmatrix} \\ \qquad\qquad + \begin{pmatrix} B_1 \\ B_1 \end{pmatrix} w + \begin{pmatrix} B_2 \\ B_2 \end{pmatrix} v \\ z = (C_1 \quad -C_1) \begin{pmatrix} x \\ p_1 \end{pmatrix} + v \\ r = (C_2 \quad -C_2) \begin{pmatrix} x \\ p_1 \end{pmatrix} + w, \quad x(0) = p_1(0) = 0. \end{cases} \tag{4.20}$$

It is easy to see that

$$\begin{pmatrix} x - p_1 \\ x \end{pmatrix} = \begin{pmatrix} A - B_1C_2 & 0 \\ B_2C_1 & A - B_2C_1 \end{pmatrix} \begin{pmatrix} x - p_1 \\ x \end{pmatrix}$$

$$+ \begin{pmatrix} 0 \\ B_1 \end{pmatrix} w + \begin{pmatrix} 0 \\ B_2 \end{pmatrix} v, \tag{4.21}$$

and since $x(0) = 0$ and $p_1(0) = 0$, we see that $x - p_1 = 0$ and so

$$\begin{pmatrix} z \\ r \end{pmatrix} = \tilde{G} \begin{pmatrix} w \\ v \end{pmatrix} = \begin{pmatrix} \tilde{G}_{11} & \tilde{G}_{12} \\ \tilde{G}_{21} & \tilde{G}_{22} \end{pmatrix} \begin{pmatrix} w \\ v \end{pmatrix} = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \begin{pmatrix} w \\ v \end{pmatrix}. \tag{4.22}$$

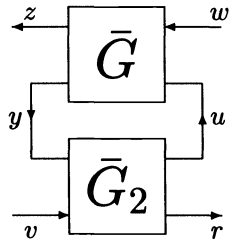


FIG. 4.3. $\tilde{G}$.

Now (4.18) follows from Figure 4.2 and (4.22), while (4.19) follows from (4.17) and (4.18). □

In fact (4.22) holds because of (4.14) and Fig. 4.3. This kind of property follows more or less from some results in [9]. There Redheffer defines the $*$-product for systems of the form $\bar{G}$ and $\bar{G}_2$ and this $*$-product corresponds to taking a feedback interconnection of $\bar{G}$ and $\bar{G}_2$ as in Fig. 4.3. A simple relation between the $*$-product inverse and the matrix inverse explains (4.22).

Using Lemma 4.3 and (4.21) we can prove the following result.

LEMMA 4.4. *Suppose that we have a controller K of the form* (4.15) *for the system $\bar{G}$ given by* (4.10) *with (M, N) exponentially stabilizable and (L, M) exponentially detectable. Furthermore, suppose that $A - B_1C_2$ and $A - B_2C_1$ both generate exponentially stable C_0-semigroups. Then K is admissible if and only if it can be realized as $\mathcal{F}(\bar{G}_2, \Lambda)$ for some Λ of the form* (4.16), *where A_Λ generates an exponentially stable C_0-semigroup and $\|\Lambda\| < 1$. In this case, $\mathcal{F}(\bar{G}_2, \Lambda)$ (with its realization on $X \times \Lambda_1$) is also itself admissible and we have $\bar{G}_{zw} = \mathcal{F}(\bar{G}, K) = \Lambda$.*

Proof. Necessity. Suppose that K of the form (4.15) is admissible for (4.10).

It follows from Lemma 4.3 that the linear map K can be expressed as $K = \mathcal{F}(\bar{G}_2, \Lambda)$ with $\Lambda = \mathcal{F}(\bar{G}, K) = \bar{G}_{zw}$. Since K is admissible, $\bar{G}_{zw}$ is of the form (2.12), where $\mathcal{A}$ is the infinitesimal generator of an exponentially stable C_0-semigroup. Hence we can realize Λ as (4.16) such that A_Λ generates an exponentially stable C_0-semigroup. Finally, $\|\Lambda\| = \|\bar{G}_{zw}\| < 1$ since K is admissible.

Sufficiency. Suppose that K of the form (4.15) allows for a realization determined by $\mathcal{F}(\bar{G}_2, \Lambda)$ with Λ of the form (4.16) such that A_Λ generates an exponentially stable C_0-semigroup and $\|\Lambda\| < 1$.

To avoid confusion we denote this realization by K_Λ, noting that $K_\Lambda y = Ky$ for all $y \in L_2^{\text{loc}}(0, \infty; Y)$).

The state-space of K_Λ is $X \times \Lambda_1$ and its realization is determined by (4.12) and (4.16) with $u_\Lambda = r$ and $y_\Lambda = v$. First we show that the closed-loop system on $X \times X \times \Lambda_1$ determined by $\bar{G}_{zw} = \mathcal{F}(\bar{G}, K_\Lambda)$ is exponentially stable and $\|\bar{G}_{zw}\| < 1$. The idea is to apply Theorem 3.2 to the right hand side of Fig. 4.2.

Using the assumption that that $A - B_1C_2$ and $A - B_2C_1$ both generate exponentially stable C_0-semigroups and (4.21), we conclude that the system $\tilde{G}$ given by (4.20) is exponentially stable (use [5, Lemmas 3.7 and 3.8], where some results of [10] are quoted). Furthermore, it follows trivially from (4.22) that $\tilde{G}$ is inner and that $(\tilde{G}_{21})^{-1}$ is i/o-stable. Now since A_Λ generates an exponentially stable semigroup and $\|\Lambda\| < 1$, we can apply Theorem 3.2 to conclude that the closed-loop system $\mathcal{F}(\bar{G}, K_\Lambda)$ on the Hilbert space $X \times X \times \Lambda_1$ is exponentially stable and that $\|\bar{G}_{zw}\| < 1$. In other words, K_Λ (with its realization on $X \times \Lambda_1$) is an admissible controller for $\bar{G}$.

We use this fact to show that K (realized as in (4.15)) is also admissible: It follows

from Lemma 2.2 that

$$\begin{pmatrix} (I-K_\Lambda\bar{G}_{22})^{-1} & (I-K_\Lambda\bar{G}_{22})^{-1}K_\Lambda \\ \bar{G}_{22}(I-K_\Lambda\bar{G}_{22})^{-1} & (I-\bar{G}_{22}K_\Lambda)^{-1} \end{pmatrix}$$

is i/o-stable and since $K_\Lambda y = Ky$ for all $y \in L_2^{\mathrm{loc}}(0,\infty;Y))$, we can replace K_Λ by K. Now the idea is to use Lemma 2.2. We have assumed that (M,N) is exponentially stabilizable and (L,M) is exponentially detectable and since $A-B_1C_2$ and $A-B_2C_1$ both generate exponentially stable semigroups, we see that also (A,B_2) is exponentially stabilizable and (C_2,A) is exponentially detectable. Hence Lemma 2.2 implies that the closed-loop system determined by (4.10) and (4.15) on the state-space $X\times P$ is exponentially stable. Now since $\|\mathcal{F}(\bar{G},K)\| = \|\mathcal{F}(\bar{G},K_\Lambda)\| = \|\bar{G}_{zw}\| < 1$, we conclude that K given by (4.15) is admissible. □

Next we define the transformed system G_{P_1} and show that it has the same admissible controllers as G, using Theorem 3.2 (this is similar to [2, Lemma 9]).

Suppose we have a system G of the form (2.8) and suppose that assumption (4.2) is satisfied (recall that without this assumption all the formulas would be more complicated). Let K again be a controller of the form (4.15). Furthermore, suppose that there exists a nonnegative definite operator $P_1 \in \mathcal{L}(X)$ that satisfies (4.5).

We define the system G_{P_1} to be

$$G_{P_1}: \begin{cases} \dot{x}_1 = (A+B_1B_1^*P_1)x_1 + B_1w_0 + B_2u, \\ u_0 = B_2^*P_1x_1 + u, \qquad x_1(0)=0 \\ y = (C_2+D_{21}B_1^*P_1)x_1 + D_{21}w_0, \end{cases} \tag{4.23}$$

or, in external representation,

$$\begin{pmatrix} u_0 \\ y \end{pmatrix} = \begin{pmatrix} {G_{P_1}}^{11} & {G_{P_1}}^{12} \\ {G_{P_1}}^{21} & {G_{P_1}}^{22} \end{pmatrix} \begin{pmatrix} w_0 \\ u \end{pmatrix}. \tag{4.24}$$

LEMMA 4.5. *We have the following equivalence. The controller K is admissible for G if and only if it is admissible for G_{P_1}.*

Proof. First, we define the auxiliary system G_I as in (4.6):

$$G_I: \begin{cases} \dot{x}_I = (A-B_2B_2^*P_1)x_I + B_1w + B_2\tilde{u}_0, \\ z = (C_1-D_{12}B_2^*P_1)x_I + D_{12}\tilde{u}_0, \qquad x_I(0)=0 \\ \tilde{w}_0 = -B_1^*P_1x_I + w \end{cases}$$

(G_I will play the role of the inner system in Theorem 3.2). It follows from Lemma 4.1 that $A-B_2B_2^*P_1$ is the infinitesimal generator of an exponentially stable C_0-semigroup and that G_I is inner. The external representation of G_I is given by

$$\begin{pmatrix} z \\ \tilde{w}_0 \end{pmatrix} = \begin{pmatrix} {G_I}^{11} & {G_I}^{12} \\ {G_I}^{21} & {G_I}^{22} \end{pmatrix} \begin{pmatrix} w \\ \tilde{u}_0 \end{pmatrix}. \tag{4.25}$$

Now ${G_I}^{21}$ is given by

$${G_I}^{21}: \begin{cases} \dot{x} = (A-B_2B_2^*P_1)x + B_1\bar{w}, \qquad x(0)=0, \\ \bar{w}_0 = -B_1^*P_1x + \bar{w}, \end{cases}$$

and it is easy to see that ${G_I}^{21}$ is invertible and that $({G_I}^{21})^{-1}$ is i/o-stable (use that $A+(B_1B_1^*-B_2B_2^*)P_1$ is exponentially stable). So we have shown that G_I satisfies the assumptions (3.9)–(3.11) of Theorem 3.2.

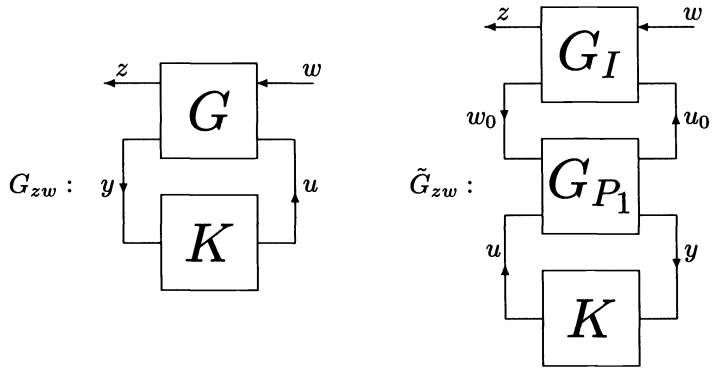


FIG. 4.4. $G_{zw} = \tilde{G}_{zw}$.

Now consider the two closed-loop systems given in Fig. 4.4, i.e., the system on the left is the interconnection of (2.8) and (4.15) with $y_1 = y$ and $u_1 = u$ and the system on the right is the interconnection of (4.6), (4.23), and (4.15) with $\tilde{w}_0 = w_0, \tilde{u}_0 = u_0, y_1 = y$, and $u_1 = u$ (recall that the system on the left is given by (2.12), (2.13)). In other words, $G_{zw} = \mathcal{F}(G, K)$ and $\tilde{G}_{zw} = \mathcal{F}(G_I, \mathcal{F}(G_{P_1}, K))$. We claim that the system on the left in Fig. 4.4 is exponentially stable if and only if the system on the right in Fig. 4.4 is and that both closed-loop maps from w to z are the same, i.e. $G_{zw} = \tilde{G}_{zw}$. Indeed, we can rewrite the state equations of the system on the right with $w = 0$ as

$$\begin{pmatrix} \dot{x_I - x_1} \\ x_1 \\ p \end{pmatrix}$$

$$= \begin{pmatrix} A + (B_1B_1^* - B_2B_2^*)P_1 & 0 & 0 \\ \times & A + B_2RC_2 & B_2L \\ \times & NC_2 & M \end{pmatrix} \begin{pmatrix} x_I - x_1 \\ x_1 \\ p \end{pmatrix}, \tag{4.26}$$

where the form of $\times$ is irrelevant. Note that the right lower 2×2-block in (4.26) represents the generator of the semigroup of the system on the left in Fig. 4.4. Since $A + (B_1B_1^* - B_2B_2^*)P_1$ generates an exponentially stable semigroup, it follows that the system on the left in Fig. 4.4 is exponentially stable if and only if the system on the right in Fig. 4.4 is exponentially stable.

Furthermore, the map $\tilde{G}_{zw}$ is determined by

$$\tilde{G}_{zw}: \begin{cases} \dot{(x_I - x_1)} &= (A + (B_1B_1^* - B_2B_2^*)P_1)(x_I - x_1), \\ \dot{x}_1 &= Ax_1 + B_1w + B_2u - B_1B_1^*P_1(x_I - x_1), \\ z &= C_1x_I + D_{12}u + D_{12}B_2^*P_1(x_1 - x_I), \\ \dot{p} &= Mp + Ny, \\ u &= Lp + Ry, \end{cases} \tag{4.27}$$

where $x_I(0) = 0, x_1(0) = 0$ and $p(0) = 0$. It follows that $(x_I - x_1) = 0$ and comparing (4.27) with (2.8) and (4.15) we see that indeed the closed-loop maps from w to z in both systems in Fig. 4.4 are the same. Therefore, K is admissible for G if and only if the system on the right in Fig. 4.4 is exponentially stable and $\|\tilde{G}_{zw}\| < 1$.

Now we can use Theorem 3.2 with G given by G_I and $G_2 = \mathcal{F}(G_{P_1}, K)$. We have already shown that G_I satisfies the assumptions (3.9)–(3.11). Hence, Theorem 3.2 implies that the system on the right in Fig. 4.4 is exponentially stable and $\|\mathcal{F}(G_I, \mathcal{F}(G_{P_1}, K))\| = \|\tilde{G}_{zw}\| < 1$ if and only if the system $\mathcal{F}(G_{P_1}, K)$ is exponentially stable and $\|\mathcal{F}(G_{P_1}, K)\| < 1$, i.e. K is admissible for G_{P_1}. This completes the proof. □

Finally, we present the main result of this section.

THEOREM 4.6. *Consider a system G of the form (2.8) and suppose that assumptions (4.1)–(4.4) are satisfied. There exists an admissible controller K (as in (4.15)) for G if and only if there exist nonnegative definite operators $P_1, P_2 \in \mathcal{L}(X)$ satisfying*

$$\begin{aligned} &\text{for all } x \in D(A),\ P_1 x \in D(A^*), \\ &(A^*P_1 + P_1A + P_1(B_1B_1^* - B_2B_2^*)P_1 + C_1^*C_1)x = 0 \\ &\text{and } A_1 := A + (B_1B_1^* - B_2B_2^*)P_1 \text{ is exponentially stable,} \end{aligned} \tag{4.5}$$

$$\begin{aligned} &\text{for all } x \in D(A^*),\ P_2 x \in D(A), \\ &((A + B_1B_1^*P_1)P_2 + P_2(A + B_1B_1^*P_1)^* + P_2(P_1B_2B_2^*P_1 - \\ &C_2^*C_2)P_2 + B_1B_1^*)x = 0 \text{ and } A_2 := A + B_1B_1^*P_1 + P_2(P_1B_2B_2^*P_1 \\ &-C_2^*C_2) \text{ is exponentially stable.} \end{aligned} \tag{4.28}$$

Moreover, in this case a controller K of the form (4.15) with (M, N) exponentially stabilizable and (L, M) exponentially detectable is admissible if and only if it can be realized as $\mathcal{F}(\tilde{K}, \Lambda)$, where $\tilde{K}$ is given by

$$\tilde{K}: \begin{cases} \dot{p}_1 = (A_1 - P_2C_2^*C_2)p_1 + P_2C_2^*y + (I + P_2P_1)B_2v, \\ u = -B_2^*P_1p_1 + v, \qquad p_1(0) = 0, \\ r = -C_2p_1 + y \end{cases} \tag{4.29}$$

and Λ is of the form

$$\Lambda: \begin{cases} \dot{\lambda} = A_\Lambda\lambda + B_\Lambda r, \qquad \lambda(0) = 0 \\ \\ v = C_\Lambda\lambda + D_\Lambda r \end{cases} \tag{4.30}$$

such that A_Λ generates an exponentially stable semigroup and $\|\Lambda\| < 1$. In this case, $\mathcal{F}(\tilde{K}, \Lambda)$ is also itself admissible.

Proof. Necessity. Suppose that K is admissible for G. It follows from Lemma 4.1 that there exists a nonnegative definite operator $P_1 \in \mathcal{L}(X)$ that satisfies (4.5). We can define G_{P_1} as in (4.23) and it follows from Lemma 4.5 that K is admissible for G_{P_1}. Lemma 4.2 implies that $K^\natural$ is admissible for $(G_{P_1})^\natural$, where $(G_{P_1})^\natural$ is given by

$$(G_{P_1})^\natural: \begin{cases} \dot{x} = (A + B_1B_1^*P_1)^*x + P_1B_2w + C_2^*u, \\ z = B_1^*x + D_{21}^*u, \qquad x(0) = 0 \\ y = B_2^*x + w \end{cases} \tag{4.31}$$

(note that now $D_{21}B_1^* = 0$).

Using assumptions (4.3), (4.4) it can be shown that $(G_{P_1})^\natural$ satisfies assumptions (4.1), (4.2), where $(G_{P_1})^\natural$ should be considered as of the form (2.8) (see [5, proof of Lemma 3.12]). Since $K^\natural$ is admissible for $(G_{P_1})^\natural$, we can therefore infer the existence of a nonnegative definite $P_2 \in \mathcal{L}(X)$ that satisfies the conditions in the theorem.

Sufficiency. Suppose that there exist nonnegative definite operators $P_1, P_2 \in \mathcal{L}(X)$ satisfying the conditions of the theorem. The existence of an admissible controller will follow from the controller parametrization part.

Controller parametrization. Suppose that there exist nonnegative definite operators $P_1, P_2 \in \mathcal{L}(X)$ satisfying the conditions of the theorem. We can define G_{P_1} as in (4.23). Then $(G_{P_1})^\natural$ is given by (4.31) and using the fact that P_2 satisfies the conditions of the theorem, we can construct $((G_{P_1})^\natural)_{P_2}$ just as we did in (4.23). Hence it follows from Lemma 4.5 that K is admissible for $(G_{P_1})^\natural$ if and only if K is admissible for $((G_{P_1})^\natural)_{P_2}$.

The transpose of $((G_{P_1})^\natural)_{P_2}$ is given by

$$(4.32)\qquad (((G_{P_1})^\natural)_{P_2})^\natural : \begin{cases} \dot{x} = ((A + B_1B_1^*P_1) + P_2P_1B_2B_2^*P_1)x \\ \qquad +P_2C_2^*w + (I + P_2P_1)B_2u, \\ z = B_2^*P_1x + u, \qquad x(0) = 0, \\ y = C_2x + w. \end{cases}$$

Now we see that $(((G_{P_1})^\natural)_{P_2})^\natural$ is of the form (4.10). We want to apply Lemma 4.4 and so we show that the system $(((G_{P_1})^\natural)_{P_2})^\natural$ satisfies the assumptions. Indeed, "$A - B_1C_2$" and "$A - B_2C_1$" are now given by $A_2 := A + B_1B_1^*P_1 + P_2(P_1B_2^*B_2P_1 - C_2^*C_2)$ and $A_1 := A + (B_1B_1^* - B_2B_2^*)P_1$. A_1 is stable because P_1 is the stabilizing solution of the Riccati equation (4.5) and A_2 is stable because P_2 is the stabilizing solution of the Riccati equation (4.28).

Hence we can apply Lemma 4.4 and it follows that a controller K of the form (4.15) with (M, N) exponentially stabilizable and (L, M) exponentially detectable is admissible for $(((G_{P_1})^\natural)_{P_2})^\natural$ if and only if it can be realized as $\mathcal{F}(\tilde{G}_2, \Lambda)$, where $\tilde{G}_2$ is constructed as $\tilde{G}_2$ in (4.12):

$$(4.33)\qquad \tilde{G}_2 : \begin{cases} \dot{p}_1 = (A + (B_1B_1^* - B_2B_2^*)P_1 - P_2C_2^*C_2)p_1 \\ \qquad + P_2C_2^*y + (I + P_2P_1)B_2v, \\ u = -B_2^*P_1p_1 + v, \qquad p_1(0) = 0, \\ r = -C_2p_1 + y, \end{cases}$$

and Λ is of the form (4.30) such that A_Λ generates an exponentially stable semigroup and $\|\Lambda\| < 1$.

Finally, the result follows from Lemmas 4.2 and 4.5. K is admissible for G if and only if K is admissible for G_{P_1} if and only if $K^\natural$ is admissible for $(G_{P_1})^\natural$ if and only if $K^\natural$ is admissible for $((G_{P_1})^\natural)_{P_2}$ if and only if K is admissible for $(((G_{P_1})^\natural)_{P_2})^\natural$. □

5. Conclusions and final remarks. In this paper we have given an infinite-dimensional time-domain version of Redheffer's lemma. The result that we have presented is a generalization of a finite-dimensional result that has been crucial in the theory of H_∞-control with measurement-feedback for finite-dimensional systems. The proof depends solely on time-domain techniques.

Using this result we have given a nice derivation of the parametrization of all controllers that solve the suboptimal regular H_∞-control problem with measurement-feedback for a large class of infinite-dimensional systems, thereby generalizing the finite-dimensional result.

Our remark in the introduction that this type of approach would be suitable for nonlinear and time-varying systems has recently been vindicated. In [8], a different approach was used to solve the finite-dimensional time-varying case.

Acknowledgment. I thank Professor Ruth Curtain for some useful comments and suggestions.

REFERENCES

[1] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite-dimensional linear systems theory*, Lecture Notes in Control and Inform. Sci., Vol. 8, Springer Verlag, Berlin, 1978.

[2] J. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.

[3] K. GLOVER AND J. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[4] C. A. JACOBSON AND C. N. NETT, *Linear state-space systems in infinite dimensional space: the role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 541–549.

[5] B. A. M. VAN KEULEN, *The H_∞-problem with measurement feedback for linear infinite-dimensional systems*, J. Math. Systems, Estim. and Control, to appear.

[6] B. A. M. VAN KEULEN, M. PETERS, AND R. F. CURTAIN, *H_∞-control with state-feedback: the infinite-dimensional case*, J. Math. Systems, Estim. Control, 3 (1993), pp. 1–39.

[7] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer Verlag, New York, 1983.

[8] R. RAVI, K. M. NAGPAL, AND P. P. KHARGONEKAR, *H_∞-control of linear time-varying systems: a state-space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.

[9] R. M. REDHEFFER, *On a certain linear fractional transformation*, J. Math. Phys., 39 (1960), pp. 269–286.

[10] H. SCHUMACHER, *Dynamic feedback in finite- and infinite-dimensional linear systems*. Math. Centre (CWI) Tracts 143, Amsterdam, 1981.

[11] A. A. STOORVOGEL, *The H_∞-Control Problem: A State-Space Approach*, Prentice Hall, New York, 1992.

[12] G. TADMOR, *Worst-case design in the time domain. the maximum principle and the standard H_∞-problem*, Math. Control, Signals, Systems, 3 (1990), pp. 301–324.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1, pp. 279–296, January 1994

016

ON DYNAMIC FEEDBACK COMPENSATION AND COMPACTIFICATION OF SYSTEMS*

JOACHIM ROSENTHAL†

Abstract. This paper introduces a compactification of the space of proper $p \times m$ transfer functions with a fixed McMillan degree n. Algebraically, this compactification has the structure of a projective variety and each point of this variety can be given an interpretation as a certain autoregressive system in the sense of Willems. It is shown that the pole placement map with dynamic compensators turns out to be a central projection from this compactification to the space of closed-loop polynomials. Using this geometric point of view, necessary and sufficient conditions are given when a strictly proper or proper system can be generically pole assigned by a complex dynamic compensator of McMillan degree q.

Key words. multivariable systems, dynamic feedback compensation, compactification, central projection, autoregressive systems

AMS subject classifications. 93B55, 93C35, 93B25, 54D35, 14M15

1. Introduction.

In this paper we investigate the pole placement problem with dynamic compensators from a geometric point of view. For this consider a multivariable, time invariant linear system Σ_n of order n with m-inputs and p-outputs. Such a system can be represented with its state space representation

$$\Sigma_n \ : \dot{x} = Ax + Bu, \qquad y = Cx. \tag{1.1}$$

From an engineering point of view, an input–output description is natural. Mathematically, this can be achieved by taking the Laplace transform. The system Σ_n is then described in the frequency domain by the following equation:

$$\hat{y} = C(sI - A)^{-1}B\,\hat{u}. \tag{1.2}$$

The strictly proper rational matrix $G(s) := C(sI - A)^{-1}B$ is called the transfer function associated to the system Σ_n. It is well known that the dynamics of the system Σ_n depends in an essential way on the location of the poles of the transfer function $G(s)$, which are exactly the eigenvalues of the matrix A. A fundamental open problem in multivariable linear system theory is the following question: Under which conditions can a p-input, m-output system $F(s)$ of McMillan degree q be constructed that stabilizes the closed-loop system $G_F(s) := (I - G(s)F(s))^{-1}G(s)$? More generally, we can ask the following question: Given an arbitrary polynomial $\phi(s) = s^{n+q} + \lambda_{n+q-1}s^{n+q-1} + \cdots + \lambda_0$, under which conditions is it always possible to find a compensator of order q such that the poles of the closed-loop system $G_F(s)$ are exactly the roots of the polynomial $\phi(s)$? Willems and Hesselink [37] called a system $G(s)$ with this property pole assignable in the class of feedback controllers of order q. Using a dimension argument they showed that

$$q(m+p) + mp \geq n + q \tag{1.3}$$

* Received by the editors October 28, 1991; accepted for publication (in revised form) August 28, 1992.

† Department of Mathematics, University of Notre Dame, Notre Dame, Indiana 46556 (Joachim.Rosenthal@nd.edu). This research was supported in part by National Science Foundation grant DMS-9201263.

is a necessary condition for any system $G(s)$ to have the pole assignability property in the class of feedback controllers of order q. In this paper we will show the new result that this numerical condition is not only necessary but also sufficient for a generic system $G(s)$ if the base field is algebraically closed. To establish this result, we will study for a generic system $G(s)$ the associated *pole placement map* ρ_G. The domain of ρ_G is the space of proper transfer functions of McMillan degree q and the range of ρ_G is the space of monic polynomials of degree $n+q$. In this language the system $G(s)$ has the pole assignability property in the class of feedback controllers of order q if and only if ρ_G is onto.

The question of pole assignability is fairly well understood if we restrict ourselves to the class of static compensators, in other words, compensators with McMillan degree $q=0$, and if we assume that the base field is algebraically closed. In this case we know that $mp \geq n$ is a necessary and sufficient condition for the pole placement map ρ_G to be onto generically. Indeed, Hermann and Martin [13] first showed that ρ_G is almost onto using the dominant morphism theorem. Brockett and Byrnes [2] later showed that ρ_G is even onto and the mapping degree of ρ_G in the case $mp=n$ is equal to the degree of the Grassmann variety $\mathrm{Grass}(p, p+m)$.

The pole placement problem with dynamic compensators ($q>0$) is much less understood. The following result of Brash and Pearson [1], published 1970, is still one of the strongest results available. For the generic situation their result can be quoted in the following manner. (See, e.g., [3].)

THEOREM 1.1 (Brash and Pearson [1]). *The generic degree n linear system with m–inputs and p–outputs can be arbitrarily pole assigned (over any field) using a compensator of order q, where q is any natural number satisfying*

$$\max(m,p)(q+1) \geq n. \tag{1.4}$$

It is interesting to see that the necessary condition $q(m+p)+mp \geq n+q$ of Willems and Hesselink [37] is also sufficient as soon as $\min(m,p)=1$. In §4 we will explain that this is essentially due to the fact that the space of proper transfer functions with fixed McMillan degree is a Zariski open subset of a projective space if $\min(m,p)=1$ and ρ_G is a linear map from this projective space to the space of closed-loop polynomials identified with a projective space as well. If $\min(m,p)>1$, however, this is not the case and ρ_G is a rather complicated morphism.

An important contribution to understanding the pole placement problem in general was done by Byrnes [4]. In this paper Byrnes introduced a compactification for the quasi-projective variety of proper transfer functions of degree q, which he denoted by $C^q_{m,p}$. He then explained the pole placement problem as an intersection problem in $C^q_{m,p}$. Using this point of view he achieved new results for pole assignment with compensators of degree $q=1$ not achieved by any other means. Our approach is guided in part by the philosophy of this paper and that is one of the reasons why we have chosen a similar title.

A great deal of research was devoted to the question of understanding the pole placement problem with static compensators over the reals. In 1975 Kimura [16] proved the result that $m+p-1 \geq n$ is a sufficient condition for the pole placement map ρ_G to be generically onto. Since that time, several authors have improved his results and methods in different directions. Using a geometric approach Wang [35] very recently achieved the strong result that the pole placement map ρ_G is generically onto over the reals as soon as $mp-1 \geq n$. A crucial part in Wang's proof is the fact that the pole placement map ρ_G is a central projection when $q=0$. As we will show in this paper, the same is true in general.

The paper is structured as follows. After explaining some mathematical preliminaries in §2, we will introduce a projective variety in §3 that can be viewed as a compactification of the space of proper $p \times m$ transfer functions of McMillan degree n that we denote with $K^n_{p,m}$. This compactification was originally introduced by Rosenthal in [27] and used in [28] to achieve new results for certain low-dimensional feedback problems. In Theorem 3.6 we will describe the defining equations of the variety $K^n_{p,m}$ and in Theorem 3.10 we give an interpretation in terms of certain autoregressive systems.

In §4 the pole placement problem is formulated in a geometric language. To deal with compensators that are not admissible, the notion of *q–degeneracy*, a generalization of the concept of *degeneracy* [2], is introduced. We will show that for a q–nondegenerate plant, the pole placement map can be extended in a continuous manner to the whole compactification.

The main results of the paper are given in §5. It is first shown that the q–degenerate systems form an algebraic subset in the quasi-projective variety of transfer functions. Then necessary and sufficient conditions are given when the q–nondegenerate systems are generic. We will show that the pole placement map is a central projection. Using this fact we are able to formulate conditions when the pole placement map for a q–nondegenerate strictly proper (or proper) system is onto (almost onto). These results constitute a generalization of the results of Hermann and Martin [13] and Brockett and Byrnes [2] from the problem of static to the problem of dynamic feedback compensation.

Finally some words about the base field. Most constructions we do in §§3 and 4 can be done over an arbitrary field $\mathbb{K}$. For most applications, of course, the relevant base fields are the real or complex numbers, i.e., $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. The results in §5 will use the projective dimension theorem (see, e.g., [9]) and this theorem is only valid if the field is algebraically closed.

2. Preliminaries. Let $\mathbb{K}$ be an arbitrary field. With $\bar{\mathbb{K}}$ we will denote the algebraic closure of $\mathbb{K}$. If V is a $\mathbb{K}$–vector space, we will denote with $\mathbb{P}(V)$ the set of one-dimensional subspaces of V. $\mathbb{P}(V)$ is called the projective space associated to V. A topology defined on V induces a topology on $\mathbb{P}(V)$, namely, the quotient topology of the canonical projection $pr : V - \{0\} \longrightarrow \mathbb{P}(V)$. As it is well known, $\mathbb{P}(\mathbb{C}^{n+1})$ and $\mathbb{P}(\mathbb{R}^{n+1})$ are compact manifolds with the induced topology coming from the natural topology on $\mathbb{C}^{n+1}$ or $\mathbb{R}^{n+1}$. If $V = \mathbb{K}^{n+1}$ we sometimes use the notation $\mathbb{P}^n_{\mathbb{K}}$ or simply $\mathbb{P}^n$.

We will identify $\mathbb{K}^n$ as a subset of $\mathbb{P}^n$ using the inclusion:

$$i : \mathbb{K}^n \longrightarrow \mathbb{P}^n_{\mathbb{K}}, \qquad (x_1, \ldots, x_n) \longmapsto (x_1, \ldots, x_n, 1). \tag{2.1}$$

In particular $\mathbb{K}$ is identified with $\{(x, 1) \mid x \in \mathbb{K}\} \subset \mathbb{P}^1_{\mathbb{K}}$. We will call the point $(1, 0)$, which is the only point in the difference set $\mathbb{P}^1_{\mathbb{K}} - i(\mathbb{K})$, the point at infinity and $\mathbb{P}^1_{\mathbb{K}}$ *the projective line* over $\mathbb{K}$.

Consider now the polynomial ring $\mathbb{K}[s]$ in one indeterminate. Assume the set of polynomials $\{f_1(s), \ldots, f_{n+1}(s)\} \subset \mathbb{K}[s]$ has no common zeroes. Then the map

$$f : \mathbb{K} \longrightarrow \mathbb{P}^n_{\mathbb{K}}, \qquad s \longmapsto (f_1(s), \ldots, f_{n+1}(s)) \tag{2.2}$$

is well defined and called a *rational map*. The *degree* of f is defined by the highest degree of the polynomials $f_i(s)$. Assume f has degree d. The homogenization of $f(s)$

is defined by

$$\hat{f}(s,t) := t^d f\left(\frac{s}{t}\right). \tag{2.3}$$

Note that $\hat{f}$ extends the rational map f to the whole projective line $\mathbb{P}^1_{\mathbb{K}}$. Moreover, if $\mathbb{K}$ is algebraically closed, the image $\mathrm{Im}(\hat{f})$ defines a rational curve in $\mathbb{P}^n_{\mathbb{K}}$ in the sense of algebraic geometry. Note that over the complex numbers the holomorphic maps from the Riemann sphere $\mathbb{P}^1_{\mathbb{C}}$ to the complex projective space $\mathbb{P}^n_{\mathbb{C}}$ are exactly the rational maps corresponding to our definition.

The degree d of the rational map f has the following geometric interpretation: Intersect the curve $\mathrm{Im}(\hat{f})$ with a generic linear hyperplane H in $\mathbb{P}^n_{\mathbb{K}}$, which can be described by a homogeneous linear equation of the form $\sum c_i x_i = 0$. By the fundamental theorem of algebra, H intersects $\mathrm{Im}(\hat{f})$ over the algebraic closure $\bar{\mathbb{K}}$ in exactly d points when counted with multiplicities. In short, the variety $\mathrm{Im}(\hat{f})$ has degree d.

Denote with $\mathrm{Rat}_d(\mathbb{P}^1, \mathbb{P}^n)$ the set of all rational maps of degree d. $\mathrm{Rat}_d(\mathbb{P}^1, \mathbb{P}^n)$ can be exhibited as a Zariski open set in $\mathbb{P}(\mathbb{K}^{d+1} \otimes \mathbb{K}^{n+1})$. For this consider a particular embedding

$$\begin{aligned} \tau &: \mathrm{Rat}_d(\mathbb{P}^1, \mathbb{P}^n) \longrightarrow \mathbb{P}(\mathbb{K}^{d+1} \otimes \mathbb{K}^{n+1}) \\ &\left(\sum_{j=0}^{d} a_{1j} s^j, \ldots, \sum_{j=0}^{d} a_{(n+1)j} s^j\right) \longmapsto (a_{10}, \ldots, a_{1d}, a_{20}, \ldots, \ldots, a_{(n+1)d}). \end{aligned} \tag{2.4}$$

The complement of the image of $\mathrm{Rat}_d(\mathbb{P}^1, \mathbb{P}^n)$ under τ in $\mathbb{P}(\mathbb{K}^{d+1} \otimes \mathbb{K}^{n+1})$ is an algebraic set already described around the turn of the century by Macaulay [21]. If $n = 1$ this algebraic set is a hypersurface described by the well-known resultant locus of two polynomials:

$$\det \mathrm{Res}(f_1, f_2) = 0. \tag{2.5}$$

A natural generalization of the projective space is the Grassmann variety. Consider again a $\mathbb{K}$–vector space V. The set of p–dimensional subspaces in V is called the Grassmann variety which we will denote by $\mathrm{Grass}(p, V)$. If $V = \mathbb{K}^n$ we will just write $\mathrm{Grass}(p, n)$. In particular, we have $\mathrm{Grass}(1, n) = \mathbb{P}^{n-1}$.

The set $\mathrm{Grass}(p, n)$ indeed has the structure of a projective variety. For this consider the Plücker embedding φ of the Grassmann variety $\mathrm{Grass}(p, n)$, which is defined in the following way:

$$\begin{aligned} \varphi &: \mathrm{Grass}(p, n) \longrightarrow \mathbb{P}(\wedge^p \mathbb{K}^n), \\ &\mathrm{span}(v_1, \ldots, v_p) \longmapsto v_1 \wedge \cdots \wedge v_p. \end{aligned} \tag{2.6}$$

It is easy to verify that φ is an embedding. Moreover, $\mathrm{Im}(\varphi)$ is irreducible and described by a famous set of quadratic relations sometimes called "shuffle relations." (See, e.g., the survey article [17] or [25] for a characteristic free approach.) Finally we say a map $h : \mathbb{K} \longrightarrow \mathrm{Grass}(p, n)$ is a rational map if $f := \varphi \circ h$ is rational according to the definition above.

The set of all rational maps of degree d from the projective line $\mathbb{P}^1_{\mathbb{K}}$ to the Grassmann variety $\mathrm{Grass}(p, n)$ will be denoted by $\mathrm{Rat}_d(\mathbb{P}^1_{\mathbb{K}}, \mathrm{Grass}(p, n))$.

3. A compactification of the space of proper transfer functions. In the following denote with $S^n_{p,m}$ the space of proper $p \times m$ transfer functions of McMillan degree n. Algebraically, the set $S^n_{p,m}$ has the structure of a quasi-projective variety of dimension $n(m+p)+mp$. This follows directly from the fact that the space $\mathrm{Rat}_{n,m,p}$ of strictly proper transfer functions is quasi-projective (even quasi-affine) [10], has dimension $n(m+p)$, and $S^n_{p,m} \cong \mathrm{Rat}_{n,m,p} \times \mathbb{K}^{mp}$. Analytically, i.e., over the complex numbers, it is well known that $S^n_{p,m}$ and $\mathrm{Rat}_{n,m,p}$ are both connected complex manifolds. Many authors already studied topological properties of the spaces $S^n_{p,m}, \mathrm{Rat}_{n,m,p}$, and very recently Mann and Milgram [22] introduced a new stratification of $\mathrm{Rat}_{n,m,p}$ enabling them to calculate the additive structure of the homology ring $H_*(\mathrm{Rat}_{n,m,p})$.

If we consider feedback problems with high gain compensators or if we want to understand partial system failures, it is of ample importance to understand the boundary structure of the space $S^n_{p,m}$. Motivated by those problems, several authors (e.g. [4], [8], [11], [12], [20], [26], [29]) considered the problem of compactifying the space $S^n_{p,m}$. In this section we will describe a compactification of the space $S^n_{p,m}$, which turns out to be suitable for the study of dynamic feedback compensation. The basic idea is to embed $S^n_{p,m}$ into a projective space. The closure of the image with respect to the Zariski topology serves as a compactification. Our approach is geometric, indeed, we will view each transfer function $G(s) \in S^n_{p,m}$ as a rational curve of degree n into a Grassmann variety. In other words we will identify each $G(s)$ with its Hermann–Martin curve [23]. Because this curve is of crucial importance for all that follows and because we want to develop our theory over an arbitrary field $\mathbb{K}$, we explain this concept in more detail.

Consider a left coprime factorization $D_L^{-1}(s)N_L(s) = G(s)$, where $D_L(s)$ and $N_L(s)$ are polynomial matrices. The following results are well known and proofs can be found, for example, in [5]. From coprimeness it follows that the $p \times (m+p)$ polynomial matrix $(\, N_L(s)\, D_L(s)\,)$ is of full rank for all $s \in \bar{\mathbb{K}}$. If $\tilde{D}_L^{-1}(s)\tilde{N}_L(s) = G(s)$ is a second coprime factorization, then there is a $p \times p$ unimodular matrix $U(s)$, i.e., $U(s) \in Gl_p(\mathbb{K}[s])$, with $(\, \tilde{N}_L(s)\, \tilde{D}_L(s)\,) = U(s)\,(\, N_L(s)\, D_L(s)\,)$; in other words, $(\, \tilde{N}_L(s)\, \tilde{D}_L(s)\,)$ is row equivalent to $(\, N_L(s)\, D_L(s)\,)$. From these remarks it now follows that every element $s \in \mathbb{K}$ is assigned a p–dimensional subspace in $\mathbb{K}^{m+p}$, namely, the rowspace of $(\, N_L(s)\, D_L(s)\,)$. Identifying each subspace with a point of the Grassmann variety $\mathrm{Grass}(p, m+p)$ we get a well-defined map h that is independent of the selected coprime factorization and just depends on the transfer function $G(s)$:

$$h : \mathbb{K} \longrightarrow \mathrm{Grass}(p, m+p), \qquad s \longmapsto \mathrm{rowsp}(\, N_L(s)\, D_L(s)\,). \tag{3.1}$$

DEFINITION 3.1. The map h is called the Hermann–Martin map associated to the transfer function $G(s)$.

We will show that h is a rational map and $\mathrm{Im}(h)$ describes a rational curve in the sense of algebraic geometry. It is not hard to see that two different transfer functions $G(s)$ and $\tilde{G}(s)$ give rise to two different maps. In this way, the space $S^n_{p,m}$ is embedded into the space of rational maps into the Grassmannian $\mathrm{Grass}(p, m+p)$. As pointed out by Martin and Hermann [23], it is possible to extend h to "infinity" if we consider a strictly proper transfer function $G(s)$ and if we work over the complex numbers. In the case of an arbitrary field $\mathbb{K}$, we can do something similar. Moreover, we do not have to restrict our considerations to strictly proper transfer functions. We will contemplate the following general setting.

Denote with $P_{p,m}$ the space of all $p \times (m+p)$ full rank polynomial matrices

$P(s)$. We say two elements $P(s)$, $\tilde{P}(s)$ in $P_{p,m}$ are (row) equivalent if there is a $p \times p$ unimodular matrix $U(s) \in Gl_p(\mathbb{K}[s])$ with the property that $\tilde{P}(s) = U(s)P(s)$. Every $p \times (m+p)$ polynomial matrix defines a system of autoregressive equations of the form

$$(P(s)) \cdot \begin{pmatrix} u \\ y \end{pmatrix}(s) \;=\; 0. \tag{3.2}$$

If $u(s)$ and $y(s)$ are solutions from the space of rational functions, it is clear that equivalent systems have the same solution set. Using the language of Willems [38], [39] (compare also with [18], [30]) we call an equivalence class in $P_{p,m}$ an *autoregressive system* and the solution set the behavior of the system. Not all autoregressive systems actually describe a left factorization of a transfer function because the last minor of $P(s)$ is not necessarily invertible. However, if the polynomial matrix $P(s)$ can be partitioned into $(P_1(s)\, P_2(s))$ with $P_2(s) \in Gl_p(\mathbb{K}(s))$ (this is the generic situation), $P(s)$ defines a proper or improper transfer function $G(s) := P_2^{-1}(s)P_1(s)$ and equivalent systems define the same transfer function.

As shown by Kuijper and Schumacher [18], [19] it is always possible to realize an autoregressive system by a not necessarily regular descriptor system of the form

$$E\dot{x} = Ax + Bu,\; y = Cx + Du. \tag{3.3}$$

An autoregressive system $P(s)$ is called irreducible or controllable if $P(s)$ has full rank for all $s \in \bar{\mathbb{K}}$. (Compare with [7], [14], [30], [39].) Every irreducible autoregressive system $P(s) \in P_{p,m}$ gives rise to a rational map

$$h:\; \mathbb{K} \longrightarrow \mathrm{Grass}(p, m+p); \qquad s \longmapsto \mathrm{rowsp}P(s) \tag{3.4}$$

and this map depends only on the equivalence class in $P_{p,m}$. (Compare with [7].) In the following we extend h to the whole projective line $\mathbb{P}^1_{\mathbb{K}}$; in other words, we extend h to "infinity."

Without loss of generality we assume $P(s)$ is row reduced (see, e.g., [15]). Denote with $h_i(s)$ the ith row of $P(s)$ and with ν_i the degree of the polynomial vector $h_i(s)$, i.e., the highest degree of all polynomial entries. Consider the homogenization

$$\hat{h}_i(s,t) := t^{\nu_i} h_i\left(\frac{s}{t}\right) \tag{3.5}$$

Denote with $\hat{P}(s,t)$ the matrix constructed from the rows $\hat{h}_i(s,t)$. In this way we receive an extended Hermann–Martin map $\hat{h}$:

$$\hat{h}\;:\mathbb{P}^1_{\mathbb{K}} \longrightarrow \mathrm{Grass}(p, m+p), \qquad (s,t) \longmapsto \mathrm{rowsp}\hat{P}(s,t). \tag{3.6}$$

The map $\hat{h}$ is in fact rational. For this consider the Plücker embedding φ of the Grassmann variety $\mathrm{Grass}(p, m+p)$ as defined in (2.6). The combined map $\hat{f} = \varphi \circ \hat{h}$ is given by $\hat{f}(s,t) := \hat{h}_1(s,t) \wedge \cdots \wedge \hat{h}_p(s,t)$. Because the entries of $\hat{f}(s,t)$ are the principal minors of $\hat{P}(s,t)$, it is immediate that $\hat{f}(s,t)$ is homogeneous in (s,t) of degree $n = \sum \nu_i$. In other words,

$$\hat{f}\;:\;\mathbb{P}^1_{\mathbb{K}} \longrightarrow \mathbb{P}(\wedge^p \mathbb{K}^{m+p}) \tag{3.7}$$

defines a rational map. Finally note that $\hat{f}(s,t)$ is exactly the homogenization of $f(s) := h_1(s) \wedge \cdots \wedge h_p(s)$.

Note that the McMillan degree of a proper or even improper transfer function $G(s)$ represented by a coprime factorization $D_L^{-1}(s)N_L(s) = G(s)$ is equal to the highest degree of the principal minors of the matrix $\begin{pmatrix} N_L(s) \\ D_L(s) \end{pmatrix}$ (see, e.g., [7], [14]). Based on this fact we define the McMillan degree of an autoregressive system in the following way.

DEFINITION 3.2. The McMillan degree of an autoregressive system $P(s)$ is given by the maximal degree of the full size minors of $P(s)$.

We are now in a position to describe a compactification of $S_{p,m}^n$, the quasi-projective variety of proper $p \times m$ transfer functions of McMillan degree n. The Hermann–Martin identification gives rise to an embedding of $S_{p,m}^n$ into the space of rational maps $\mathrm{Rat}_n(\mathbb{P}^1_{\mathbb{K}}, \mathrm{Grass}(p, m+p))$. Using the Plücker embedding (2.6), this set can be identified with a set of rational maps into a projective space, and, as outlined earlier, this set is contained in a Zariski open set of a projective space. All those maps can be summarized by the following diagram of maps [27]:

$$\begin{aligned} S_{p,m}^n &\xrightarrow{\text{Her.–Mar.}} \mathrm{Rat}_n(\mathbb{P}^1, \mathrm{Grass}(p, m+p)) \\ &\xrightarrow{\text{Plücker}} \mathrm{Rat}_n(\mathbb{P}^1, \mathbb{P}(\wedge^p \mathbb{K}^{m+p})) \\ &\xrightarrow{\tau} \mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p}). \end{aligned} \tag{3.8}$$

DEFINITION 3.3. $K_{p,m}^n$ is defined as the Zariski closure of $S_{p,m}^n$ in $\mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p})$.

$K_{p,m}^n$ is an algebraic set of a projective space by definition. Over the reals ($\mathbb{K} = \mathbb{R}$) or over the complex numbers ($\mathbb{K} = \mathbb{C}$) we have already mentioned in §2 that $\mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p})$ is compact with the induced topology. In this way we can view $K_{p,m}^n$ as a compactification of $S_{p,m}^n$. Note also that $K_{p,m}^0 = \mathrm{Grass}(p, m+p)$. In other words, our compactification reduces to the Grassmannian model already widely used to study static output feedback problems (see, e.g., [3], [34]). The following theorem states that $K_{p,m}^n$ is a projective variety for all natural numbers m, p, n.

THEOREM 3.4. *$K_{p,m}^n$ is a projective variety of dimension $n(m+p)+mp$. If $S_{p,m}^n$ is irreducible then $K_{p,m}^n$ is irreducible as well.*

Proof. Because $S_{p,m}^n$ is quasi-projective the dimension of $S_{p,m}^n$ and its Zariski closure $K_{p,m}^n$ are the same. The irreducibility of $K_{p,m}^n$ follows directly from the irreducibility of $S_{p,m}^n$. Indeed, consider a decomposition $K_{p,m}^n = Y^1 \bigcup Y^2$ into Zariski closed subsets. Then $S_{p,m}^n = (S_{p,m}^n \cap Y^1) \bigcup (S_{p,m}^n \cap Y^2)$. By irreducibility of $S_{p,m}^n$ it follows that $S_{p,m}^n \subset Y^1$ or $S_{p,m}^n \subset Y^2$. But then $K_{p,m}^n$ is also contained in one of the sets Y^1, Y^2. □

Remark 3.5. 1. Over an algebraically closed field $S_{p,m}^n$ is always irreducible [10].

2. By a dimension argument it is clear that $K_{p,m}^n$ is a proper subset of $\mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p})$ as soon as $\min(m, p) \geq 2$. On the other hand, we have:

$$K_{1,m}^n \cong K_{m,1}^n \cong \mathbb{P}^{mn+m+n}. \tag{3.9}$$

In the following we want to describe a specific set of equations that generate the homogeneous ideal $I(K_{p,m}^n)$. For this consider a polynomial vector $f(s) \in \mathrm{Rat}_n(\mathbb{P}^1, \mathbb{P}(\wedge^p \mathbb{K}^{m+p}))$ and expand it in terms of its Plücker coordinates with respect to the standard basis:

$$f(s) = \sum_{\underline{i} \in \binom{m+p}{p}} f_{\underline{i}}(s) \cdot e_{i_1} \wedge \cdots \wedge e_{i_p}. \tag{3.10}$$

To say the map $f(s)$ factors over the Grassmannian it is necessary that the Plücker coordinates satisfy the "shuffle relations" (QR) (see, e.g., [17] or [25]), when considered as equations of the polynomial ring $\mathbb{K}[s]$:

$$(QR) \qquad \sum_{\lambda=1}^{p+1}(-1)^{\lambda} \cdot f_{i_1,\ldots,i_{p-1},j_\lambda}(s) \cdot f_{j_1,\ldots,\hat{j_\lambda},\ldots,j_{p+1}}(s) = 0. \tag{3.11}$$

In these equations, $i_1,\ldots,i_{p-1}$ and $j_1,\ldots,j_{p+1}$ are any sequence of integers with $1 \leq i_\alpha, j_\beta \leq m+p$ and the symbol $\hat{}$ means that j_λ must be removed (compare [17]). As shown in [25] the quadratic equations (QR) generate the homogeneous ideal if the base field is arbitrary but infinite. Equating polynomial coefficients we receive a set of necessary quadratic equations in $\mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p})$. The following theorem states that those equations are also sufficient; in other words, they really "cut out" $K^n_{p,m}$.

THEOREM 3.6. *Let $\mathbb{K}$ be an infinite field. Then the variety $K^n_{p,m} \subset \mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p})$ is the zero set of the ideal generated by the set of quadratic relations obtained from equating the coefficients in the shuffle relations (QR).*

Proof. Denote with $\wp$ the homogeneous ideal generated by the equations obtained when equating (QR). Because the polynomials of $\wp$ vanish on $S^n_{p,m}$, i.e., $\wp \subseteq I(S^n_{p,m})$, it follows for the sets of zeros that $Z(I(S^n_{p,m})) = K^n_{p,m} \subseteq Z(\wp)$. It therefore remains to show that $Z(\wp) \subseteq K^n_{p,m}$. For this consider in $\mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p})$ the Zariski open subset Y corresponding to all polynomial vectors $f(s) = (\ldots, f_i(s), \ldots)$ that have the property that $f(s) \neq 0$ for all $s \in \bar{\mathbb{K}}$ and that have the property that the last Plücker coordinate $f_{m+1,\ldots,m+p}(s)$ has degree n. Assume now that a point $f(s) \in Y \subset \mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p})$ satisfies the equations coming from (QR) for all $s \in \bar{\mathbb{K}}$. Viewing the entries of $f(s)$ as elements in the field $\mathbb{K}(s)$, it is immediate that there is a rational $p \times (m+p)$ matrix $R(s)$ that is mapped under the Plücker embedding on the vector $f(s)$. Using the row reduction process introduced by Forney [6], we find a $p \times (m+p)$ polynomial matrix $P(s)$ with minimal row indices and a rational matrix $Q(s) \in Gl_p(\mathbb{K}(s))$ with $P(s) = Q(s)R(s)$. The Plücker coordinates $p(s)$ of the polynomial matrix $P(s)$ are clearly given by $p(s) = \det Q(s)\, f(s)$. However, it then follows from the assumptions we made that $\det Q(s) \in \mathbb{K}$, the last entry of $p(s)$ is a polynomial of degree n, and $P(s)$ is mapped onto $f(s)$ viewed as a point of projective space. In other words, $f(s)$ describes a point of $S^n_{p,m}$. In short, $Z(\wp) \cap Y \subseteq S^n_{p,m}$, but it is then clear that $Z(\wp) \subseteq K^n_{p,m}$. □

The following example illustrates how it is possible to find a describing set of equations in a concrete case.

Example 3.7 (see [27]). $K^1_{2,2} \hookrightarrow \mathbb{P}^{11}$ is the complete intersection of three quadrics.

Indeed, $K^1_{2,2}$ is defined by $\{(f_{1,2}(s),\ldots,f_{3,4}(s)) \mid f_{1,2}(s)f_{3,4}(s) - f_{1,3}(s)f_{2,4}(s) + f_{1,4}(s)f_{2,3}(s) \equiv 0$ and $f_{i,j}(s) = a_{i,j} + b_{i,j}s\ 1 \leq i < j \leq 4\}$. In $\mathbb{P}^{11}$, we therefore have the following equations:

$$a_{1,2}a_{3,4} - a_{1,3}a_{2,4} + a_{1,4}a_{2,3} = 0, \tag{3.12}$$

$$b_{1,2}b_{3,4} - b_{1,3}b_{2,4} + b_{1,4}b_{2,3} = 0, \tag{3.13}$$

$$a_{1,2}b_{3,4} + a_{3,4}b_{1,2} - a_{1,3}b_{2,4} - a_{2,4}b_{1,3} + a_{1,4}b_{2,3} + a_{2,3}b_{1,4} = 0. \tag{3.14}$$

Because $\dim K^1_{2,2} = 8$ the intersection must be complete. In particular the degree of $K^1_{2,2}$ is equal to 8 by the classical Bézout theorem.

In the remaining part of this section we want to give a system theoretic interpretation of the boundary points that were added in the compactification $K^n_{p,m}$. For this consider a polynomial vector $f(s) \in K^n_{p,m}$. We now distinguish two cases.

Case 1. Assume $f(s) \neq 0$ for all $s \in \bar{\mathbb{K}}$. From the proof of Theorem 3.6 it immediately follows that we find a $p \times (m+p)$ polynomial matrix $P(s)$ that is mapped onto $f(s)$ under the Plücker embedding. Because $P(s)$ has full rank for all $s \in \bar{\mathbb{K}}$, it follows that the Kronecker row indices are equal to the minimal row indices in the sense of Forney [6]. (Compare [15].) In other words, if $\tilde{P}(s)$ is another polynomial matrix that is mapped onto $f(s)$, then $P(s)$ and $\tilde{P}(s)$ are row equivalent, i.e., there is a unimodular matrix $U(s)$ with $\tilde{P}(s) = U(s) \cdot P(s)$.

Case 2. There is an $s_o \in \bar{\mathbb{K}}$ with $f(s_o) = 0$. Because the minimal polynomial of s_o over $\mathbb{K}$ divides each coordinate, we find a polynomial $g(s) \in \mathbb{K}[s]$ with $f(s) = g(s)\tilde{f}(s)$ and $\tilde{f}(s) \neq 0$ for all $s \in \bar{\mathbb{K}}$. It is obvious that we again find a polynomial $p \times (m+p)$ matrix $P(s)$ that is mapped onto $f(s)$. Note that $P(s_o)$ does not have full rank. To describe all other polynomial matrices that are mapped onto $f(s)$ we introduce the following group:

$$H := \{A \in Gl_p(\mathbb{K}(s)) \mid \det A \in \mathbb{K}\backslash\{0\}\}. \tag{3.15}$$

Clearly the unimodular group is a subgroup of H consisting of all elements in H that have polynomial entries. This group enables us to introduce the following equivalence relation.

DEFINITION 3.8. Two polynomial matrices $P(s)$ and $\tilde{P}(s)$ are called H-equivalent if there is an element $U \in H$ with $\tilde{P}(s) = U \cdot P(s)$.

Note that row equivalent matrices are always H-equivalent. Moreover, if $P(s)$ has full row rank for all $s \in \bar{\mathbb{K}}$, it then follows from the proof of Theorem 3.6 that $P(s)$ and $\tilde{P}(s)$ are row equivalent if and only if they are H-equivalent. In other words, the concept of row equivalence and H-equivalence are the same for the generic set. The following example illustrates the difference of the two concepts.

Example 3.9. The following two matrices have the same Plücker coordinates and are therefore H–equivalent:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2s & 3s \end{pmatrix}, \qquad B = \begin{pmatrix} s & 0 & 0 \\ 0 & 2 & 3 \end{pmatrix}. \tag{3.16}$$

On the other hand, it is immediate that the matrices A, B are not row equivalent; that is, there is no unimodular matrix $U(s)$ with $B = UA$.

From the above it is now clear that every point of $K^n_{p,m}$ can be viewed as an H-equivalence class of $p \times (m+p)$ polynomial matrices and every H-equivalence class consists of one (generically) or several autoregressive systems. At this point we want to mention that $K^n_{p,m}$ has singularities and those singularities occur at points where several autoregressive systems form one H-equivalence class. As shown by Ravi and Rosenthal [26] the set of all "homogeneous autoregressive systems" of degree n constitutes a desingularisation of $K^n_{p,m}$ and we refer to [26] for details.

We summarize this section with the following theorem.

THEOREM 3.10. *$K^n_{p,m}$ consists of all H-equivalence classes of autoregressive systems of size $p \times (m+p)$ and degree less than or equal to n.*

4. Dynamic feedback and q-nondegeneracy. In the last section we introduced a compactification (denoted by $K^n_{p,m}$) of the space of proper transfer functions $S^n_{p,m}$. In this section we will show that the pole placement problem with dynamic compensators can be studied as an intersection problem in the variety $K^n_{p,m}$.

For this consider a proper transfer function $G(s) \in S^n_{p,m}$ describing the behavior between an input $\hat{u}$ and an output $\hat{y}$ in the frequency domain:

$$\hat{y} = G(s)\hat{u}. \tag{4.1}$$

The feedback compensators that we will consider are proper transfer functions $F(s) \in S^q_{m,p}$. The plant and the compensator are combined through the feedback law:

$$\hat{u} = F(s)\hat{y} + \hat{v}. \tag{4.2}$$

If the characteristic matrix $(I - G(s)F(s))$ is invertible (this is always the case if $G(s)$ is strictly proper) it is well known that the transfer function between the new input $\hat{v}$ and the output $\hat{y}$ is well defined and given by

$$G_F(s) := (I - G(s)F(s))^{-1}G(s). \tag{4.3}$$

The stability of equilibria or periodic motions of the closed-loop system depends on the position of the poles of $G_F(s)$. To describe the poles of the closed-loop transfer function, we introduce a left coprime factorization of $G(s)$ and a right coprime factorization of $F(s)$:

$$G(s) = D^{-1}_{LG}(s)N_{LG}(s), \qquad F(s) = N_{RF}(s)D^{-1}_{RF}(s). \tag{4.4}$$

A straightforward calculation results in the following form for the closed-loop transfer function:

$$G_F(s) = D_{RF}(s)(D_{LG}(s)D_{RF}(s) - N_{LG}(s)N_{RF}(s))^{-1}N_{LG}(s). \tag{4.5}$$

Note that every pole of $G_F(s)$ is a zero of the polynomial

$$\phi(s) = \det(D_{LG}(s)D_{RF}(s) - N_{LG}(s)N_{RF}(s)) \tag{4.6}$$

and every zero of $\phi(s)$ is a pole of $G_F(s)$ if no pole-zero cancellation occurred. Moreover, if $G(s)$ is a strictly proper system of McMillan degree n and $F(s)$ is a proper compensator of McMillan degree q, then $\phi(s)$ is a polynomial of degree $n+q$. Identifying the vector space $\mathbb{K}^{n+q}$ with all monic polynomials of degree $n+q$ we define the pole placement map for a strictly proper system $G(s)$ by:

$$\rho_G : S^q_{m,p} \longrightarrow \mathbb{K}^{n+q}, \qquad F(s) \longmapsto \phi(s)_{\text{monic}}. \tag{4.7}$$

This definition is in many ways unsatisfactory if $G(s)$ is proper. Indeed, if $G(s)$ is proper it is possible that $\phi(s)$ is not of degree $n+q$ anymore, in particular if $(I - G(s)F(s))$ is not invertible $\phi(s) \equiv 0$.

To extend the definition of the pole placement map to proper systems we first introduce the following set, which Ghosh [8] called the *base locus*:

$$B_G := \{F(s) \in S^q_{m,p} \mid \det(I - G(s)F(s)) \equiv 0\}. \tag{4.8}$$

To avoid difficulties with low-degree polynomials, we identify the space of polynomials with the projective space $\mathbb{P}^{n+q}$ and use the following definition.

DEFINITION 4.1. The pole placement map for a proper transfer function $G(s)$ is given by

$$\rho_G : S^q_{m,p} - B_G \longrightarrow \mathbb{P}^{n+q}, \qquad F(s) \longmapsto \phi(s). \tag{4.9}$$

It is, of course, an important problem in multivariable linear control theory: under which condition is ρ_G onto or at least almost onto? In particular, it would be of great interest to know the minimum order q of a compensator that pole assigns

or stabilizes a given generic system of order n. Using a dimension argument, we immediately obtain the following necessary condition for ρ_G to be onto:

$$q(m+p) + mp \geq n+q. \tag{4.10}$$

One of our main goals in this paper is to show that this condition is also sufficient when the field is algebraically closed and the plant $G(s)$ is generic. To achieve this result, we first give a new description of the polynomial $\phi(s)$ and this will enable us to reformulate the problem geometrically.

If $F(s) = D_{LF}^{-1}(s)N_{LF}(s) = N_{RF}(s)D_{RF}^{-1}(s)$ are a left and a right coprime factorization of $F(s)$ it is obvious that

$$(N_{LF}(s)\, D_{LF}(s)) \begin{pmatrix} D_{RF}(s) \\ -N_{RF}(s) \end{pmatrix} \equiv 0_{m\times p}. \tag{4.11}$$

In some sense we can view the matrix

$$\begin{pmatrix} D_{RF}(s) \\ -N_{RF}(s) \end{pmatrix}$$

as the dual curve of the Hermann–Martin curve $(N_{LF}(s)\, D_{LF}(s))$ of $F(s)$. The following lemma, which is well known if the compensator is static [2], is now easy to verify and the proof will be omitted.

LEMMA 4.2. *For a particular point $s_i \in \bar{\mathbb{K}}$ the following conditions are equivalent:*

$$\det\left((D_{LG}(s_i)\, N_{LG}(s_i)) \begin{pmatrix} D_{RF}(s_i) \\ -N_{RF}(s_i) \end{pmatrix}\right) = 0, \tag{4.12}$$

$$\det \begin{pmatrix} D_{LG}(s_i) & N_{LG}(s_i) \\ N_{LF}(s_i) & D_{LF}(s_i) \end{pmatrix} = 0, \tag{4.13}$$

$$\operatorname{rowsp}(D_{LG}(s_i)\, N_{LG}(s_i)) \cap \operatorname{rowsp}(N_{LF}(s_i)D_{LF}(s_i)) \neq \{0\}. \tag{4.14}$$

Note that two polynomials with the same roots are multiples of each other. In other words the following corollary holds.

COROLLARY 4.3.

$$\phi(s) = \det(D_{LG}(s)D_{RF}(s) - N_{LG}(s)N_{RF}(s)) = \mathrm{c}\cdot \det \begin{pmatrix} D_{LG}(s) & N_{LG}(s) \\ N_{LF}(s) & D_{LF}(s) \end{pmatrix} \tag{4.15}$$

The $(m+p)\times(m+p)$ matrix appearing in this equation has many nice properties. On one side the equation

$$\begin{pmatrix} D_{LG}(s) & N_{LG}(s) \\ N_{LF}(s) & D_{LF}(s) \end{pmatrix} \cdot \begin{pmatrix} y \\ -u \end{pmatrix}(s) = 0 \tag{4.16}$$

gives a combined description of the plant and the compensator equations by means of autoregressive equations. This point of view can be found, e.g., in [31], [39].

Geometrically, $(D_{LG}(s)N_{LG}(s))$ defines a rational curve $\zeta \epsilon \mathrm{Rat}_n(\mathbb{P}^1, \mathrm{Grass}(p, m+p))$ and $(N_{LF}(s)D_{LF}(s))$ defines a rational curve $\psi \epsilon \mathrm{Rat}_q(\mathbb{P}^1, \mathrm{Grass}(m, m+p))$. Using the Plücker embedding (2.6) we can represent ζ by

$$g(s) := g_1(s) \wedge \cdots \wedge g_p(s), \tag{4.17}$$

where again $g_i(s)$ denotes the ith row of $(D_{LG}(s)N_{LG}(s))$. Similarly, ψ has a representation

$$f(s) := f_1(s) \wedge \cdots \wedge f_m(s). \tag{4.18}$$

Finally the poles of the closed-loop system are the zeros of the polynomial

$$\tilde{\phi}(s) := g_1(s) \wedge \cdots \wedge g_p(s) \wedge f_1(s) \wedge \cdots \wedge f_m(s). \tag{4.19}$$

Note that $\tilde{\phi}(s)$ is, of course, a multiple of the polynomial $\phi(s)$. In addition the wedge product $g(s) \wedge f(s)$ defines a bilinear pairing $\langle\ ,\ \rangle$ that extends linearly to the product space $\mathbb{P}(\mathbb{K}^{n+1} \otimes \wedge^p \mathbb{K}^{m+p}) \times \mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p})$.

We are now in a position to formulate the pole placement problem with dynamic compensators in a geometric language.

Geometric problem. Given a rational curve $\zeta \epsilon \mathrm{Rat}_n(\mathbb{P}^1, \mathrm{Grass}(p, m+p))$ and a divisor $P = \{s_1, \ldots, s_{n+q}\}$. Is there a curve $\psi \epsilon \mathrm{Rat}_q(\mathbb{P}^1, \mathrm{Grass}(m, m+p))$ such that $\psi(s_i) \cap \zeta(s_i) \neq \{0\}$ for all $s_i \in P$? What is the minimal degree q needed?

Remark 4.4. Not all geometric solutions enable us to construct a proper compensator although it is always possible to represent such a solution by an autoregressive system. In addition, we want to find solutions that are admissible (compare with [31]). In geometric terms, we want to exclude a Hermann–Martin curve $\psi(s)$ with the property that $\psi(s) \cap \zeta(s) \neq \{0\}$ for all $s \in \mathbb{K}$.

To handle these difficulties we make the following definition.

DEFINITION 4.5. A rational curve $\zeta \epsilon \mathrm{Rat}_n(\mathbb{P}^1, \mathrm{Grass}(p, m+p))$ is called q-degenerate if there is a rational curve $\psi \epsilon \mathrm{Rat}_i(\mathbb{P}^1, \mathrm{Grass}(m, m+p))$ with $i \leq q$ and $\psi(s) \cap \zeta(s) \neq \{0\}$ for all $s \in \mathbb{K}$. A curve that is not q-degenerate is called q-nondegenerate. A system $G(s)$ is called q-(non)degenerate if the corresponding Hermann–Martin curve is q-(non)degenerate.

Note that our definition is a natural generalization of the concept of the degenerate system as introduced in [2], and this concept itself generalizes the concept of a degenerate curve in projective space. In a concrete example we can use the equivalent formulations in Lemma 4.2 to decide if a particular plant $G(s)$ is q-degenerate.

From the definition it now follows immediately that the pole placement map ρ_G introduced in (4.7) and (4.9) can be extended in a continuous manner to a morphism $\bar{\rho_G}$ defined on the whole compactification $K^q_{m,p}$ if the system $G(s)$ is q-nondegenerate. In other words, all autoregressive systems $P(s) \in K^q_{m,p}$ are admissible and the base locus set B_G introduced in (4.8) is empty:

$$\begin{array}{rcc} \bar{\rho_G}\ : & K^q_{m,p} & \longrightarrow \mathbb{P}^{n+q} \\ & \uparrow & \uparrow \\ \rho_G\ : & S^q_{m,p} & \longrightarrow \mathbb{K}^{n+q} \end{array} \tag{4.20}$$

The concept of q-degeneracy will be of crucial importance in the next section. The following example will illustrate the concept of q–degeneracy on a 3-input, 1-output system.

Example 4.6.

1. $G(s) = (1/s^5, 1/s^3, 1/s)$ defines a system of order 5, which is 1-degenerate. Indeed, use the first condition in Lemma 4.2 to construct a covector which will make the inner product $\langle\ ,\ \rangle$ identically zero:

$$\langle\ (1, s^2, s^4, s^5)\ ,\ (0, 0, s, -1)\ \rangle \equiv 0 \tag{4.21}$$

2. $\tilde{G}(s) = (1/s^6, 1/s^4, 1/s^2)$ defines a system of order 6 which is 1-nondegenerate because

$$\langle (1, s^2, s^4, s^6), (a_1 + b_1 s, a_2 + b_2 s, a_3 + b_3 s, a_4 + b_4 s) \rangle \equiv 0 \tag{4.22}$$

implies $a_i = b_i = 0, \quad i = 1, \ldots, 4$. Actually, we will show in the next section that the generic 1-input, 3-output system of order 6 is 1-nondegenerate.

5. On the minimal order dynamic compensator. In this section we will assume that the ground field $\mathbb{K}$ is algebraically closed. The following theorem, called the projective dimension theorem, will be used several times in this section. Our formulation can be found in Hartshorne [9], where a proof is also given.

THEOREM 5.1. *Let Y, Z be varieties of dimension r, s in $\mathbb{P}^N$. Then every irreducible component of $Y \cap Z$ has dimension $\geq r+s-N$. Furthermore, if $r+s-N \geq 0$, then $Y \cap Z$ is nonempty.*

The next theorem that we present is a strong version of the classical Bézout theorem, which we will need to prove Theorem 5.7. The theorem was originally formulated and proven by Weil [36]. The crucial part for the formulation of the theorem was the "right" definition of the intersection multiplicity i. For a broader discussion of this theorem and its generalizations we refer the reader to Vogel [33]. The following theorem is a reformulation of [33, Prop. 3.26].

THEOREM 5.2. *Let Y, Z be varieties of dimension r, s in $\mathbb{P}^N$. Assume the intersection $Y \cap Z$ is proper, i.e., $\dim(Y \cap Z) = r + s - N$. Denote with Ω the set of irreducible components of $Y \cap Z$ and with $i(Y, Z; C)$ the intersection multiplicity of Y and Z along C. Then we have*

$$\deg Y \cdot \deg Z = \sum_{C \epsilon \Omega} i(Y, Z; C) \cdot \deg C. \tag{5.1}$$

Another important concept in all that follows is the notion of a central projection. Assume E, H are linear subspaces of dimension $r, N - r - 1$ and $E \cap H = \emptyset$. In this case we can define the following map, which is well defined by basic facts of linear algebra:

$$\pi : \mathbb{P}^N - E \longrightarrow H, \qquad x \longmapsto \operatorname{span}(x, E) \cap H. \tag{5.2}$$

π is called a central projection onto H with center E. As shown by Wang [34], the pole placement map with static compensators is a central projection. As we will show, the same is true in the dynamic case.

Our first goal is a characterization of the q-nondegenerate systems.

LEMMA 5.3. *The set of q-degenerate systems is algebraic in the quasi-projective variety $S^n_{p,m}$ of proper systems with McMillan degree n.*

Proof. Consider in $S^n_{p,m} \times K^q_{m,p}$ the coincidence set

$$S := \{(N_{LG}(s)\, D_{LG}(s)), (N_{RF}(s)\, D_{RF}(s)) \mid \det(D_{LG}(s) D_{RF}(s) - N_{LG}(s) N_{RF}(s)) \equiv 0\}, \tag{5.3}$$

which defines an algebraic set in the product. Because $K^q_{m,p}$ is projective, the projection on the first factor is still an algebraic set by the main theorem of elimination theory (see, e.g., [24]). □

The next lemma shows that every system is q-degenerate for some large natural number $q \in \mathbb{N}$.

LEMMA 5.4. *If $q(m+p)+mp > n+q$, every $p \times m$ system of order n is q-degenerate.*

Proof. Assume $g(s)$ are the Plücker coordinates of a plant $G(s)$ with McMillan degree n. Consider in $\mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p})$ the set

$$E_G := \{f(s) \mid \langle g(s), f(s) \rangle \equiv 0\}. \tag{5.4}$$

E_G defines a plane of codimension at most $q(m+p)+mp$, the dimension of the variety $K^q_{m,p}$. The plane E_G intersects $K^q_{m,p}$ by the projective dimension theorem. □

So far it has only been shown (Lemma 5.3) that the set of q-nondegenerate systems form a Zariski-open (possibly empty) set in $S^n_{p,m}$. Using the following theorem we will be able to show that this Zariski-open set is nonempty in $S^n_{p,m}$ if q is small enough.

THEOREM 5.5. *The dimension of the coincidence set $S \subset S^n_{p,m} \times K^q_{m,p}$ introduced in* (5.3) *is given by*

$$\dim S = \dim S^n_{p,m} + \dim K^q_{m,p} - n - q - 1. \tag{5.5}$$

Proof. Consider an element of S given by

$$\det \begin{pmatrix} D_{LG}(s) & N_{LG}(s) \\ N_{LF}(s) & D_{LF}(s) \end{pmatrix} \equiv 0. \tag{5.6}$$

Without loss of generality we assume that the system $(D_{LG}(s)N_{LG}(s))$ and the compensator $(N_{LF}(s)D_{LF}(s))$ are both row reduced with minimal indices $\nu_1 \geq \cdots \geq \nu_p$, respectively, $\mu_1 \geq \cdots \geq \mu_m$ and $\nu_1 \geq \mu_1$. In particular we have $n = \sum_{i=1}^p \nu_i$ and $q = \sum_{j=1}^m \mu_j$. As explained in [30] we have a free action on $(N_{LF}(s)D_{LF}(s))$ with an algebraic group of dimension at least m^2. This group is characterized as the subgroup of the unimodular group $Gl_m(\mathbb{K}[s])$ which leaves the row indices $\mu_1, \ldots, \mu_m$ invariant. Similarly there is a free action on $(D_{LG}(s)N_{LG}(s))$ with an algebraic group of dimension at least p^2.

Denote with S_1 the parameter space of all $(m+p-1) \times (m+p)$ polynomial matrices having row indices $\nu_2, \ldots, \nu_p, \mu_1, \ldots, \mu_m$. S_1 is a vector space of dimension

$$\dim S_1 = (\sum_{i=2}^{p} \nu_i + \sum_{j=1}^{m} \mu_j + m + p - 1)(m+p). \tag{5.7}$$

Assume now that the last $m+p-1$ rows form a minimal basis in the sense of Forney [6] of the $\mathbb{K}(s)$-vector space, which these rows generate. Equivalently, the greatest common divisor of the full size $(m+p-1) \times (m+p-1)$ minors is 1. In the following we restrict the dimension calculation to this Zariski-open subset of S_1 because it is not difficult to show that the other cases lead to lower-dimensional subsets. From (5.6) it then follows that the first row is a linear combination of the last $m+p-1$ rows

$$g_1(s) = \sum_{i=2}^{p} x_i(s)g_i(s) + \sum_{j=1}^{m} y_j(s)h_j(s), \qquad x_i(s), y_j(s) \in \mathbb{K}(s). \tag{5.8}$$

From the main theorem in Forney [6] it follows that $x_i(s), y_j(s)$ are even elements of $\mathbb{K}[s]$. Moreover $\deg x_i(s) \leq \nu_1 - \nu_i$ and $\deg y_j(s) \leq \nu_1 - \mu_j$.

Denote with S_2 all polynomial vectors $g_1(s)$ of degree ν_1 that are in the rowspace of a given set of vectors $\{g_2(s), \ldots, g_p(s), \ldots h_m(s)\}$. From the above follows that

$$\dim S_2 = \sum_{i=2}^{p}(\nu_1 - \nu_i + 1) + \sum_{j=1}^{m}(\nu_1 - \mu_j + 1) = (m+p)\nu_1 - n - q + m + p - 1. \tag{5.9}$$

Finally, taking into consideration the free action of the above-mentioned groups, we obtain

$$\dim S \leq \dim S_1 + \dim S_2 - m^2 - p^2 \tag{5.10}$$

$$= n(m+p) + mp + q(m+p) + mp - n - q - 1 \tag{5.11}$$

$$= \dim S^n_{p,m} + \dim K^q_{m,p} - n - q - 1. \tag{5.12}$$

Finally, (5.6) imposes at most $n+q+1$ algebraic conditions because the characteristic equation is a polynomial of degree at most $n+q$. The inequality in (5.10) is therefore an equality. □

COROLLARY 5.6. *If $q(m+p) + mp \leq n+q$, the generic $p \times m$ proper system of order n is q-nondegenerate.*

Proof. Because $\dim K^q_{m,p} = q(m+p) + mp$ it follows from (5.5) that $\dim S \leq \dim S^n_{p,m} - 1$. In particular the projection of S onto $S^n_{p,m}$ is a proper algebraic subset in $S^n_{p,m}$. □

The previous corollary was proven for $q = 0$ (static feedback) by Brockett and Byrnes [2], from which it then followed that the pole placement map with static compensators is generically onto if $mp = n$. In the following we will extend this result to the dynamic case. The proof that we present combines ideas from a proof given by Rosenthal in [27] and a proof given by Wang in [34] for the case of static feedback.

THEOREM 5.7. *If a system $G(s)$ is q-nondegenerate and $q(m+p)+mp = n+q$, then the pole placement map*

$$\bar{\rho_G} : K^q_{m,p} \longrightarrow \mathbb{P}^{n+q} \tag{5.13}$$

is onto of degree $d_{m,p,q}$, where $d_{m,p,q}$ is the degree of the variety $K^q_{m,p}$.

Proof. Consider in $\mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p})$ again the linear subspace

$$E_G = \{f(s) \mid \langle g(s), f(s) \rangle \equiv 0\}. \tag{5.14}$$

Because $G(s)$ is q-nondegenerate it follows that $E_G \cap K^q_{m,p} = \emptyset$ and the codimension of E_G is equal to $q(m+p) + mp + 1$. The linear pairing $\langle , \rangle$ induces a linear map

$$\begin{aligned} L \; : \; \mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p}) - E_G &\longrightarrow \mathbb{P}^{n+q} \\ f(s) &\longmapsto \langle g(s), f(s) \rangle, \end{aligned} \tag{5.15}$$

which has to be onto by a linear argument. Note that $\bar{\rho_G} = L \mid_{K^q_{m,p}}$. Denote with H any linear subspace of $\mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p})$ for which $\dim H = n+p$ and $L(H) = \mathbb{P}^{n+q}$. We have a central projection

$$\pi \; : \; \mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p}) - E_G \longrightarrow H. \tag{5.16}$$

If $y \in H$ is a particular point, it follows by linear equation theory that the whole fiber $\pi^{-1}(y)$ (which is a linear plane in $\mathbb{P}(\mathbb{K}^{q+1} \otimes \wedge^m \mathbb{K}^{m+p})$) is mapped under L onto $L(y)$. In other words, we have $L = L \circ \pi$ and $\bar{\rho_G}$ is onto if and only if $\pi \mid_{K^q_{m,p}}$ is onto.

By the projective dimension theorem $\pi^{-1}(y) \cap K^q_{m,p} \neq \emptyset$. Finally every fiber $\pi^{-1}(y)$ intersects $K^q_{m,p}$ properly [32, p. 48]. By Theorem 5.2, $\pi^{-1}(y) \cap K^q_{m,p}$ consists in this case of exactly $d_{m,p,q}$ points when counted with multiplicities. □

If the system $G(s)$ is strictly proper and the compensator $F(s)$ is admissible and proper, it follows from Corollary 4.3 that the closed-loop characteristic polynomial $\phi(s)$ has degree exactly equal to $n+q$, the sum of the McMillan degrees of $G(s)$ and $F(s)$. In other words the "infinite points," that is, the points in the set $K^q_{m,p} - S^q_{m,p}$, are mapped onto the closed-loop characteristic polynomials of degree strictly less than $n+q$. We therefore obtain the following corollary.

COROLLARY 5.8. *If $G(s)$ is q-nondegenerate and strictly proper and $q(m+p)+mp = n+q$, then the pole placement map $\rho_G : S^q_{m,p} \longrightarrow \mathbb{K}^{n+q}$ introduced in* (4.7) *is onto. Moreover, if counted with multiplicities there are exactly $d_{m,p,q}$ different compensators $F(s)$ assigning a specific closed-loop characteristic polynomial.*

The degree of the variety $K^q_{m,p}$ is therefore equal to the number of compensators that will place the poles of the closed-loop system at a desired location. In particular, if $G(s)$ is a real plant and the number $d_{m,p,q}$ would turn out to be odd for certain m,p,q, we would be able to predict the existence of a real compensator because the solution set must be invariant under complex conjugation. In the case of static feedback, i.e., $q=0$, we have $K^0_{m,p} = \mathrm{Grass}(m, m+p)$ and it is well known when the degree of the Grassmann variety is odd. (Compare, e.g., [3].) As shown in [28] the degree of $K^1_{2,3}$ is equal to 55 and so such (nontrivial) cases also exist if $q > 0$. The following corollary explains the proper situation.

COROLLARY 5.9. *If $G(s)$ is q-nondegenerate and proper and $q(m+p)+mp = n+q$, then the pole placement map $\rho_G : S^q_{m,p} \longrightarrow \mathbb{P}^{n+q}$ introduced in* (4.9) *is almost onto.*

Proof. Because $G(s)$ is q-nondegenerate, the lifted map $\bar{\rho_G} : K^q_{m,p} \to \mathbb{P}^{n+q}$ exists and is onto by Theorem 5.7. The difference set $K^q_{m,p} - S^q_{m,p}$ has dimension strictly less than $n+q$. Because $\bar{\rho_G}(K^q_{m,p}) = \mathbb{P}^{n+q}$ and $\bar{\rho_G}\,|_{S^q_{m,p}} = \rho_G$ the statement follows. □

Remark 5.10. From the proof it follows in particular that those closed-loop characteristic polynomials that cannot be achieved with a proper compensator can always be achieved with a general autoregressive compensator. (Compare with Remark 4.4.)

So far we have provided only positive results, that is, results when the dimension of the domain and the range are equal. The following theorem explains the situation when the dimension of the domain is larger than the dimension of the range.

THEOREM 5.11. *If $G(s) \in S^n_{p,m}$ is a generic plant and if*

$$q(m+p)+mp \geq n+q, \tag{5.17}$$

then the pole placement map

$$\rho_G : S^q_{m,p} - B_G \longrightarrow \mathbb{P}^{n+q} \tag{5.18}$$

introduced in Definition 4.1 *is almost onto. Moreover, the extended map*

$$\bar{\rho_G} : K^q_{m,p} - E_G \longrightarrow \mathbb{P}^{n+q} \tag{5.19}$$

is onto.

Proof. Consider again the coincidence set $S \subset S^n_{p,m} \times K^q_{m,p}$ introduced in the proof of Lemma 5.3. Denote with $pr : S \to S^n_{p,m}$ the projection onto the first factor. From

Theorem 5.5 it then follows that for a generic element $G(s) \in S^n_{p,m}$ the dimension of the fiber $pr^{-1}(G(s))$ is bounded by

$$\dim(pr^{-1}(G(s))) \leq \dim K^q_{m,p} - n - q - 1. \tag{5.20}$$

In particular, using earlier notation we have

$$\dim(E_G \cap K^q_{m,p}) \leq q(m+p) + mp - n - q - 1. \tag{5.21}$$

Following the proof of Theorem 5.7 and using again the projective dimension theorem it follows that

$$\dim(\pi^{-1}(y) \cap K^q_{m,p}) > q(m+p) + mp - n - q - 1. \tag{5.22}$$

From above two inequalities it now follows in particular that for every closed-loop polynomial $p(s)$ there is an admissible autoregressive system $F(s) \in K^q_{m,p} - E_G$ with $\bar{\rho_G}(F(s)) = p(s)$. The map $\bar{\rho_G}$ is therefore onto. Finally because the fibers of $\bar{\rho_G}$ in $K^q_{m,p} - E_G$ have dimension at least $q(m+p) + mp - n - q$ and the dimension of the range of $\bar{\rho_G}$ is $n+q$, the map $\rho_G : S^q_{m,p} - B_G \to \mathbb{P}^{n+q}$ is almost onto by a dimension argument. □

Acknowledgments. The author thanks U. Helmke, M. S. Ravi, and X. Wang for helpful conversations and suggestions.

REFERENCES

[1] F. M. Brash and J. B. Pearson, *Pole placement using dynamic compensators*, IEEE Trans. Automat. Control, AC-15, (1970), pp. 34–43.

[2] R. W. Brockett and C. I. Byrnes, *Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, AC-26, (1981), pp. 271–284.

[3] C. I. Byrnes, *Pole assignment by output feedback*, Lecture Notes in Control and Information Sci., 135, Springer-Verlag, Berlin, Heidelberg, New York, 1989, pp. 31–78.

[4] ———, *On compactifications of spaces of systems and dynamic compensation*, in Proc. IEEE Conference on Decision and Control, San Antonio, TX, 1983, pp. 889–894.

[5] D. F. Delchamps, *State Space and Input–Output Linear Systems*, Springer-Verlag, New York, 1988.

[6] G. D. Forney, *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control Optim., 13 (1975), pp. 493–520.

[7] H. Glüsing-Lüerssen, *Gruppenaktionen in der Theorie Singulärer Systeme,* Ph.D. thesis, Universität Bremen, Bremen, Germany, 1991.

[8] B. K. Ghosh, *An approach to simultaneous system design. Part* II: *Nonswitching gain and dynamic feedback compensation by algebraic geometric methods*, SIAM J. Control Optim., 26 (1988), pp. 919–963.

[9] R. Hartshorne, *Algebraic Geometry,* Springer-Verlag, Berlin, 1977.

[10] M. Hazewinkel, *Moduli and canonical forms for linear dynamical systems* III: *The algebraic geometric case*, in Proc. of the 1976 Ames Research Center (NASA) Conference on Geometric Control Theory, C. F. Martin and R. Hermann, eds., Math.-Sci. Press, Brookline, MA, 1977, pp. 291–336.

[11] ———, *On families of linear systems: Degeneration phenomena*, Algebraic and Geometric Methods in Linear Systems Theory, C. I. Byrnes and C. F. Martin, eds., American Mathematical Society, Providence, RI, 1980, pp. 157–189.

[12] U. Helmke, *A compactification of the space of rational transfer functions by singular systems*, J. Math. Systems Estim. Control, to appear.

[13] R. Hermann and C. F. Martin, *Applications of algebraic geometry to system theory part* I, IEEE Trans. Automat. Control, AC-22, (1977), pp. 19–25.

[14] P. Janssen, *General results on the McMillan degree and the Kronecker indices of* ARMA *and* MFD *models*, Internat. J. Control, 48 (1988), pp. 591–608.

[15] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[16] H. KIMURA, *Pole assignment by gain output feedback*, IEEE Trans. Automat. Control, 20 (1975), pp. 509–516.
[17] S. L. KLEIMAN AND D. LAKSOV, *Schubert Calculus*, Amer. Math. Monthly, 79 (1972), pp. 1061–1082.
[18] M. KUIJPER AND J. M. SCHUMACHER, *Realization of autoregressive equations in pencil and descriptor form*, SIAM J. Control Optim., 28 (1990), pp. 1162–1189.
[19] ———, *Realization and partial fractions*, Linear Algebra Appl., 169 (1992), pp. 195–222.
[20] V. G. LOMADZE, *Finite-Dimensional Time-Invariant Linear Dynamical Systems: Algebraic Theory*, Acta Appl. Math., 19 (1990), pp. 149–201.
[21] F. S. MACAULAY, *Some formulae in elimination*, Proc. London Math. Soc., 1903, pp. 3–27.
[22] B. M. MANN AND R. J. MILGRAM, *Some spaces of holomorphic maps to complex Grassmann manifolds*, J. Differential Geom., 33 (1991), pp. 301–324.
[23] C. F. MARTIN AND R. HERMANN, *Applications of algebraic geometry to system theory: The McMillan degree and Kronecker indices as topological and holomorphic invariants*, SIAM J. Control, 16 (1978), pp. 743–755.
[24] D. MUMFORD, *Algebraic Geometry I: Complex Projective Varieties*, Springer-Verlag, Berlin, New York, 1976.
[25] C. PROCESI, *A Primer of Invariant Theory*, Notes by G.Boffi, Brandeis Lecture Notes, Brandeis University, Waltham, MA, 1982.
[26] M. S. RAVI AND J. ROSENTHAL, *A smooth compactification of the space of transfer functions with fixed McMillan degree*, preprint, 1992.
[27] J. ROSENTHAL, *Geometric Methods for Feedback Stabilization of Multivariable Linear Systems*, Department of Mathematics, Ph.D. thesis, Arizona State University, 1990.
[28] ———, *On minimal order dynamical compensators of low order systems*, in Proc. European Control Conference, Grenoble, France, 1991, pp. 374–378.
[29] ———, *A compactification of the space of multivariable linear systems using geometric invariant theory*, J. Math. Systems, Estim. Control, 2 (1992), pp. 111–121.
[30] J. ROSENTHAL, M. SAIN, AND X. WANG, *Topological considerations for autoregressive systems with fixed Kronecker indices*, Circuits Systems Signal Process., to appear.
[31] J. M. SCHUMACHER, *A pointwise criterion for controller robustness*, Systems Control Lett., 18 (1992), pp. 1–8.
[32] I. R. SHAFAREVICH, *Basic Algebraic Geometry*, Springer-Verlag, Berlin, New York, 1974.
[33] W. VOGEL, *Results on Bézout's Theorem*, Tata Institute of Fundamental Research, Springer-Verlag, Berlin, New York, 1981.
[34] X. WANG, *On output feedback via Grassmannians*, SIAM J. Control Optim., 29 (1991), pp. 926–935.
[35] ———, *Pole placement by static output feedback*, J. Math. Systems Estim. Control, 2 (1992), pp. 205–218.
[36] A. WEIL, *Foundations of Algebraic Geometry*, Amer. Math. Soc. Coll. Publ., Vol. XXIX, Providence, R.I., 1946.
[37] J. C. WILLEMS AND W. H. HESSELINK, *Generic properties of the pole placement problem*, in Proc. of the IFAC, Helsinki, Finland, 1978.
[38] J. C. WILLEMS, *Input-output and state-space representations of finite-dimensional linear time-invariant systems*, Linear Algebra Appl., 50 (1983), pp. 581–608.
[39] ———, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 1 pp. 297–310, January 1994

017

FINITE-DIMENSIONAL FILTERS WITH NONLINEAR DRIFT II: BROCKETT'S PROBLEM ON CLASSIFICATION OF FINITE-DIMENSIONAL ESTIMATION ALGEBRAS*

WEN-LIN CHIOU† AND STEPHEN S.-T. YAU‡

Abstract. The idea of using estimation algebras to construct finite-dimensional nonlinear filters was first proposed by Brockett and Mitter independently. It turns out that the concept of estimation algebra plays a crucial role in the investigation of finite-dimensional nonlinear filters. In his talk at the International Congress of Mathematics in 1983, Brockett proposed classifying all finite-dimensional estimation algebras. In this paper, all finite-dimensional algebras with maximal rank are classified if the dimension of the state space is less than or equal to two. Therefore, from the Lie algebraic point of view, all finite-dimensional filters are understood generically in the case where the dimension of state space is less than three.

Key words. nonlinear filters, estimation algebra, Wei–Norman approach

AMS subject classifications. 17B30, 35J15, 60G35, 93E11

1. Introduction. In a previous paper [Ya], Yau has studied the general class of nonlinear filtering systems that include both Kalman–Bucy and Benes filtering systems as special cases. Simple algebraic necessary and sufficient conditions were proved for an estimation algebra of such filtering system to be finite-dimensional. Using the Wei–Norman approach, he constructed explicitly finite-dimensional recursive filters for such nonlinear filtering systems. This paper is, in essence, a continuation of [Ya] and we strongly recommend that readers familiarize themselves with the results in [Ya]. However, every effort will be made to make this paper as self-contained as possible without too much duplication of the previous paper.

The idea of using estimation algebras to construct finite-dimensional nonlinear filters was first proposed in Brockett and Clark [Br-Cl], Brockett [Br1], and Mitter [Mi]. The concept of estimation algebras has proved to be an invaluable tool in the study of nonlinear filtering problems. In his famous talk at the International Congress of Mathematics in 1983, Brockett proposed classifying all finite-dimensional estimation algebras. There were some interesting results in 1987 due to Wong [Wo] under the assumptions that the observation $h(x)$ and drift term $f(x)$ are real analytic functions on $\mathbf{R}^n$, and f satisfies the following growth conditions: for any i, all the first-, second-, and third-order partial derivatives of f_i are bounded functions. Under all these conditions, Wong provides partial information toward the classification of finite-dimensional estimation algebra. Namely, he showed that if the estimation algebra is finite-dimensional, then the degree of h in x is at most one, and the estimation algebra has a basis consisting of one second-degree differential operator, L_0 (see (2.1)), first-

*Received by the editors July 8, 1991; accepted for publication (in revised form) October 29, 1992. This research was supported by Army grant DAAL–3-89K-0123.

† Department of Mathematics, Fu Jen University, College of Science and Engineering, Hsinchuang, (24205), Taipei, Taiwan.

‡ Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Box 4348 – M/C 249, Chicago, Illinois 60680.

degree differential operators of the form

$$\sum_{i=1}^{n} \alpha_i \left(\frac{\partial}{\partial x_i} - f_i \right) + \sum_{i=1}^{n} \beta_i \frac{\partial \eta}{\partial x_i},$$

where α_i and β_i are constants and

$$\eta = -\frac{1}{2} \left(\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2 \right),$$

and zero-degree differential operators affine in x. In [T-W-Y], Tam, Wong, and Yau have introduced the concept of an estimation algebra with maximal rank. This is one of the most important general subclass of estimation algebras. Let n be the dimension of the state space. It turns out that all nontrivial finite-dimensional estimation algebras are automatically exact with maximal rank if $n = 1$. It follows from the works of Ocone [Oc], Tam, Wong, and Yau [T-W-Y], and Dong et al. [D-T-W-Y] that the finite-dimensional estimation algebras are completely classified if $n = 1$. In fact, Tam, Wong, and Yau have classified all finite-dimensional exact estimation algebras with maximal rank of arbitrary dimension. In this paper, we classify all finite-dimensional estimation algebras with maximal rank if $n = 2$. The novelty of the problem is that there is no assumption on the drift term of the nonlinear filtering system. The following is our main theorem.

MAIN THEOREM. *Suppose that the state space of the filtering system* (2.0) *below is of dimension two. If E is the finite-dimensional estimation algebra with maximal rank, then the drift term f must be linear vector field plus gradient vector field, and E is a real vector space of dimension 6 with basis given by* 1, x_1, x_2, D_1, D_2, *and* L_0.

This kind of nonlinear filtering systems was studied by Yau [Ya]. Therefore, from the Lie algebraic point of view, we have shown that the finite-dimensional filters considered in [Ya] are the most general finite-dimensional filters.

2. Basic concepts. In this section, we will recall some basic concepts and results from [Ya]. Consider a filtering problem based on the following signal observation model:

$$\begin{aligned} dx(t) &= f(x(t))dt + g(x(t))dv(t), \qquad x(0) = x_0, \\ dy(t) &= h(x(t))dt + dw(t), \qquad y(0) = 0 \end{aligned} \tag{2.0}$$

in which x, v, y, and w are, respectively, $\mathbf{R}^n$, $\mathbf{R}^p$, $\mathbf{R}^m$, and $\mathbf{R}^m$ valued processes, and v and w have components that are independent, standard Brownian processes. We further assume that $n = p$, f, h are C^∞ smooth and that g is an orthogonal matrix. We will refer to $x(t)$ as the state of the system at time t and to $y(t)$ as the observation at time t.

Let $\rho(t, x)$ denote the conditional density of the state given the observation $\{y(s) : 0 \leq s \leq t\}$. It is well known (see [Da-Ma], for example) that $\rho(t, x)$ is given by normalizing a function, $\sigma(t, x)$, which satisfies the following Duncan–Mortensen–Zakai equation:

$$d\sigma(t, x) = L_0\sigma(t, x)dt + \sum_{i=1}^{m} L_i\sigma(t, x)dy_i(t), \qquad \sigma(0, x) = \sigma_0, \tag{2.1}$$

where

$$L_0 = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^{n} f_i\frac{\partial}{\partial x_i} - \sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i} - \frac{1}{2}\sum_{i=1}^{m} h_i^2$$

and for $i = 1, \dots, m$, L_i is the zero-degree differential operator of multiplication by h_i. σ_0 is the probability density of the initial point x_0. In this paper, we will assume σ_0 is a C^∞ function.

Equation (2.1) is a stochastic partial differential equation. In real applications, we are interested in constructing state estimators from observed sample paths with some property of robustness. Davis in [Da] studied this problem and proposed some robust algorithms. In our case, his basic idea reduces to defining a new unnormalized density

$$\xi(t,x) = \exp\Big(-\sum_{i=1}^{m} h_i(x)y_i(t)\Big)\sigma(t,x).$$

It is easy to show that $\xi(t,x)$ satisfies the following time-varying partial differential equation

$$\begin{aligned}\frac{\partial\xi}{\partial t}(t,x) &= L_0\xi(t,x) + \sum_{j=1}^{m} y_i(t)[L_0, L_i]\xi(t,x)\\ &\quad + \frac{1}{2}\sum_{i,j=1}^{m} y_i(t)y_j(t)[[L_0, L_i], L_j]\xi(t,x),\\ \xi(0,x) &= \sigma_0,\end{aligned} \tag{2.2}$$

where $[\cdot\,,\cdot]$ is the Lie bracket defined as follows.

DEFINITION. If X and Y are differential operators, the Lie bracket of X and Y, $[X,Y]$, is defined by $[X,Y]\varphi = X(Y\varphi) - Y(X\varphi)$ for any C^∞ function φ.

Recall that a real vector space $\mathcal{F}$, with an operation $\mathcal{F}\times\mathcal{F}\to\mathcal{F}$ denoted $(x,y)\mapsto[x,y]$ and called the Lie bracket of x and y, is called a Lie algebra if the following axioms are satisfied:

(1) The Lie bracket operation is bilinear;

(2) $[x,y] = 0$ for all $x\in\mathcal{F}$;

(3) $[x,[y,z]] + [y,[z,x]] + [z,[x,y]] = 0 \quad (x,y,z\in\mathcal{F})$.

DEFINITION. The estimation algebra E of a filtering problem (2.0) is defined to be the Lie algebra generated by $\{L_0, L_1, \dots, L_m\}$ or $E = \langle L_0, L_1, \dots, L_m\rangle_{L.A.}$. If, in addition, there exists a potential function φ such that $f_i = \partial\varphi/\partial x_i$ for all $1\le i\le n$, then the estimation algebra is called exact.

In [Ya], the following proposition is proven.

PROPOSITION 1. *$\partial f_j/\partial x_i - \partial f_i/\partial x_j = c_{ij}$ are constants for all i and j if and only if $(f_1,\dots,f_n) = (\ell_1,\dots,\ell_n) + (\partial\varphi/\partial x_1,\dots,\partial\varphi/\partial x_n)$, where $\ell_1,\dots,\ell_n$ are polynomials of degree one and φ is a C^∞ function.*

Define

$$D_i = \frac{\partial}{\partial x_i} - f_i$$

and

$$\eta = \sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2.$$

Then

$$L_0 = \frac{1}{2}\left(\sum_{i=1}^{n} D_i^2 - \eta\right).$$

We need the following basic results for later discussion.

THEOREM 2 (Ocone). *Let E be a finite-dimensional estimation algebra. If a function ξ is in E, then ξ is a polynomial of degree less than or equal to 2.*

Ocone's theorem ([Oc], see [Co] for an extension) says that $h_1, \dots, h_m$ in a finite-dimensional estimation algebra are polynomials of degree less than or equal to 2.

The following theorem proved in [Ya] plays a fundamental role in the classification of finite-dimensional estimation algebra.

THEOREM 3. *Let E be a finite-dimensional estimation algebra of (2.0) satisfying $\partial f_j/\partial x_i - \partial f_i/\partial x_j = c_{ij}$, where c_{ij} are constants for all $1 \le i, j \le n$. Then $h_1, \dots, h_m$ are polynomials of degree at most one.*

In view of the above theorem, we introduce the following definition.

DEFINITION. The estimation algebra E of a filtering problem (2.0) is said to be the estimation algebra with maximal rank if $x_i + c_i$ is in E for all $1 \le i \le n$ where c_i is a constant.

In [Ya], the following theorem was also proved.

THEOREM 4. *Let $F(x_1, \dots, x_n)$ be a polynomial on $\mathbf{R}^n$. Suppose that there exists a polynomial path $c : \mathbf{R} \to \mathbf{R}^n$ such that $\lim_{t\to\infty} \|c(t)\| = \infty$ and $\lim_{t\to\infty} F \circ c(t) = -\infty$. Then there is no C^∞ functions $f_1, f_2, \dots, f_n$ on $\mathbf{R}^n$ satisfying the equation*

$$\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 = F.$$

We recall the following simple lemma proved in [Ya].

LEMMA 5. (i) $[XY, Z] = X[Y, Z] + [X, Z]Y$, *where X, Y and Z are differential operators.*

(ii) $[gD_i, h] = g\ (\partial h/\partial x_i)$, *where $D_i = \partial/\partial x_i - f_i$, g and h are functions defined on $\mathbf{R}^n$.*

(iii) $[gD_i, hD_j] = -gh\omega_{ij} + g(\partial h/\partial x_i)D_j - h(\partial g/\partial x_j)\ D_i$, *where* $\omega_{ji} = [D_i, D_j] = (\partial f_i/\partial x_j) - (\partial f_j/\partial x_i)$.

(iv) $[gD_i^2, h] = 2g(\partial h/\partial x_i)\ D_i + g(\partial^2 h/\partial x_i^2)$.

(v) $[D_i^2, hD_j] = 2(\partial h/\partial x_i)\ D_iD_j - 2h\omega_{ij}D_i + (\partial^2 h/\partial x_i^2)\ D_j - h(\partial\omega_{ij}/\partial x_i)$.

LEMMA 6.

(i) $$[D_i^2, D_j^2] = 4\omega_{ji}D_jD_i + 2(\partial\omega_{ji}/\partial x_j)D_i + (\partial\omega_{ji}/\partial x_i)D_j + (\partial^2\omega_{ji}/\partial x_i\partial x_j) + 2\omega_{ji}^2.$$

(ii) $$\begin{aligned}[D_k^2, hD_iD_j] = {} & 2\ (\partial h/\partial x_k)D_kD_iD_j + 2h\omega_{jk}D_iD_k + 2h\omega_{ik}D_kD_j \\ & + (\partial^2 h/\partial x_k^2)D_iD_j + 2h(\partial\omega_{jk}/\partial x_i)D_k + h(\partial\omega_{jk}/\partial x_k)D_i \\ & + h(\partial\omega_{ik}/\partial x_k)D_j + h(\partial^2\omega_{jk}/\partial x_i\partial x_k).\end{aligned}$$

(iii) $$\begin{aligned}[D_iD_j, hD_k] = {} & (\partial h/\partial x_j)D_iD_k + (\partial h/\partial x_i)D_jD_k + h\omega_{kj}D_i + h\omega_{ki}D_j \\ & + (\partial^2 h/\partial x_i\partial x_j)D_k + h(\partial\omega_{kj}/\partial x_i).\end{aligned}$$

Proof.

(i)
$$\begin{aligned}[D_i^2, D_j^2] &= D_i[D_i, D_j^2] + [D_i, D_j^2]D_i \\ &= -D_i\left[2\omega_{ij}D_j + \frac{\partial \omega_{ij}}{\partial x_j}\right] - \left[2\omega_{ij}D_j + \frac{\partial \omega_{ij}}{\partial x_j}\right]D_i \\ &= -2\frac{\partial \omega_{ij}}{\partial x_i}D_j - 2\omega_{ij}D_iD_j - \frac{\partial^2 \omega_{ij}}{\partial x_i \partial x_j} - \frac{\partial \omega_{ij}}{\partial x_j}D_i \\ &\quad - 2\omega_{ij}D_jD_i - \frac{\partial \omega_{ij}}{\partial x_j}D_i \\ &= 4\omega_{ji}D_jD_i + 2\frac{\partial \omega_{ji}}{\partial x_j}D_i + 2\frac{\partial \omega_{ji}}{\partial x_i}D_j + \frac{\partial^2 \omega_{ji}}{\partial x_i \partial x_j} + 2\omega_{ji}^2.\end{aligned}$$

(ii)
$$\begin{aligned}[D_k^2, hD_iD_j] &= -[hD_iD_j, D_k^2] \\ &= -h[D_iD_j, D_k^2] - [h, D_k^2]D_iD_j \\ &= -h\{D_i[D_j, D_k^2] + [D_i, D_k^2]D_j\} + \left(\frac{\partial^2 h}{\partial x_k^2} + 2\frac{\partial h}{\partial x_k}D_k\right)D_iD_j \\ &= -h\left\{D_i\left(-2\omega_{jk}D_k - \frac{\partial \omega_{jk}}{\partial x_k}\right) + \left(-2\omega_{ik}D_k - \frac{\partial \omega_{ik}}{\partial x_k}\right)D_j\right\} \\ &\quad + \left(\frac{\partial^2 h}{\partial x_k^2} + 2\frac{\partial h}{\partial x_k}D_k\right)D_iD_j \\ &= -h\left\{-2\frac{\partial \omega_{jk}}{\partial x_i}D_k - 2\omega_{jk}D_iD_k - \frac{\partial^2 \omega_{jk}}{\partial x_i \partial x_k} - \frac{\partial \omega_{jk}}{\partial x_k}D_i\right. \\ &\qquad \left. - 2\omega_{ik}D_kD_j - \frac{\partial \omega_{ik}}{\partial x_k}D_j\right\} + \left(\frac{\partial^2 h}{\partial x_k^2} + 2\frac{\partial h}{\partial x_k}D_k\right)D_iD_j \\ &= 2\frac{\partial h}{\partial x_k}D_kD_iD_j + 2h\omega_{jk}D_iD_k + 2h\omega_{ik}D_kD_j + \frac{\partial^2 h}{\partial x_k^2}D_iD_j \\ &\quad + 2h\frac{\partial \omega_{jk}}{\partial x_i}D_k + h\frac{\partial \omega_{jk}}{\partial x_k}D_i \\ &\quad + h\frac{\partial \omega_{ik}}{\partial x_k}D_j + h\frac{\partial^2 \omega_{jk}}{\partial x_i \partial x_k}.\end{aligned}$$

(iii)
$$\begin{aligned}[D_iD_j, hD_k] &= -h[D_k, D_iD_j] - [h, D_iD_j]D_k \\ &= h[D_iD_j, D_k] + [D_iD_j, h]D_k \\ &= h\{D_i[D_j, D_k] + [D_i, D_k]D_j\} \\ &\quad + D_i[D_j, h]D_k + [D_i, h]D_jD_k \\ &= h\{D_i\omega_{kj} + \omega_{ki}D_j\} \\ &\quad + D_i\frac{\partial h}{\partial x_j}D_k + \frac{\partial h}{\partial x_i}D_jD_k \\ &= h\,\omega_{kj}D_i + h\,\frac{\partial \omega_{kj}}{\partial x_i} \\ &\quad + h\omega_{ki}D_j + \frac{\partial h}{\partial x_j}D_iD_k\end{aligned}$$

$$
\begin{aligned}
&+\frac{\partial^2 h}{\partial x_i \partial x_j} D_k + \frac{\partial h}{\partial x_i} D_j D_k \\
&= \frac{\partial h}{\partial x_j}\, D_i D_k + \frac{\partial h}{\partial x_i}\, D_j D_k \\
&+ h\,\omega_{kj} D_i + h\,\omega_{ki} D_j + \frac{\partial^2 h}{\partial x_i \partial x_j} D_k \\
&+ h\,\frac{\partial \omega_{kj}}{\partial x_i}. \qquad \square
\end{aligned}
$$

LEMMA 7. *Let $\tilde{x} = Rx$ be an orthogonal change of coordinate, i.e., R is an orthogonal matrix. Then*

(1) $\tilde{f}(\tilde{x}) = Rf(x)$;

(2) $\tilde{L}_0 = L_0$;

(3) $(\tilde{\omega}_{ji}) = R(\omega_{\ell k})R^T$ *where* $\tilde{L}_0 = \frac{1}{2}\big(\sum_{i=1}^n \tilde{D}_i^2 - \tilde{\eta}(\tilde{x})\big)$, $\tilde{D}_i = \partial/\partial\tilde{x}_i - \tilde{f}_i$, $\tilde{h}(\tilde{x}) = h(x)$, $\tilde{\eta}(\tilde{x}) = \sum_{i=1}^n (\partial\tilde{f}_i(\tilde{x})/\partial\tilde{x}_i) + \tilde{f}(\tilde{x}) \cdot \tilde{f}(\tilde{x}) + \sum_{i=1}^m \tilde{h}_i^2(\tilde{x})$, *and* $\tilde{\omega}_{ji} = (\partial\tilde{f}_i/\partial\tilde{x}_j) - (\partial\tilde{f}_j/\partial\tilde{x}_i)$;

(4) $\tilde{E}$ *is isomorphic to* E *as Lie algebra, where* $\tilde{E}$ *is the Lie algebra generated by* $\tilde{L}_0, \tilde{h}_1, \ldots, \tilde{h}_m$.

Proof. Statement (1) is obvious. For (2), observe that

$$
\tilde{L}_0 = \frac{1}{2}\sum_{i=1}^{2} \frac{\partial^2}{\partial\tilde{x}_i} - \sum_{i=1}^{2} \tilde{f}_i(\tilde{x}) \frac{\partial}{\partial\tilde{x}_i} - \sum_{i=1}^{2} \frac{\partial\tilde{f}_i}{\partial\tilde{x}_i} - \frac{1}{2}\sum_{i=1}^{m} \tilde{h}_i^2.
$$

Let S be the inverse matrix of R. Then $x = S\tilde{x}$ and $S = (s_{ij}) = R^T = (r_{ji})$.

$$
\begin{aligned}
\tilde{L}_0 &= \frac{1}{2}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n} \frac{\partial x_j}{\partial\tilde{x}_i}\frac{\partial}{\partial x_j}\Big)^2 - \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} r_{ij} f_j(x)\Big)\Big(\sum_{j=1}^{n} \frac{\partial x_j}{\partial\tilde{x}_i}\frac{\partial}{\partial x_j}\Big) \\
&\quad - \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}\frac{\partial f_j}{\partial\tilde{x}_i}(x) - \frac{1}{2}\sum_{i=1}^{m} h_i^2(x) \\
&= \frac{1}{2}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n} s_{ji}\frac{\partial}{\partial x_j}\Big)^2 - \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} r_{ij} f_j(x)\Big)\Big(\sum_{j=1}^{n} s_{ji}\frac{\partial}{\partial x_j}\Big) \\
&\quad - \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}\sum_{k=1}^{n}\frac{\partial f_j}{\partial x_k}\frac{\partial x_k}{\partial\tilde{x}_i} - \frac{1}{2}\sum_{i=1}^{m} h_i^2(x) \\
&= \frac{1}{2}\sum_{i,j,k=1}^{n} s_{ji}s_{ki}\frac{\partial^2}{\partial x_j \partial x_k} - \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} r_{ik} f_k(x) s_{ji}\frac{\partial}{\partial x_j} \\
&\quad - \sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}\sum_{k=1}^{n} s_{ki}\frac{\partial f_j}{\partial x_k} - \frac{1}{2}\sum_{i=1}^{m} h_i^2(x) \\
&= \frac{1}{2}\sum_{j,k=1}^{n} \delta_{jk}\frac{\partial^2}{\partial x_j \partial x_k} - \sum_{j,k=1}^{n} \delta_{jk} f_k(x)\frac{\partial}{\partial x_j} - \sum_{j,k=1}^{n} \delta_{jk}\frac{\partial f_j}{\partial x_k} - \frac{1}{2}\sum_{i=1}^{m} h_i^2(x) \\
&= \frac{1}{2}\sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2} - \sum_{j=1}^{n} f_j(x)\frac{\partial}{\partial x_j} - \sum_{j=1}^{n} \frac{\partial f_j}{\partial x_j} - \frac{1}{2}\sum_{i=1}^{m} h_i^2(x)
\end{aligned}
$$

$= L_0.$

Statement (3) follows from the following computation:

$$\begin{aligned}\widetilde{\omega}_{ji} = \frac{\partial \widetilde{f}_i}{\partial \widetilde{x}_j} - \frac{\partial \widetilde{f}_j}{\partial \widetilde{x}_i} &= \sum_{k=1}^{n} r_{ik}\frac{\partial f_k}{\partial \widetilde{x}_j} - \sum_{k=1}^{n} r_{jk}\frac{\partial f_k}{\partial \widetilde{x}_i} \\ &= \sum_{k=1}^{n} r_{ik}\sum_{\ell=1}^{n}\frac{\partial x_\ell}{\partial \widetilde{x}_j}\frac{\partial f_k}{\partial x_\ell} - \sum_{k=1}^{n} r_{jk}\sum_{\ell=1}^{n}\frac{\partial x_\ell}{\partial \widetilde{x}_i}\frac{\partial f_k}{\partial x_\ell} \\ &= \sum_{k,\ell=1}^{n} r_{ik}\frac{\partial f_k}{\partial x_\ell}s_{\ell j} - \sum_{k,\ell=1}^{n} r_{jk}\frac{\partial f_k}{\partial x_\ell}s_{\ell i} \\ &= \sum_{k,\ell=1}^{n} s_{ki}\frac{\partial f_k}{\partial x_\ell}r_{j\ell} - \sum_{k,\ell=1}^{n} r_{j\ell}\frac{\partial f_\ell}{\partial x_k}s_{ki} \\ &= \sum_{k,\ell=1}^{n} r_{ik}r_{j\ell}\left(\frac{\partial f_k}{\partial x_\ell} - \frac{\partial f_\ell}{\partial x_k}\right) \\ &= \sum_{k,\ell=1}^{n} r_{ik}r_{j\ell}\omega_{\ell k}\end{aligned}$$

Statement (4) is a particular case of Brockett's result in [Br3].

3. Classification theorems. Let us first recall that the following two theorems were stated in Ocone [Oc].

THEOREM 8 (Ocone). *With the notation in* §2, *let* $n = m = p = 1$, $g = 1$. *Then* $\dim E$ *is finite only if* (i)

$$(*) \qquad h(x) = \alpha x, \quad \textit{and} \quad f' + f^2 = ax^2 + bx + c$$

or

(ii) $h(x) = \alpha x^2 + \beta x$, $\alpha \neq 0$ *and*

$$(**) \qquad f' + f^2 = -h^2 + a(2\alpha x + \beta)^2 + b + c(2\alpha x + \beta)^{-2}$$

$$(***) \qquad \textit{or} \quad f' + f^2 = -h^2 + ax^2 + bx + c.$$

THEOREM 9 (Ocone). *If* f *satisfies* $(*)$, f *must have a singularity in any unbounded interval.*

The following theorem follows easily from Ocone's Theorem 8 and Theorem 9 in the case where $m = 1$. Since Theorem 8 was stated without proof in [Oc], it is interesting to know that Theorem 9 follows from the proof of Theorem A as well. In fact we do not need to assume $m = 1$.

THEOREM A. *Suppose that the state space of the filtering system* (2.0) *is of dimension one. If the estimation algebra* E *is finite-dimensional, then one of the following holds:* (i) E *is a real vector space of dimension* 4 *with basis given by* 1, x, $D = (\partial/\partial x) - f$ *and* $L_0 = \frac{1}{2}(D^2 - \eta)$ *or* (ii) E *is a real vector space of dimension* 2 *with basis given by* 1, *and* $L_0 = \frac{1}{2}(D^2 - \eta)$ *or* (iii) E *is a real vector space of dimension* 1 *with basis given by* $L_0 = \frac{1}{2}(D^2 - \eta)$.

Proof. In view of Theorem 3, all the observation terms h_i $1 \leq i \leq m$ are necessarily affine polynomials. So we have only three cases.

If all the h_i for $1 \leq i \leq m$ are actually zero, then obviously we are in case (iii) above.

If all the h_i for $1 \leq i \leq m$ are at most constants and one of them is nonzero, then $1 \in E$. By Lemma 5 (iv), we have

$$[L_0, 1] = \frac{1}{2}[D^2 - \eta, 1] = 0.$$

Therefore we are in case (ii) above.

Finally we may assume that there is a constant c such that $x + c$ is in E. In view of Lemma 5, we have

$$[L_0, x + c] = \tfrac{1}{2}[D^2 - \eta, x + c] = D, \tag{3.1}$$

$$[D, x + c] = 1, \tag{3.2}$$

$$[L_0, D] = \frac{1}{2}[D^2 - \eta, D] = \frac{1}{2}\frac{d\eta}{dx}. \tag{3.3}$$

$d\eta/dx \in E$ implies η is a polynomial of degree at most 3 by Theorem 2. Recall that

$$\frac{df}{dx} + f^2 = \eta - \sum_{i=1}^{m} h_i^2. \tag{3.4}$$

If η is a polynomial of degree 3, then $\eta - \sum_{i=1}^{m} h_i^2$ is also a polynomial of degree 3. According to Theorem 4, (3.4) has no C^∞ solution f since

$$\lim_{x \to +\infty} (\eta - \sum_{i=1}^{m} h_i^2) = -\infty \quad \text{or} \quad \lim_{x \to -\infty} (\eta - \sum_{i=1}^{m} h_i^2) = -\infty.$$

This leads to a contradiction. Therefore, we have shown that η is a polynomial of degree 2. In view of (3.1)–(3.3), E is four-dimensional real vector space with basis 1, x, $D = (d/dx) - f$ and $L_0 = \frac{1}{2}(D^2 - \eta)$. □

THEOREM B. *Suppose that the state space of the filtering system* (2.0) *is of dimension two. If E is the finite-dimensional estimation algebra with maximal rank, then E is a real vector space of dimension* 6 *with basis given by* 1, x_1, x_2, D_1, D_2, *and* L_0.

Proof. Since E is a finite-dimensional estimation algebra with maximal rank, there are constants c_i's such that $x_i + c_i$ is in E for $i = 1, 2$. In view of Lemma 5, we have the following

$$[L_0, x_j + c_j] = \frac{1}{2}\left[\sum_{i=1}^{2} D_i^2 - \eta, x_j\right] = \frac{1}{2}\sum_{i=1}^{2}[D_i^2, x_j] = D_j \in E, \tag{3.5}$$

$$\omega_{ji} = [D_i, D_j] \in E \tag{3.6}$$

(3.7)

$$Y_j = [L_0, D_j] = \frac{1}{2}\left[\sum_{i=1}^{2} D_i^2 - \eta, D_j\right] = -\sum_{i=1}^{2}\left(\omega_{ij} D_i + \frac{1}{2}\frac{\partial \omega_{ij}}{\partial x_i}\right) + \frac{1}{2}\frac{\partial \eta}{\partial x_j}$$

$$= \sum_{i=1}^{2}\left(\omega_{ji} D_i + \frac{1}{2}\frac{\partial \omega_{ji}}{\partial x_i}\right) + \frac{1}{2}\frac{\partial \eta}{\partial x_j} \in E,$$

$$[Y_j, \omega_{k\ell}] = \left[\sum_{i=1}^{2} \omega_{ji} D_i + \frac{1}{2}\sum_{i=1}^{2} \frac{\partial \omega_{ji}}{\partial x_i} + \frac{1}{2}\frac{\partial \eta}{\partial x_j},\ \omega_{k\ell}\right] = \sum_{i=1}^{2} \omega_{ji} \frac{\partial \omega_{k\ell}}{\partial x_i} \in E, \tag{3.8}$$

$$[Y_j, D_k] = \left[\sum_{i=1}^{2} \omega_{ji} D_i + \frac{1}{2}\sum_{i=1}^{2} \frac{\partial \omega_{ji}}{\partial x_i} + \frac{1}{2}\frac{\partial \eta}{\partial x_j},\ D_k\right] = \sum_{i=1}^{2}\left(\omega_{ji}\omega_{ki} - \frac{\partial \omega_{ji}}{\partial x_k} D_i\right) - \frac{1}{2}\sum_{i=1}^{2} \frac{\partial^2 \omega_{ji}}{\partial x_k \partial x_i} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_k \partial x_j}. \tag{3.9}$$

By Theorem 2 and (3.6), ω_{ij}'s are polynomials of degree less than or equal to 2. Recall that $\omega_{11} = 0 = \omega_{22}$. By (3.8), we have

$$\omega_{12}\frac{\partial \omega_{12}}{\partial x_1} \in E \quad \text{and} \quad \omega_{12}\frac{\partial \omega_{12}}{\partial x_2} \in E,$$

which implies that

$$\frac{\partial \omega_{12}^2}{\partial x_1} \in E \quad \text{and} \quad \frac{\partial \omega_{12}^2}{\partial x_2} \in E.$$

If ω_{12} were polynomial of degree 2, then there would be a nonzero polynomial of degree 3 in E, which contradicts Theorem 2. Therefore, we conclude that ω_{12} is a polynomial of degree at most 1. We will prove that ω_{12} is actually a constant. From (3.9) and (3.5), we have

$$\sum_{i=1}^{2} \omega_{ji}\omega_{ki} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_k \partial x_j} \in E. \tag{3.10}$$

Since

$$\sum_{i=1}^{2} \omega_{ji}\omega_{ki} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_k \partial x_j}$$

is a polynomial of degree at most 2 for all $1 \le j, k \le 2$ we deduce easily that η is a polynomial of degree at most 4. Assume that $\eta = a_{40}x_1^4 + a_{31}x_1^3 x_2 + a_{22}x_1^2 x_2^2 + a_{13}x_1 x_2^3 + a_{04}x_2^4 +$ degree 3 polynomial and $\omega_{12} = ax_1 + bx_2 + c$. Equation (3.10) implies that

$$\omega_{12}^2 - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2}, \quad \frac{\partial^2 \eta}{\partial x_1 \partial x_2}, \quad \text{and} \quad \omega_{12}^2 - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_2^2}$$

are in E. Hence we have

$$\begin{aligned} E \ni \omega_{12}^2 - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2} &= a^2 x_1^2 + 2abx_1x_2 + b^2 x_2^2 \\ &\quad - (6a_{40}x_1^2 + 3a_{31}x_1x_2 + a_{22}x_2^2) \\ &\quad + \text{ polynomial of degree one} \\ &= (a^2 - 6a_{40})x_1^2 + (2ab - 3a_{31})x_1x_2 + (b^2 - a_{22})x_2^2 \\ &\quad + \text{ polynomial of degree one.} \end{aligned} \tag{3.11}$$

$$E \ni \frac{\partial^2 \eta}{\partial x_1 \partial x_2} = 3a_{31}x_1^2 + 4a_{22}x_1x_2 + 3a_{13}x_2^2 + \text{polynomial of degree one.} \tag{3.12}$$

(3.13)
$$\begin{aligned} E \ni \omega_{12}^2 - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2} &= a^2x_1^2 + 2abx_1x_2 + b^2x_2^2 - (a_{22}x_1^2 + 3a_{13}x_1x_2 + 6a_{04}x_2^2) \\ &\quad + \text{polynomial of degree one} \\ &= (a^2 - a_{22})x_1^2 + (2ab - 3a_{13})x_1x_2 + (b^2 - 6a_{04})x_2^2 \\ &\quad + \text{polynomial of degree one.} \end{aligned}$$

Since $1 = [D_1, x_1 + c_1] \in E$, we have $1, x_1, x_2 \in E$. It follows from (3.11)–(3.13) that

$$(a^2 - 6a_{40})x_1^2 + (2ab - 3a_{31})x_1x_2 + (b^2 - a_{22})x_2^2 \in E, \tag{3.14}$$

$$3a_{31}x_1^2 + 4a_{22}x_1x_2 + 3a_{13}x_2^2 \in E, \tag{3.15}$$

$$(a^2 - a_{22})x_1^2 + (2ab - 3a_{13})x_1x_2 + (b^2 - 6a_{04})x_2^2 \in E. \tag{3.16}$$

We will prove that ω_{12} is a constant. If there is no polynomial of degree 2 in E, then we have $a = b = a_{22} = 0$. This implies that ω_{12} is a constant.

Suppose that there is a polynomial of degree 2 in E. Then, by using the affine transformation $\widetilde{x} = Rx$, where R is an orthogonal matrix, we may assume that there exists a degree 2 polynomial in E of the form $k_1x_1^2 + k_2x_2^2 +$ polynomial of degree one, where either $k_1 \neq 0$ or $k_2 \neq 0$. This can be seen by using Lemma 7 because $\widetilde{\omega}_{12} = \sum_{k,\ell=1}^{2} r_{2k}r_{1\ell}\omega_{\ell k}$ is still a polynomial in x_i of degree at most one. As $1, x_1, x_2 \in E$, we deduce that there exists a polynomial in E of the form $k_1x_1^2 + k_2x_2^2$, where either $k_1 \neq 0$ or $k_2 \neq 0$. Without loss of generality we may assume that $k_1 \neq 0$. So we have $p(x)$

$$p(x) = x_1^2 + kx_2^2 \in E \text{ where } k = k_2/k_1. \tag{3.17}$$

Case 1. $k \neq 0$.

We observe that

$$\begin{aligned} [Y_1, p(x)] &= \left[\omega_{12}D_2 + \frac{1}{2}\frac{\partial \omega_{12}}{\partial x_2} + \frac{1}{2}\frac{\partial \eta}{\partial x_1}\ ,\ x_1^2 + kx_2^2\right] \\ &= 2k\omega_{12}x_2 = 2k(ax_1x_2 + bx_2^2 + cx_2) \in E, \end{aligned} \tag{3.18}$$

$$\begin{aligned} [Y_2, p(x)] &= \left[\omega_{21}D_1 + \frac{1}{2}\frac{\partial \omega_{21}}{\partial x_1} + \frac{1}{2}\frac{\partial \eta}{\partial x_2}\ ,\ x_1^2 + kx_2^2\right] \\ &= -2\omega_{12}x_1 = -2(ax_1^2 + bx_1x_2 + cx_1) \in E. \end{aligned} \tag{3.19}$$

It follows from (3.18) and (3.19) that we have

$$ax_1x_2 + bx_2^2 \in E, \tag{3.20}$$

$$ax_1^2 + bx_1x_2 \in E. \tag{3.21}$$

Equations (3.20) and (3.21) imply that

$$a^2x_1^2 - b^2x_2^2 \in E. \tag{3.22}$$

On the other hand, we have

$$\begin{aligned}[L_0, p(x)] &= \frac{1}{2}\left[\sum_{i=1}^{2} D_i^2 - \eta,\ x_1^2 + kx_2^2\right] \\ &= \frac{1}{2}[D_1^2,\ x_1^2 + kx_2^2] + \frac{1}{2}[D_2^2,\ x_1^2 + kx_2^2] \\ &= \frac{1}{2}(2 + 4x_1D_1) + \frac{1}{2}(2k + 4kx_2D_2) \\ &= 2\left(x_1D_1 + kx_2D_2 + \frac{k+1}{2}\right) \in E,\end{aligned}$$

$$\left[x_1D_1 + kx_2D_2 + \frac{k+1}{2}\ ,\ x_1^2 + kx_2^2\right] = 2x_1^2 + 2k^2x_2^2 \in E.$$

So we have

$$x_1^2 + k^2x_2^2 \in E. \tag{3.23}$$

Equations (3.17) and (3.23) imply $(k^2 - k)x_2^2 \in E$. So if $k \neq 1$, then both x_1^2 and x_2^2 are in E. If $k = 1$, then it follows from (3.17) and (3.22) that $(a^2 + b^2)x_2^2 \in E$. If $a^2 + b^2 = 0$, then ω_{12} is constant as claimed. On the other hand, if $a^2 + b^2 \neq 0$, then we conclude that x_1^2, x_2^2 are in E. Therefore in view of Lemma 5, we have

$$\left[L_0,\ \frac{1}{2}x_1^2\right] = \frac{1}{4}[D_1^2 + D_2^2 - \eta,\ x_1^2] = \frac{1}{4}[D_1^2,\ x_2^2] = x_1D_1 + \frac{1}{2} \in E, \tag{3.24}$$

$$\left[L_0,\ \frac{1}{2}x_2^2\right] = \frac{1}{4}[D_1^2 + D_2^2 - \eta,\ x_2^2] = \frac{1}{4}[D_2^2,\ x_2^2] = x_2D_2 + \frac{1}{2} \in E, \tag{3.25}$$

$$\left[x_1D_1 + \frac{1}{2},\ x_2D_2 + \frac{1}{2}\right] = -x_1x_2\omega_{12} \in E. \tag{3.26}$$

By Theorem 2, $x_1x_2\omega_{12}$ is a polynomial of degree 2. So ω_{12} is a constant.

Case 2. $k = 0$.

By (3.19) we have $ax_1^2 + bx_1x_2 \in E$ which implies $bx_1x_2 \in E$. If $b \neq 0$, then $x_1x_2 \in E$. It follows that

$$\begin{aligned}[L_0, x_1x_2] &= \frac{1}{2}\left[\sum_{i=1}^{2} D_i^2 - \eta, x_1x_2\right] = \frac{1}{2}[D_1^2, x_1x_2] + \frac{1}{2}[D_2^2, x_1x_2] \\ &= x_2D_1 + x_1D_2 \in E,\end{aligned} \tag{3.27}$$

$$[x_2D_1 + x_1D_2,\ x_1x_2] = x_1^2 + x_2^2 \in E. \tag{3.28}$$

We deduce from (3.28) that x_1^2 and x_2^2 are in E. Hence (3.24)–(3.26) imply that ω_{12} must be a constant as claimed.

From now on, we assume that $b = 0$, i.e., $\omega_{12} = ax_1 + c$, and $p(x) = x_1^2$.

Let $Z_0 = \frac{1}{2}p(x) = \frac{1}{2}x_1^2 \in E$ and $Z_k = [L_0, Z_{k-1}]$. Then by (3.24) $Z_1 = [L_0, Z_0] = x_1D_1 + \frac{1}{2}$. In view of Lemma 5, we have

$$\begin{aligned} Z_2 = [L_0, Z_1] &= \frac{1}{2}\left[D_1^2 + D_2^2 - \eta,\ x_1D_1 + \frac{1}{2}\right] \\ &= \frac{1}{2}[D_1^2,\ x_1D_1] + \frac{1}{2}[D_2^2, x_1D_1] + \frac{1}{2}[x_1D_1,\ \eta] \\ &= D_1^2 + x_1\omega_{12}D_2 + \frac{1}{2}E_1(\eta) \text{ where } E_1 = x_1\frac{\partial}{\partial x_1}. \end{aligned}$$

Let U^k be the space of differential operators of order up to and including k. Then

$$\begin{aligned} Z_3 = [L_0, Z_2] &= \frac{1}{2}\left[D_1^2 + D_2^2 - \eta,\ D_1^2 + x_1\omega_{12}D_2 + \frac{1}{2}E_1(\eta)\right] \\ &= \frac{1}{2}[D_1^2, x_1\omega_{12}D_2] + \frac{1}{2}[D_2^2, D_1^2] + \frac{1}{2}[D_2^2, x_1\omega_{12}D_2] \bmod U^1 \\ &= \frac{\partial(x_1\omega_{12})}{\partial x_1}D_1D_2 + 2\omega_{12}D_1D_2 + \frac{\partial(x_1\omega_{12})}{\partial x_2}\ D_2^2 \bmod U^1 \\ &= (4ax_1 + 3c)D_1D_2 \bmod U^1. \end{aligned}$$

Here $(\ \cdot\ )$ mode U^k signifies a member of the affine class of operators obtained by adding members of U^k to the argument. Suppose $a \neq 0$. Then $A = Z_3/4a = \left(x_1 + 3c/4a\right)D_1D_2 \bmod U^1$ is an element in E. We claim that $(-1)^{k+1}Ad_A^kZ_2 = 2^kD_1^2D_2^k \bmod U^{k+1}$. For $k = 1$,

$$\begin{aligned} (-1)Ad_AZ_2 = [Z_2, A] &= \left[D_1^2 \bmod U^1, \left(x_1 + \frac{3c}{4a}\right)D_1D_2 \bmod U^1\right] \\ &= \left[D_1^2,\ \left(x_1 + \frac{3c}{4a}\right)D_1D_2\right] \bmod U^2 \\ &= 2D_1^2D_2 \bmod U^2. \end{aligned}$$

Suppose that it is true for $k-1$, i.e., $(-1)^kAd_A^{k-1}Z_2 = 2^{k-1}D_1^2D_2^{k-1} \bmod U^k$. Then

$$\begin{aligned} (-1)^{k+1}Ad_A^kZ_2 &= (-1)Ad_A[(-1)^kAd_A^{k-1}Z_2] \\ &= [2^{k-1}D_1^2D_2^{k-1} \bmod U^k,\ \left(x_1 + \frac{3c}{4a}\right)D_1D_2 \bmod U^1] \\ &= 2^{k-1}\left[D_1^2D_2^{k-1},\ \left(x_1 + \frac{3c}{4a}\right)D_1D_2\right] \bmod U^{k+1} \\ &= -\left(x_1 + \frac{3c}{4a}\right)[D_1D_2,\ 2^{k-1}D_1^2D_2^{k-1}] \\ &\quad - \left[x_1 + \frac{3c}{4a},\ 2^{k-1}D_1^2D_2^{k-1}\right]D_1D_2 \bmod U^{k+1}. \end{aligned} \tag{3.29}$$

We show that $[D_1D_2, D_1^2D_2^{k-1}] \equiv 0 \bmod U^{k+1}$. This can be seen easily by induction as follows. For $k = 1$, this follows from Lemma 6 (ii):

$$\begin{aligned}[D_1D_2, D_1^2D_2^{k-1}] &= -[D_1^2D_2^{k-2}D_2, D_1D_2] \\ &= -D_1^2D_2^{k-1}[D_2, D_1D_2] \\ &\quad - [D_1^2D_2^{k-2}, D_1D_2]D_2 = 0 \bmod U^{k+1}\end{aligned}$$

in view of Lemma 6 (iii) and induction hypothesis. Put this into (3.29), and we obtain

$$\begin{aligned}(-1)^{k+1}Ad_A^k Z_2 &= 2^{k-1}\left[D_1^2D_2^{k-1},\ x_1 + \frac{3c}{4a}\right]D_1D_2 \bmod U^{k+1} \\ &= 2^{k-1}D_1^2\left[D_2^{k-1},\ x_1 + \frac{3c}{4a}\right]D_1D_2 + 2^{k-1}\left[D_1^2,\ x_1 + \frac{3c}{4a}\right]D_2^{k-1}D_1D_2 \\ &= 2^{k-1}D_1^2\left\{D_2\left[D_2^{k-2}, x_1 + \frac{3c}{4a}\right] + \left[D_2,\ x_1 + \frac{3c}{4a}\right]D_2^{k-2}\right\}D_1D_2 \\ &\quad + 2^{k-1}\cdot 2D_1D_2^{k-1}D_1D_2 \bmod U^{k+1} \\ &= 2^kD_1D_2^{k-1}D_1D_2 \bmod U^{k+1} \\ &= 2^kD_1^2D_2^k \bmod U^{k+1}.\end{aligned}$$

This proves our claim. We have shown that if $a \neq 0$, then E is infinite-dimensional. Hence the finite-dimensionality of E implies that $a = 0$, i.e., ω_{12} is a constant. We can apply Theorem 6 of [Ya] to deduce our result. □

REFERENCES

[Br-Cl] R. W. Brockett and J. M. C. Clark, *The geometry of the conditional density functions*, in Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs, et. al., eds., Academic Press, New York, 1980, pp. 299–309.

[Br1] R. W. Brockett, *Nonlinear systems and nonlinear estimation theory*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dordrecht, 1981.

[Br2] ———, *Nonlinear Control Theory and Differential Geometry*, Proceedings of the International Congress of Mathematics, 1983, pp. 1357–1368.

[Br3] ———, *Classification and Equivalence in Estimation Theory*, Proceedings, 18th IEEE Conference on Decision and Control, Fort Lauderdale, FL, December, 1979.

[Co] P. C. Collingwood, *Some remarks on estimation algebras*, Systems Control Lett., 7 (1986), pp. 217–224.

[Da] M. H. A. Davis, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 125–139.

[Da-Ma] M. H. A. Davis and S. I. Marcus, *An introduction to nonlinear filtering*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds.,, eds., Reidel, Dordrecht, 1981.

[D-T-W-Y] R. T. Dong, L. F. Tam, W. S. Wong and S. S.-T. Yau,, *Structure and classification theorems of finite dimensional exact estimation algebras*, SIAM J. Control Optim., to appear.

[Oc] D. L. Ocone, *Finite dimensional estimation algebras in nonlinear filtering*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dorrecht, 1981.

[Mi] S. K. Mitter, *On the analogy between mathematical problems of nonlinear filtering and quantum physics*, Ricerche di Automatica, 10 (1979), pp. 163–216.

[T-W-Y] L. F. TAM, W. S. WONG, AND S. S.-T. YAU, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.

[Wo] W. S. WONG, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117–124.

[Ya] S. S.-T. YAU, *Finite Dimensional Filters with Nonlinear Drift I: A Class of Filters Including both Kalman-Bucy Filters and Benes Filters*, J. Math. Systems, Estimation Control,, to appear.

[Ya-Ch] S. S.-T. YAU AND W.-L. CHIOU, *Recent results on classification of finite dimensional estimation algebras: dimension of state space* ≤ 2, Proceedings of 30th Conf. on Decision and Control, Brighton, England, Dec. 11-13, 1991, pp. 2758–2760.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 311–331, March 1994

EXISTENCE THEORY AND THE MAXIMUM PRINCIPLE FOR RELAXED INFINITE-DIMENSIONAL OPTIMAL CONTROL PROBLEMS*

H. O. FATTORINI†

Abstract. Existence theorems are considered for relaxed optimal control problems described by semilinear systems in Banach spaces. Relaxed controls are used whose values are finitely additive probability measures; this class of relaxed controls does not require special assumptions (such as compactness) on the control set. Under suitable conditions, relaxed trajectories coincide with those obtained from differential inclusions. Existence theorems for relaxed controls are obtained that apply to distributed parameter systems described by semilinear parabolic and wave equations, as well as a version of Pontryagin's maximum principle for relaxed optimal control problems.

Key words. relaxed controls, optimal controls, relaxation

AMS subject classifications. 93E20, 93E25

1. Introduction. Consider a finite-dimensional control system described by a vector differential equation

$$y'(t) = f(t, y(t), u(t)) \tag{1.1}$$

in m-dimensional Euclidean space $\mathbb{R}^m$, with cost functional

$$y_0(t) = \int_0^t f_0(\sigma, y(\sigma), u(\sigma))d\sigma \tag{1.2}$$

and control constraint $u(t) \in U$. It has long been known that optimal controls may fail to exist unless the set $f(t, y, U)$ is convex in $\mathbb{R}$ and the functional is weakly lower semicontinuous. A classical example is the system $y'(t) = u(t)$ in $\mathbb{R}^1$ with $U = [-1, 1]$, terminal time $\bar{t} = 1$ and cost functional

$$y_0(t) = \int_0^t \{y(\sigma)^2 + (u(\sigma)^2 - 1)^2\}d\sigma. \tag{1.3}$$

We can construct a sequence of "handsaw" functions with derivative $y'(t) = u(t) = \pm 1$ and teeth height tending to zero (number of teeth tending to infinity) for which the cost functional (1.3) is arbitrarily small; however, the value zero is not attained by any control (see [4] for additional details and other examples). This difficulty was surmounted independently by Filippov [26], Warga [34], [35], and Gamkrelidze [29] by means of extensions of the class of trajectories of (1.1). Warga's extension uses probability measure-valued controls $\mu(t, du)$ (called *relaxed* controls) and replaces the original equation by

$$y'(t) = \int_U f(t, y(t), u)\mu(t, du), \tag{1.4}$$

with a correspondingly relaxed cost functional. Relaxed controls are a natural generalization of *Young measures* [37], [38] from calculus of variations to control problems. Under suitable assumptions on the control set, limits of relaxed trajectories are relaxed trajectories themselves, which is the basis of existence theorems. For full expositions of

* Received by the editors October 7, 1991; accepted for publication (in revised form) August 10, 1992. This work was supported in part by National Science Foundation grant DMS-9001793.

† Department of Mathematics, University of California, Los Angeles, California 90024.

finite-dimensional relaxed control theory see [36] and [38]; since this paper is primarily on infinite-dimensional problems, we have not included in the references any sample of the large existing bibliography on finite-dimensional relaxed controls. See also §3 and [19] and [21] for additional details on Filippov's solution by means of differential inclusions and on its relation to measure valued controls, and [25] for applications to existence theory of optimal fluid flow problems.

Both approaches to relaxation have been extended to semilinear systems in Banach spaces,

$$y'(t) = Ay(t) + f(t, y(t), u(t)), \tag{1.5}$$

where A is the infinitesimal generator of a strongly continuous semigroup, the measure-valued control approach in [2] and [33] and the differential inclusion approach in [3] and [28]. A different class of measure-valued relaxed controls was introduced in [19], where compactness assumptions in the control set U or weak measurability assumptions on $f(t, y, u)$ are avoided by using finitely additive measures. These measures have been used in a different vein in minimization problems; see, for instance, [7].

We present in this paper various results on the relaxed controls introduced in [19], whose definition is reproduced in §2. Sections 4–6 are on existence of relaxed optimal controls for systems described by (1.5). All existence theorems follow the same pattern: existence of a minimizing sequence with suitable properties is assumed and the optimal relaxed control is obtained by taking limits. However, the success of the finite-dimensional theory has no complete counterpart here: in finite dimensions, convergence of trajectories follows from the Arzelà–Ascoli theorem, while in infinite-dimensional spaces we must rely on compactness properties of the equation or the control operator. We present essentially two examples: the first is an abstract parabolic system where A generates a compact semigroup (§5), the second (§6) a semilinear wave equation treated as an abstract differential equation. In these applications, the setting is a reflexive separable Banach space. We cover in §7 an extension of the theory to nonreflexive Banach spaces with applications to parabolic systems in L^1 spaces and spaces of continuous functions.

The last section is on Pontryagin's maximum principle. We show that, using the theory for ordinary controls, understood in sufficient generality, the maximum principle for relaxed controls results. (In the finite-dimensional case, this is stressed in [38, §39].)

2. Spaces of ordinary and relaxed controls. The control system is

$$y'(t) = Ay(t) + f(t, y(t), u(t)), \qquad y(0) = \zeta \tag{2.1}$$

in a time interval $0 \le t \le T$, where A is the infinitesimal generator of a strongly continuous semigroup $S(t)$ in the Banach space E and $\zeta \in E$. The *control set* U is a normal topological space. By definition, solutions of (2.1) are elements of the space $C(0, T; E)$ of continuous E-valued functions defined in $0 \le t \le T$ satisfying

$$y(t) = S(t)\zeta + \int_0^t S(t-\sigma) f(\sigma, y(\sigma), u(\sigma))\, d\sigma \tag{2.2}$$

in $0 \le t \le T$, where the integral is understood in the sense of Lebesgue–Bochner. The *admissible control space* $U_{ad}(0, T; U)$ is an arbitrary space of U-valued functions $t \to u(t)$ such that

(I) for every $y(\cdot) \in C(0, T; E)$ and every $u(\cdot) \in U_{ad}(0, T; U)$ the function $t \to f(t, y(t), u(t))$ is (strongly measurable and) integrable in $0 \le t \le T$.

The key result in the definition of relaxed controls is Theorem 2.1, due to Dieudonné [9], [10], on the dual of $L^1(0,T;E)$ for an arbitrary Banach space E. The space $L^\infty_w(0,T;E^*)$ consists of all E^*-valued E-weakly measurable functions $g(\cdot)$ such that there exists C with

$$|\langle g(t), y\rangle| \le C\|y\| \qquad \text{a.e. in } 0 \le t \le T, \quad y \in E \tag{2.3}$$

(the null set implicit in "a.e." may depend on y). The norm $\|\cdot\|$ in $L^\infty_w(0,T;E^*)$ is the least C such that (2.3) holds. The equivalence relation in $L^\infty_w(0,T;E^*)$ is $f(\cdot) \approx g(\cdot)$ if and only if $\langle f(t), y\rangle = \langle g(t), y\rangle$ almost everywhere for every $y \in E$.

THEOREM 2.1. *The dual space $L^1(0,T;E)^*$ is isometrically isomorphic to $L^\infty_w(0,T;E^*)$ through the pairing*

$$\langle g(\cdot), f(\cdot)\rangle = \int_0^T \langle g(\sigma), f(\sigma)\rangle d\sigma. \tag{2.4}$$

For a complete proof see [9] and [10]. Theorem 2.1 is a simple consequence of the Dunford–Pettis theorem [11], [12, Th. 6, p. 503 and Lemma 8, p. 504], [31, Cor. 1, p. 89] on bounded linear operators from $L^1(0,T)$ into the dual E^* of a Banach space E. This theorem provides additional information on the space $L^\infty_w(0,T;E^*)$, for instance, the following result.

THEOREM 2.2. *There exists a linear operator $S : L^\infty_w(0,T;E^*) \to L^\infty_w(0,T;E^*)$ such that* (a) *Sg belongs to the equivalence class of g;* (b) *the function $t \to \|(Sg)(t)\|$ is measurable in $0 \le t \le T$;* (c) $\sup_{0\le t\le T} \|(Sg)(t)\| = \|g\|$.

For a proof see [31, Cor. 1, p. 89]. In particular, Theorem 2.2 shows that each equivalence class in $L^\infty_w(0,T;E^*)$ contains an element $g(\cdot)$ such that

$$\|g(t)\| \le \|g\| \qquad (0 \le t \le T). \tag{2.5}$$

We shall use Theorem 2.1 with various measure spaces taking the role of E^*. Let U be an arbitrary set, Φ a field of subsets of U. We denote by $\Sigma_{ba}(U,\Phi)$ the Banach space of all bounded finitely additive measures defined in Φ endowed with the total variation norm. If U is a normal topological space and Φ_c is the field generated by the closed sets of U, the space $\Sigma_{rba}(U,\Phi_c)$ of all regular, bounded finitely additive measures is a closed subspace of $\Sigma_{ba}(U,\Phi_c)$ (thus a Banach space) under the total variation norm. Finally, if Φ_b is the Borel field of $U, \Sigma_{rca}(U,\Phi_b)$ denotes the space of all regular bounded countably additive measures defined in Φ_b, and is also a Banach space under the total variation norm.

We denote by $B(U)$ the space of all real valued bounded functions in U endowed with the supremum norm: if U is a topological space, $BC(U)$ denotes the subspace of all real valued bounded continuous functions in U. Both spaces are Banach spaces; if U is compact, $BC(U) = C(U)$, the space of all continuous functions in U. Finally, given an arbitrary field Φ of subsets of U, the space $B(U,\Phi)$ is defined as the closure in $B(U)$ of the set of all finite linear combinations of characteristic functions of sets in Φ.

THEOREM 2.3. (a) *The dual space $B(U,\Phi)^*$ of $B(U,\Phi)$ is isometrically isomorphic to $\Sigma_{ba}(U,\Phi)$, the duality pairing given by*

$$\langle \mu, f\rangle = \int_U f(u)\mu(du). \tag{2.6}$$

In particular, $B(U)^* = \Sigma_{ba}(U,\Phi), \Phi$ the field of all subsets of U. (b) *If U is a normal topological space, the dual space $BC(U)^*$ of $BC(U)$ is isometrically isomorphic to $\Sigma_{rba}(U,\Phi_c)$ with the same duality pairing, Φ_c the field generated by the closed sets of U.*

(c) *If U is a compact topological space, $C(U)^*$ is isometrically isomorphic to $\Sigma_{rca}(U, \Phi_b)$ with the same duality pairing.*

For the proof of (a)–(c), see [12, pp. 258, 262, 265].

The *relaxed control space* $V_r(0,T;U)$ consists of all elements $\mu(\cdot) \in L^1(0,T;BC(U))^* = L^\infty_w(0,T;\Sigma_{rba}(U,\Phi_c))$ that satisfy the following three conditions: (i)

$$\|\mu(\cdot)\| \leq 1 \tag{2.7}$$

(norm in $L^\infty_w(0,T;\Sigma_{rba}(U,\Phi_c))$. (ii) If $f(\cdot) \in L^1(0,T;BC(U))$ is such that $f(t,u) \geq 0$ for $u \in U$ almost everywhere in $0 \leq t \leq T$, then

$$\int_0^T \int_U f(t,u)\mu(t,du)dt \geq 0. \tag{2.8}$$

(iii) If e is a measurable set in $[0,T]$ and $\chi(t,u)$ is the characteristic function of $e \times U$, then

$$\int_0^T \int_U \chi(t,u)\mu(t,du)dt = \text{meas}(e). \tag{2.9}$$

In view of (2.5) we may always assume (if necessary selecting another element of the equivalence class) that an element of $V_r(0,T;U)$ satisfies

$$\|\mu(t)\| \leq 1 \qquad (0 \leq t \leq T). \tag{2.10}$$

LEMMA 2.4. *Let $\mu(\cdot) \in V_r(0,T;U)$ be such that* (2.10) *is satisfied almost everywhere. Then*

$$\mu(t) \geq 0, \qquad \mu(t,U) = \|\mu(t)\| = 1 \quad \textit{a.e. in } 0 \leq t \leq T. \tag{2.11}$$

In fact, let $\mathbf{1}(u) \equiv 1$. Since $\mu(t,U) = \langle \mu(t), \mathbf{1} \rangle$, the function $t \to \mu(t,U)$ is measurable. If $\mu(t,U) < 1$ in a set of positive measure, we may find $\varepsilon > 0$ and a set e of positive measure such that

$$\mu(t,U) \leq 1 - \varepsilon \qquad (t \in e).$$

If $\chi(t,u)$ is the characteristic function of $e \times U$, then

$$\int_0^T \int_U \chi(t,u)\mu(t,du) = \int_0^T \chi(t)\mu(t,U) \leq \text{meas}(e)(1-\varepsilon),$$

which contradicts condition (iii), and shows the second equality (2.11). Let t belong to the set e where $\mu(t,U) = 1, \|\mu(t)\| \leq 1$. Assume there exists a set $V \subseteq U$ such that $\mu(t,V) < 0$, and let W be the complement of V. Then $1 = \mu(t,U) = \mu(t,V) + \mu(t,W)$, whereas $|\mu|(t,U) \geq |\mu(t,V)| + |\mu(t,W)| > 1$, a contradiction. Accordingly, the first equality (2.11) holds as well in e. The fact that $\|\mu(\cdot)\|$ is measurable follows from the fact that $\|\mu(t)\| = \mu(t,U)$ almost everywhere.

We note that condition (i) is a consequence of (ii) and (iii).

The *relaxed control system* associated with (2.1) is, formally,

$$y'(t) = Ay(t) + \int_U f(t,y(t),u)\mu(t,du), \qquad y(0) = \zeta, \tag{2.12}$$

with $\mu(\cdot) \in V_r(0,T;U)$ (or with $\mu(\cdot) \in L_w^\infty(0,T;\Sigma_{rba}(U,\Phi_c))$ in general). To give sense to this equation we assume that E is reflexive and separable and the following conditions hold.

(II) (a) $f(t,y,\cdot)$ is continuous and bounded in U for t,y fixed; moreover, for every compact set $K \subseteq E$ there exists $\alpha(\cdot) = \alpha(K,\cdot) \in L^1(0,T)$ such that

$$\|f(t,y,u)\| \le \alpha(t) \qquad (0 \le t \le T, y \in K, u \in U). \tag{2.13}$$

(b) If $y(\cdot) \in C(0,T;E)$ and $y^* \in E^*$ then $t \to \langle y^*, f(t,y(t),\cdot)\rangle$ is a strongly measurable $BC(U)$-valued function.

Note that (b) and the bound (2.13) imply that $t \to \langle y^*, f(t,y(t),\cdot)\rangle$ belongs to $L^1(0,T;BC(U))$; in fact, we have $\|\langle y^*, f(t,y(t),\cdot)\rangle\| \le \|y^*\|\alpha(K,t)$, where $K = \{y(t); 0 \le t \le T\}$.

Equation (2.12) is understood in the following way. Define a function $\mathbf{f}$: $[0,T] \times E \times \Sigma_{rba}(U,\Phi_c) \to E$ by: $\mathbf{f}(t,y)\mu$ is the unique element of E satisfying

$$\langle y^*, \mathbf{f}(t,y)\mu\rangle = \int_U \langle y^*, f(t,y,u)\rangle \mu(du)$$

for all $y^* \in E^*$. The integral is well defined since, in view of (a) $\langle y^*, f(t,y,\cdot)\rangle \in BC(U)$; moreover, if $y \in K$ (K a compact set in E), $\|\langle y^*, f(t,y,\cdot)\rangle\|_{BC(U)} \le \alpha(K,t)\|y^*\|$ so that

$$\|\mathbf{f}(t,y)\mu\| \le \alpha(K,t)\|\mu\|_{\Sigma_{rba}(U,\Phi_c)} \qquad (0 \le t \le T, y \in K). \tag{2.14}$$

Let $\mu(\cdot) \in L_w^\infty(0,T;\Sigma_{rba}(U,\Phi_c)), y(\cdot) \in C(0,T;E)$ and $y^* \in E^*$. By virtue of (b) the function

$$\langle y^*, \mathbf{f}(\cdot, y(\cdot))\mu(\cdot)\rangle = \int_U \langle y^*, f(\cdot, y(\cdot), u)\rangle \mu(\cdot, du)$$

belongs to $L^1(0,T)$; since y^* is arbitrary, $\mathbf{f}(\cdot, y(\cdot))\mu(\cdot)$ is E^*-weakly measurable and by separability of E $\mathbf{f}(\cdot, y(\cdot))\mu(\cdot)$ is strongly measurable [30, p. 73]. We note that $\mathbf{f}(\cdot, y(\cdot))\mu(\cdot)$ depends only on the equivalence class of $\mu(\cdot)$ in $L_w^\infty(0,T;\Sigma_{rba}(U,\Phi_c))$. We recast (2.12) in the form

$$y'(t) = Ay(t) + \mathbf{f}(t, y(t))\mu(t), \qquad y(0) = \zeta. \tag{2.15}$$

The function $\mathbf{f}(t,y,u)$ satisfies (I) with respect to the control space $V_r(0,T;U)$. This makes it possible to interpret (2.15), the same as (2.1), as the integral equation

$$y(t) = S(t)\zeta + \int_0^t S(t-\sigma)\mathbf{f}(\sigma, y(\sigma))\mu(\sigma)d\sigma. \tag{2.16}$$

In the case where U is compact, the basic space is $L_w^\infty(0,T;\Sigma_{rca}(U,\Phi_b))$ instead of $L_w^\infty(0,T;\Sigma_{rba}(U,\Phi_c))$; this corresponds to using (c) of Theorem 2.3.

A requirement of any reasonable definition of relaxed control is that $U_{ad}(0,T;U) \subseteq V_r(0,T;U)$, that is, that every ordinary control $u(\cdot)$ can be replicated by a relaxed control $\mu(\cdot)$ in the sense that

$$f(t, y(t), u(t)) = \mathbf{f}(t, y(t))\mu(t). \tag{2.17}$$

We achieve this with $\mu(t) = \delta(\cdot - u(t))$, δ the Dirac delta. This relaxed control will belong to $V_r(0,T;U)$ if $U_{ad}(0,T;U)$ satisfies the following condition:

(I′) Let $u(\cdot) \in U_{ad}(0,T;U), y(\cdot) \in BC(U)$. Then

$$t \to y(u(t)) \tag{2.18}$$

is measurable in $0 \leq t \leq T$.

This will be satisfied, for instance, if U is a subset of a Banach space F and the elements of $U_{ad}(0,T;U)$ are strongly measurable.

We may base our definition of relaxed controls on the space $B(U)$ rather than $BC(U)$. In this case, U is just an arbitrary set and the space of relaxed controls $V_r(0,T;U)$ consists of all elements $\mu(\cdot) \in L^1(0,T;B(U))^* = L^\infty_w(0,T;\Sigma_{ba}(U,\Phi))$ that satisfy (i)–(iii). The relaxed control system is again (2.15) and the only assumptions on $f(t,y,u)$ are (a) and (b) of (II), with "continuous and bounded" replaced by "bounded" and the space $BC(U)$ replaced by $B(U)$. The price we pay for this enormous generality is that, if we insist on the inclusion $U_{ad}(0,T;U) \subseteq V_r(0,T;U)$, admissible controls must make (2.18) measurable for every $y \in B(U)$, which restricts severely the space $U_{ad}(0,T;U)$. Possibly, definitions using intermediate spaces $B(U,\Phi)$ for a suitable field Φ may be useful. We do the theory below for $BC(U)$ and $\Sigma_{rba}(U,\Phi_c)$ and use the shorthand $\Sigma_{rba}(U)$ for this space; every result in the rest of the paper can be immediately translated to $B(U)$ and $\Sigma_{ba}(U,\Phi) = \Sigma_{ba}(U)$.

We shall call $y(t,\mu)$ the trajectory of (2.15) corresponding to $\mu(\cdot) \in L^\infty_w(0,T;\Sigma_{rba}(U))$. The assumptions in this section only guarantee that solutions of (2.1) and (2.15) can be defined in their integral versions; existence of solutions is not guaranteed, even locally.

The requirement that $f(t,y,\cdot) \in CB(U)$ for (t,y) fixed leaves out such natural control terms as $f(t,y,u) = Lu$ with L a bounded operator and u in an unbounded control set U. Such control terms (and more general ones) appear in viscous flow problems [24], [25]. The theory of relaxed controls can be extended to cases like this requiring that controls in $V_r(0,T;U)$ satisfy suitable integrability conditions. See [24] and [25] for details.

3. Differential inclusions, relaxation theorems. Relaxed controls and trajectories have been defined in nonparametric form (that is, using differential inclusions) in [3] and [28]. In the formulation of [28] (somewhat rephrased) a pair $(y(\cdot),g(\cdot))$, where $y(\cdot) \in C(0,T;E)$ and $g(\cdot) \in L^1(0,T;E)$ is a *trajectory* of the differential inclusion

$$y'(t) \in Ay(t) + \overline{\text{conv}}\ f(t,y(t),U) \tag{3.1}$$

($\overline{\text{conv}}$ denotes closed convex hull) if and only if

$$g(t) \in \overline{\text{conv}}\ \{f(t,y(t),U)\} \qquad \text{a.e. in } 0 \leq t \leq T \tag{3.2}$$

$$y(t) = S(t)\zeta + \int_0^t S(t-\sigma)g(\sigma)d\sigma \qquad (0 \leq t \leq T). \tag{3.3}$$

A basic question is whether the solutions $y(\cdot)$ of (3.1) and those of (2.15) are the same. The answer is affirmative, as we see in the following theorem.

THEOREM 3.1. *Assume the space E is separable and that assumption* (II) *holds. Let $y(t,\mu)$ be a solution of* (2.15) *for $\mu(\cdot) \in V_r(0,T;U)$. Then $(y(t,\mu),\mathbf{f}(t,y(t))\mu(t))$ is a trajectory of the differential inclusion* (3.1).

For a proof see [21, Th. 4.1]. The result also holds (under modified definitions and assumptions) in certain nonreflexive spaces; see [21, §6]. Another important problem is that of establishing *relaxation* theorems, that is, showing that every solution $y(t,\mu)$ of (2.15) can be uniformly approximated in their interval of existence by solutions $y(t,u)$ of the original equation (2.1). For two results of this type in reflexive spaces, see [21, Ths. 5.4 and 5.5]; a generalization for certain nonreflexive spaces can be found in [21, §6].

4. Optimal control problems. Relaxation. Optimal control problems for (2.1) include a *cost functional* $y_0(t,u)$ defined in $U_{ad}(0,T;U)$ with values in $\mathbb{R}$. In many applications, the cost functional has the form

$$y_0(t,u) = \int_0^t f_0(\sigma, y(\sigma,u), u(\sigma))d\sigma + \phi_0(t, y(t,u)), \tag{4.1}$$

where $y(t,u)$ is the trajectory of (2.1) corresponding to $u(\cdot)$, $f_0 : [0,T] \times E \times U \to \mathbb{R}$, $\phi : [0,T] \times E \to \mathbb{R}$. Also, there may be a *target condition*

$$y(\bar{t},u) \in Y = \begin{matrix}\text{target set} \\ \text{cost functional}\end{matrix} \subseteq E, \tag{4.2}$$

where the *arrival time* $\bar{t}$ (the endpoint of the control interval $0 \le t \le \bar{t}$) may be free or fixed. The companion of assumption (I) for f in §2 is

(I_0) For every $y(\cdot) \in C(0,T;E)$ and every $u(\cdot) \in U_{ad}(0,T;U)$ the function $t \to f_0(t, y(t), u(t))$ is integrable in $0 \le t \le T$.

The original optimal control problem is that of minimizing $y_0(\bar{t},u)$ among all $u \in U_{ad}(0,\bar{t};U)$ whose corresponding trajectory $y(t,u)$ satisfies the target condition (4.2). Define

$$m = \inf y_0(\bar{t},u), \tag{4.3}$$

the infimum taken over all $u \in U_{ad}(0,\bar{t};U)$, such that $y(\cdot,u)$ satisfies (4.2); for the free arrival time problem $\bar{t}$ varies in $\mathbb{R}_+$. A natural restriction on the minimum is

$$-\infty < m < \infty. \tag{4.4}$$

The second inequality simply means that there exists some control $u \in U_{ad}(0,\bar{t};U)$ such that the trajectory $y(\cdot,u)$ satisfies the target condition (4.2). On the other hand, $m = -\infty$ means we can reach the target with arbitrarily low values of the functional, hence there is no optimal control. A *minimizing sequence* for the original problem is a sequence $\{u^n(\cdot)\} \subset U_{ad}(0,t_n;U)$ such that

$$\lim_{n\to\infty} \text{dist } (y(t_n,u^n), Y)) = 0, \tag{4.5}$$

$$\limsup_{n\to\infty} y_0(t_n,u^n) \le m. \tag{4.6}$$

The *relaxed control problem* is that described by (2.15), with controls $\mu(\cdot) \in V_r(0,T;U)$ and *relaxed cost functional*

$$y_0(t,\mu) = \int_0^t \mathbf{f}_0(\sigma, y(\sigma,\mu))\mu(\sigma)d\sigma + \phi_0(t, y(t,\mu)). \tag{4.7}$$

where $y(t,\mu)$ is the trajectory of (2.15) corresponding to the relaxed control $\mu(\cdot) \in V_r(0,\bar{t};U)$ and $\mathbf{f}_0$ is defined by

$$\mathbf{f}_0(t,y)\mu = \int_U f_0(t,y,u)\mu(du). \tag{4.8}$$

The target condition is the same. The companion of assumption (II) is

(II_0) (a) $f_0(t,y,\cdot)$ is bounded and continuous in U for t,y fixed. (b) If $y(\cdot)\in C(0,T;E)$ then $t\to f_0(t,y(t),\cdot)\in L^1(0,T;BC(U))$.

Assumption (II_0) makes possible the definition of $\mathbf{f}_0(t,y,u)$ and assures that $t\to \mathbf{f}_0(t,y(t))\mu(\cdot)\in L^1(0,T)$ for every $\mu(\cdot)\in L^\infty_w(0,T;\Sigma_{rba}(U))$, in particular for $\mu(\cdot)\in V_r(0,T;U)$, so that $y_0(t,\mu)$ can be defined. Since $y(t,\mu)$ may not exist or be unique, the same applies to $y_0(t,\mu)$.

Corresponding to the relaxed control problem we define

$$\mathbf{m}=\inf y_0(\bar t,\mu), \tag{4.9}$$

the infimum taken over all $\mu\in V_r(0,t;U)$ whose trajectories $y(\cdot,\mu)$ exist in $0\le t\le\bar t$ and satisfy the target condition (4.2). The observations about m apply to $\mathbf{m}$ as well. Since there are more relaxed than ordinary controls, we will always have

$$\mathbf{m}\le m. \tag{4.10}$$

In principle, strict inequality is possible, including situations where $\mathbf{m}<\infty, m=\infty$. (The target may be attained by a trajectory $y(\cdot,\mu)$ of the relaxed system but not by a trajectory $y(\cdot,u)$ of the original system.) It is desirable that

$$\mathbf{m}=m, \tag{4.11}$$

for if $\mathbf{m}<m$ it could be maintained that the relaxed problem "generalizes too much" the ordinary problem.

The basis of all our existence theorems will be the *closure theorem* below.

THEOREM 4.1. (closure of set of trajectories). *Let* $\{y(t,\mu_n)\}$ *be a sequence of trajectories of the relaxed system* (2.15) *in* $0\le t\le\bar t$. *Assume that* (a) *there exists* $y(\cdot)\in C(0,T;E)$ *such that* $y(t,\mu_n)\to y(t)$ *weakly in* E *for each* $t, 0\le t\le\bar t$; *and* (b) *for every* $y^*\in E^*$

$$\langle y^*, f(\cdot,y(\cdot,\mu_n),\cdot)\rangle\to\langle y^*,f(\cdot,y(\cdot),\cdot)\rangle \tag{4.12}$$

in $L^1(0,T;BC(U))$. *Then* $y(t)$ *is a trajectory of* (2.15), *that is, there exists a relaxed control* $\mu(\cdot)\in V_r(0,T;U)$ *such that* $y(t)=y(t,\mu)$.

Proof. We have

$$y(t,\mu_n)=S(t)\zeta+\int_0^t\int_U S(t-\sigma)f(\sigma,y(\sigma,\mu_n),u)\mu_n(\sigma,du)d\sigma.$$

Apply a functional $y^*\in E^*$:

$$\begin{aligned}\langle y^*,y(t,\mu_n)\rangle&=\langle y^*,S(t)\zeta\rangle+\int_0^t\int_U\langle S(t-\sigma)^*y^*,f(\sigma,y(\sigma,\mu_n),u)\rangle\mu_n(\sigma,du)d\sigma\\&=\langle y^*,S(t)\zeta\rangle+\int_0^{\bar t}\int_U\langle\chi(t-\sigma)S(t-\sigma)^*y^*,f(\sigma,y(\sigma,\mu_n),u)\rangle\mu_n(\sigma,du)d\sigma,\end{aligned} \tag{4.13}$$

where $\chi(\cdot)$ is the characteristic function of $t\ge 0$. Since E is reflexive the adjoint semigroup $S(t)^*$ is strongly continuous, thus, for each t we may approximate $\chi(t-\cdot)S(t-\cdot)^*y^*$ uniformly by step functions and use (4.12) to deduce that

$$\langle\chi(t-\cdot)S^*(t-\cdot)y^*,f(\sigma,y(\cdot,\mu_n),\cdot)\rangle\to\langle\chi(t-\cdot)S^*(t-\cdot)y^*,f(\sigma,y(\cdot),\cdot)\rangle$$

in $L^1(0,T;C(U))$. Select a $L^1(0,T;C(U))$-weakly convergent (generalized) subsequence of $\{\mu_n(\cdot)\}$, denoted in the same way. Taking limits and using the fact that y^* is arbitrary,

$$y(t) = S(t)\zeta + \int_0^t \int_U S(t-\sigma) f(\sigma, y(\sigma), u)\mu(\sigma, du)d\sigma,$$

which ends the proof.

More than a usable result, Theorem 4.1 is a "template" by means of which we shall cut all of our existence theorems. In these, condition (b) will result from assumptions on $f(t,y,u)$ and convergence of $y(t,\mu_n)$.

5. Abstract parabolic equations. We examine the existence problem for (2.15) when $S(t)$ is *compact*. In applying Theorem 4.1, we use the result below [22, §3] on the operator

$$(\Lambda g)(t) = \int_0^t S(t-\sigma)g(\sigma)d\sigma, \tag{5.1}$$

where the Banach space E is completely arbitrary.

LEMMA 5.1. *The operator Λ is bounded from $L^1(0,T;E)$ into $C(0,T;E)$. If $S(t)$ is compact for $t>0$ and $\{g_n(\cdot)\}$ is a sequence in $L^1(0,T;E)$ such that the integrals of $\|g_n(\cdot)\|$ are equicontinuous in $0 \le t \le T$, then $\{\Lambda g_n(\cdot)\}$ has a convergent subsequence in $C(0,T;U)$.*

In the following result the space E is again assumed to be reflexive and separable. The following assumptions on f, f_0, reinforce (II) and (II_0), respectively, with an assumption of continuity in y. In these assumptions, the space $C(0,T;E)$ is endowed with its usual supremum norm.

(III) (a) as in (II).

(b) as in (II) but with K bounded.

(c) If $\{y_n(\cdot)\} \subset C(0,T;E)$ is such that $y_n(\cdot) \to y(\cdot)$ in $C(0,T;E)$ then

$$\langle y^*, f(\cdot, y_n(\cdot), \cdot)\rangle \to \langle y^*, f(\cdot, y(\cdot), \cdot)\rangle \tag{5.2}$$

in $L^1(0,T;BC(U))$ (or, equivalently, almost everywhere, in view of the bound (2.13) that now holds for K bounded).

(III_0) (a) and (b) as in (II_0).

(c) If $\{y_n(\cdot)\} \subset C(0,T;E)$ is such that $y_n(\cdot) \to y(\cdot)$ in $C(0,T;E)$ then

$$f_0(\cdot, y_n(\cdot), \cdot) \to f_0(\cdot, y(\cdot), \cdot) \tag{5.3}$$

in $L^1(0,T)$.

(d) $\phi_0(t,y)$ is continuous.

THEOREM 5.2. *Let E be reflexive and separable, let $S(t)$ be a compact semigroup, and let the target set Y be closed. Assume that* (III) *and* (III_0) *hold and that there exists a minimizing sequence $\{\mu_n(\cdot)\}, \mu_n(\cdot) \in V_r(0,t_n;U)$ of relaxed controls with $\{t_n\}$ bounded and $y(\cdot,\mu_n)$ uniformly bounded in $0 \le t \le t_n$. Then there exists a relaxed solution $\overline{\mu}(\cdot)$ to the relaxed optimal control problem.*

Proof. Passing if necessary to a subsequence we may assume that $t_n \to \bar{t}$. For each n, if $t_n < \bar{t}$ extend $\mu_n(t)$ to $t_n < t \le \bar{t}$ setting $\mu(t) = \delta(t-u), u \in U$ arbitrary there; if $t_n \ge \bar{t}$ chop off $\mu_n(t)$ at $t = \bar{t}$. Call $\overline{\mu}_n(\cdot)$ the extended/chopped-off control. Since the sequence $\mathbf{f}(\cdot, y(\cdot,\mu_n))\mu_n(\cdot)$ satisfies (2.14), by Lemma 5.1 there exits a generalized subsequence, which we denote with the same symbol, such that $\{y(\cdot,\mu_n)\}$ is uniformly convergent in $C(0,\bar{t};U)$ to a function $y(\cdot)$. We now apply (III) and Theorem 4.1 and

obtain $\overline{\mu}(\cdot) \in V_r(0,\bar{t};U)$ such that $y(t) = y(t,\overline{\mu})$. The fact that the approximate target conditions (4.5) imply the exact target condition (4.2) follows from uniform convergence of the sequence of trajectories $\{y(t,\mu_n)\}$ and closedness of the target set Y. It only remains to show that $\overline{\mu}$ is an optimal relaxed control. To do this it suffices to prove that $y_0(\bar{t},\overline{\mu}) = \mathbf{m}$, $\mathbf{m}$ the minimum in (4.9). This follows using (III_0) and taking limits.

Theorem 5.2 reduces the existence problem to that of finding a minimizing sequence $\{\mu_n(\cdot)\}$ having the required properties. Existence of this sequence is certainly not guaranteed by hypothesis (III), which does not even imply local existence of the trajectory for each control $\mu(\cdot) \in V_r(0,T;U)$. Local existence is implied by the following local boundedness/local Lipschitz continuity assumption.

(IV) (a) $f(t,y,\cdot)$ is continuous and bounded in U for t,y fixed.

(b) If $y(\cdot) \in C(0,T;E)$ and $y^* \in E^*$ then $t \to \langle y^*, f(t,y(t),\cdot)\rangle$ is a strongly measurable $BC(U)$-valued function.

(c) For every $c > 0$ there exist $\alpha(\cdot) = \alpha(c,\cdot)$ and $\beta(\cdot) = \beta(c,\cdot) \in L^1(0,T)$ with

$$\|f(t,y,u)\| \le \alpha(t) \qquad (0 \le t \le T, \|y\| \le c, u \in U), \tag{5.4}$$

$$\|f(t,y',u) - f(t,y,u)\| \le \beta(t)\|y' - y\| \qquad (0 \le t \le T, \|y\|, \|y'\| \le c, u \in U). \tag{5.5}$$

Assumption (IV) implies corresponding local boundedness and local Lipschitz continuity properties for $\mathbf{f}(t,y)\mu$:

$$\|\mathbf{f}(t,y)\mu\| \le \alpha(t)\|y\|\ \|\mu\|_{\Sigma_{rba}(U)} \qquad (0 \le t \le T, \|y\| \le c, \mu \in \Sigma_{rba}(U)), \tag{5.6}$$

$$\begin{gathered}\|\mathbf{f}(t,y')\mu - f(t,y)\mu\| \le \beta(t)\|y' - y\|\ \|\mu\|_{\Sigma_{rba}(U)} \\ (0 \le t \le T, \|y\|, \|y'\| \le c, \mu \in \Sigma_{rba}(U)).\end{gathered} \tag{5.7}$$

Under (5.6) and (5.7), the integral equation (2.16) can be uniquely solved in some interval $[0,T'], 0 \le T' \le T$ by successive approximations.

It is plain that (IV) $\Rightarrow$ (III) $\Rightarrow$ (II).

Under stronger hypotheses, a priori bounds can be obtained that guarantee the following statement.

(V) For every $\mu(\cdot) \in V_r(0,T;U), y(t,\mu)$ exists in $0 \le t \le T$ and $\{y(t,\mu)\}$ is bounded independently of μ.

For results of this type under somewhat different conditions see [18, Lemma 5.1]. A priori bounds can also be obtained using Lyapunov functions or, in certain cases, by means of energy estimates [32].

If (V) holds, the only requirement on the minimizing sequence in Theorem 5.2 is boundedness of $\{t_n\}$. Even this follows automatically for some cost functionals.

COROLLARY 5.3. *Assume that*

(a) (III), (III_0), *and* (V) *hold;*

(b) *there exists a relaxed or ordinary control* $\mu(\cdot)$ *such that* $y(t,\mu)$ *satisfies the target condition* (4.2);

(c) $f_0(t,y,u) \ge \delta > 0$, (d) $\phi_0(t,y) \ge 0$. *Then there exists a relaxed solution* $\overline{\mu}(\cdot)$ *to the optimal control problem.*

In fact, condition (a) allows the construction of a minimizing sequence $\{\mu_n(\cdot)\}$. If $f_0(t,y,u) \ge \delta > 0$ and $\phi_0(t,y) \ge 0$ we obtain from (4.7) that $y_0(t_n,\mu_n) \ge t_n\delta$ which shows that $\{t_n\}$ must be bounded. Property (V) applied in some interval $[0,T]$ with $T > t_n$ then shows that $y(\cdot,\mu_n)$ is uniformly bounded in $0 \le t \le t_n$, so that Theorem 5.2 applies.

Verification of (b) is a *controllability* problem of interest in its own right. In some cases, the solution is obvious, for instance, where the target set Y is a ball with center at the origin and for some $\mu(\cdot) \in V_r(0,T;U)$ all solutions of (2.1) tend to zero as $t \to \infty$.

Many compact semigroups (such as those generated by uniformly elliptic partial differential operators) are also *holomorphic*. Modulo a translation we may assume that the origin belongs to the resolvent set of the infinitesimal generator A, and fractional powers $(-A)^\alpha$ can be defined for α real: $(-A)^\alpha S(t)$ is bounded in $t > 0$ for all $\alpha > 0$ and

$$\|(-A)^\alpha S(t)\| \le C_\alpha t^{-\alpha} e^{\omega t} \qquad (t \ge 0). \tag{5.8}$$

Writing the integral equation (2.15) in the form

$$y(t) = S(t)\zeta + \int_0^t (-A)^\alpha S(t-\sigma)(-A)^{-\alpha}\mathbf{f}(t, y(\sigma))\mu(\sigma)d\sigma, \tag{5.9}$$

we may prove an analog of Theorem 5.2 where (III) is required of $(-A)^{-\alpha}f(t,y,u)$ rather than of $f(t,y,u)$.

6. A hyperbolic distributed parameter system. Existence theorems for relaxed controls can be established without compactness assumptions on the semigroup $S(\cdot)$. As an example, we consider the semilinear wave equation

$$y_{tt}(t,x) = \sum_{j=1}^{m}\sum_{k=1}^{m} \partial^j(a_{jk}(x)\partial^k y(t,x)) - \phi(y(t,x),u(t)) \qquad (x \in \Omega), \tag{6.1}$$

$$y(t,x) = 0 \qquad (x \in \Gamma) \tag{6.2}$$

in an arbitrary domain Ω with boundary Γ in m-dimensional Euclidean space $\mathbb{R}^m$; the notations are $x = (x_1, x_2, \ldots, x_m), \partial^j = \partial/\partial x_j$. We assume that $a_{jk} = a_{kj}$ and that the operator A is uniformly elliptic: $\Sigma\Sigma a_{jk}(x)\xi_j\xi_k > \kappa|\xi|^2 (\xi \in \mathbb{R}^m, x \in \overline{\Omega}, \kappa > 0)$. The nonlinear term $\phi(y,u)$ is defined in $\mathbb{R} \times U$, where U is a normal topological space. Precise assumptions on ϕ will be given below.

We reduce (6.1) to a first-order system for a two-dimensional vector function $\mathbf{y} = (y, y_t) = (y, y_1)$ in the usual way:

$$y_t(t,x) = y_1(t,x), \tag{6.3}$$

$$y_{1t}(t,x) = \sum\sum \partial^j(a_{jk}(x)\partial^k y(t,x)) - \phi(y(t,x),u(t)) \tag{6.4}$$

and examine this system as an equation of the form (2.1),

$$\mathbf{y}'(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{f}(\mathbf{y}(t), u(t)) \tag{6.5}$$

in the space $E = H_0^1(\Omega) \times L^2(\Omega)$, where

$$\mathbf{A} = \begin{pmatrix} 0 & I \\ \sum\sum \partial^j(a_{jk}(x)\partial^k) & 0 \end{pmatrix},$$

$$\mathbf{f}((y,y_1),u) = \begin{pmatrix} 0 \\ \phi(y,u) \end{pmatrix}.$$

For the linear case $\phi = 0$ see for instance [16], where it is shown that $\mathbf{A}$ generates a strongly continuous group $\mathbf{S}(\cdot)$ in $E = H_0^1(\Omega) \times L^2(\Omega)$ and a precise description of the domain $D(\mathbf{A})$ is given. The assumption on the nonlinear term ϕ is (VI) below. This assumption has two parts, the latter depending on the dimension m.

(VI) (a) For each $y \in \mathbb{R}, \phi(y, \cdot) \in BC(U)$.

(b) Dimension $m > 2$. Let $\alpha = m/(m-2)$. Then

$$|\phi(y,u)| \leq C(1 + |y|^\alpha) \qquad (y \in \mathbb{R}, u \in U), \tag{6.6}$$

$$|\phi(y',u) - \phi(y,u)| \leq K(1 + |y|^{\alpha-1} + |y'|^{\alpha-1})|y' - y| \qquad (y, y' \in \mathbb{R}, u \in U). \tag{6.7}$$

Dimension $m = 2$. There exists some $\alpha > 0$ such that (6.6) and (6.7) hold.

Dimension $m = 1$. For each $c > 0$ there exist $C = C(c), K = K(c)$ such that

$$|\phi(y,u)| \leq C \qquad (y \in \mathbb{R}, |y| \leq c, u \in U), \tag{6.8}$$

$$|\phi(y',u) - \phi(y,u)| \leq K|y' - y| \qquad (y, y' \in \mathbb{R}, |y|, |y'| \leq c, u \in U). \tag{6.9}$$

We check that assumption (VI) implies (III), the local Lipschitz condition (IV), and the global existence-boundedness condition (V) in §5 beginning with the case $m > 2$. We shall use (a particular case of) Sobolev's imbedding theorem [1, p. 97]; assuming Ω satisfies the cone property [1, p. 66], the imbedding $W^{k,p}(\Omega) \to L^q(\Omega)$ holds for $kp < m, 1 \leq p \leq q \leq mp/(m-kp)$; in particular, we have the imbedding $H^1(\Omega) \to L^{2m/(m-2)}(\Omega)$; hence, if $y(\cdot) \in H^1(\Omega)$ (6.6) implies that $\phi(y(\cdot), u) \in L^{2m/(m-2)\alpha}(\Omega) = L^2(\Omega)$ with

$$\|\phi(y(\cdot),u)\|_{L^2(\Omega)} \leq C\|y(\cdot)\|_{H^1(\Omega)}, \tag{6.10}$$

so that $y(\cdot) \to \mathbf{f}(y(\cdot), u)$ maps E into E. On the other hand (6.7) implies that if $y(\cdot), y'(\cdot) \in H^1(\Omega)$,

$$\begin{aligned} &\|\phi(y'(\cdot),u) - \phi(y(\cdot),u)\|_{L^2(\Omega)} \\ &\qquad \leq K\left(\int_\Omega |y'(x) - y(x)|^2(1 + |y(x)|^{\alpha-1} + |y'(x)|^{\alpha-1})^2 dx\right)^{1/2}. \end{aligned} \tag{6.11}$$

Since $y(\cdot), y'(\cdot) \in L^{2m/(m-2)}(\Omega)$, we can apply Hölder's inequality with exponents $p = \alpha = m/(m-2), q = p/(p-1) = m/2$. The result is

$$\begin{aligned} &\|\phi(y'(\cdot),u) - \phi(y(\cdot),u)\|_{L^2(\Omega)} \\ &\leq K\left(\int_\Omega |y'(x) - y(x)|^{2m/(m-2)}\, dx\right)^{(m-2)/2m} \\ &\qquad \times \left(\int_\Omega (1 + |y(x)|^{\alpha-1} + |y'(x)|^{\alpha-1})^m\, dx\right)^{1/m}. \end{aligned}$$

Using the inequality $(a+b+c)^\beta \leq 3^\beta(a^\beta + b^\beta + c^\beta)(a, b, c, \beta \geq 0)$ and noting that $(\alpha - 1)m = 2m/(m-2)$, we obtain

$$\begin{aligned} &\|\phi(y'(\cdot),u) - \phi(y(\cdot),u)\|_{L^2(\Omega)} \\ &\qquad \leq K'(1 + \|y(\cdot)\|_{L^{2m/(m-2)}(\Omega)} + \|y'(\cdot)\|_{L^{2m/(m-2)}(\Omega)})^{2/(m-2)} \\ &\qquad\qquad \times \|y'(\cdot) - y(\cdot)\|_{L^{2m/(m-2)}(\Omega)} \\ &\qquad \leq K'(1 + \|y(\cdot)\|_{H^1(\Omega)} + \|y'(\cdot)\|_{H^1(\Omega)})^{2/(m-2)} \|y'(\cdot) - y(\cdot)\|_{H^1(\Omega)}. \end{aligned} \tag{6.12}$$

We check that $\mathbf{f}((y(\cdot), y_1(\cdot)), u) = \mathbf{f}(y(\cdot), u) = (0, \phi(y(\cdot), u)$ satisfies assumption (IV). Obviously, (a) of (VI) is (a) of (IV). The local Lipschitz condition (6.12) implies that $t \to \mathbf{f}(y, \cdot)$ is a continuous $BC(U)$-valued function of y, so that part (b) of (IV) is satisfied to excess. Finally, (6.10) and (6.12) imply the two inequalities in (c) of (IV),

$$\|\mathbf{f}((y, y_1), u)\| \le C\|(y, y_1)\|_E, \qquad (y, y_1) \in E,$$
$$\|\mathbf{f}((y', y_1'), u) - \mathbf{f}((y, y_1), u)\|_E$$
$$\le K(1 + \|(y, y_1)\|_E + \|(y', y_1')\|_E)^{2/(m-2)}\|(y', y_1') - (y, y_1)\|_E.$$

For $m = 2$, we have the imbedding $H^1(\Omega) \to L^q(\Omega)$ for every $q \ge 1$ [1, p. 97] and the arguments are essentially the same; for $m = 1$ the imbedding is $H^1(\Omega) \to C(\overline{\Omega})$. We omit the details.

Finally, the global existence-boundedness property (V) in §5 can be deduced from the present hypotheses using as a priori energy bounds: these take the form

$$\|y(t, \mu_n)\|^2_{H^1(\Omega)} + \|y_t(t, \mu_n)\|^2_{L^2(\Omega)} \le C \qquad (0 \le t \le T'), \tag{6.13}$$

where $[0, T']$ is any interval where the solution $y(t, \mu)$ exists. For details, see [32].

We consider below an optimal control problem for the nonlinear wave equation (6.1). The cost functional is arbitrary and only assumed to satisfy condition (III_0). The conditions on the nonlinear term will have to be slightly stiffened only for $m > 2$.

THEOREM 6.1. *Assume ϕ satisfies assumption* (VI) *with $\alpha < m/(m-2)$ in case $m > 2$. Let $\{\mu_n(\cdot)\}$ be a minimizing sequence with $\{t_n\}$ bounded. Then there exists a relaxed solution $\mu(\cdot)$ to the optimal control problem.*

Proof. The Rellich–Kondrachev theorem [1, p. 146] implies that the imbedding $H^1(\Omega) \to L^{2m/(m-2)}(\Omega)$ is compact. This and (6.13) justify the application of the Arzelà–Ascoli theorem below. Selecting if necessary a subsequence we may assume $\{y(\cdot, \mu_n)\}$ is uniformly convergent in $C(0, T; L^{2m/(m-2)}(\Omega))$ to $y(\cdot) \in C(0, T; L^{2m/(m-2)}(\Omega))$. Since $\phi(y(\cdot), u)$ is a locally Lipschitz operator from $L^{2m/(m-2)}(\Omega)$ into $L^2(\Omega)$ uniformly with respect to u, it follows that

$$f(y(t, \mu_n), u) \to f(y(t), u)$$

uniformly with respect to $u \in U$ and to $t \in [0, T]$. This is in excess of what is required in Theorem 4.1 and thus ends the proof.

See [16] for the treatment of other (variational) boundary conditions, addition of lower order terms, etc.

7. Abstract parabolic equations in nonreflexive spaces. We outline in this section a theory of relaxed systems that does not require the space E to be reflexive. This theory is also expounded in [21, §6] (with different objectives in mind); thus we only sketch the main facts here.

Let E be a Banach space, $S(t)$ a strongly continuous semigroup in E, and $E^\odot \subseteq E^*$ the closure of the domain of $D(A^*)$ in E^* or, equivalently, the maximal subspace where the semigroup $S(t)^*$ is strongly continuous. The restriction of the semigroup $S(t)^*$ to $E^\odot$ is a strongly continuous semigroup $S^\odot(t)$ called the *Phillips adjoint* of $S(t)$ [30, Chap. 14] The infinitesimal generator of $S^\odot(t)$ is called $A^\odot$ and is the restriction of A^* with domain $D(A^\odot) = \{y \in D(A^*); A^*y \in E^\odot\}$.

The norm $\|y\|_0 = \{\sup\langle y^*, y\rangle; y^* \in E^\odot, \|y^*\| \le 1\}$ in E is equivalent to the original norm of E; precisely, $\|y\|_0 \le \|y\| \le M\|y\|_0$ where $M = \liminf_{\lambda\to\infty} \|\lambda R(\lambda; A)\| < \infty$ [30 p. 423]. It follows that the canonical pairing of E and $E^\odot$ produces a bicontinuous

linear imbedding of E into $(E^{\odot})^*$. In fact, the imbedding is into $(E^{\odot})^{\odot} = E^{\odot\odot}$ and we have $A \subseteq (A^{\odot})^{\odot} = A^{\odot\odot}$ [30, p. 430].

Among the motivations for the theory of Phillips adjoints is the study of uniformly elliptic partial differential operators

$$Ay(x) = \sum_{j=1}^{m}\sum_{k=1}^{m} \partial^j(a_{jk}(x)\partial^k y(x)) + \sum_{j=1}^{m} b_j(x)\partial^j y(x) + c(x)y(x) \tag{7.1}$$

in a bounded domain Ω of class $C^{(2)}$, associated with either the Dirichlet boundary condition $y(x) = 0$ or a variational boundary condition $\partial^\nu y(x) = \gamma(x)y(x)$ on the boundary Γ with γ continuously differentiable, where ∂^ν denotes the *conormal derivative* $\partial^\nu = \Sigma\Sigma a_{jk}\partial^j\eta_k$ (with $(\eta_1, \ldots, \eta_m)$ the outer normal vector on Γ). The a_{jk} and the b_j are continuously differentiable and c is continuous. The *formal adjoint* A' of A is $A'y = \sum\sum \partial^j(a_{jk}(x)\partial^k y) - \sum \partial^j(b_j(x)y) + c(x)y$ and the *adjoint boundary condition* β' is $\beta' = \beta$ if β is the Dirichlet boundary condition; for a variational boundary condition, β' is $\partial^\nu y(x) = (\gamma(x) + b(x))y(x)$ with $b(x) = \Sigma b_j(x)\eta_j$.

The nonreflexive spaces of interest (the first in diffusion processes, the second in heat propagation) are $L^1(\Omega)$ and the space $C(\overline{\Omega})$ of continuous functions in $\overline{\Omega}$ endowed with the supremum norm. Given a boundary condition β, the operator A admits an extension $A_1(\beta)$ in $E = L^1(\Omega)$ that generates a compact analytic semigroup $S_1(t; A, \beta)$. The domain $D(A_1(\beta))$ of this extension can be characterized as follows: $D(A_1(\beta))$ consists of all elements $y \in L^1(\Omega)$ such that there exists $z(= A_1(\beta)y)$ in $L^1(\Omega)$ with

$$\int_\Omega y(x)(A'(\beta')v)(x)dx = \int_\Omega z(x)v(x)dx$$

for every $v \in C^{(2)}(\overline{\Omega})_\beta$, where $C^{(2)}(\overline{\Omega})$ is the space of all twice continuously differentiable functions defined in $\overline{\Omega}$ and $C^{(2)}(\overline{\Omega})_\beta$ is the subspace of $C^{(2)}(\overline{\Omega})$ consisting of all functions y that satisfy the boundary condition β on Γ.

When β is a condition of variational type A admits an extension $A_C(\beta)$ to the space $E = C(\overline{\Omega})$ which generates a compact analytic semigroup $S_C(t; A, \beta)$. The domain $D(A_C(\beta))$ can be characterized as

$$D(A_C) = \left\{ y \in \bigcap_{p\geq 1} W^{2,p}(\Omega)_\beta;\, Ay \in C(\overline{\Omega}) \right\},$$

where $W^{2,p}(\Omega)_\beta$ is the subspace of $W^{2,p}(\Omega)$ consisting of all $y(\cdot)$ that satisfy the boundary condition β on Γ. The same considerations apply when β is the Dirichlet boundary condition, but the space E is $C_0(\overline{\Omega})$, consisting of all $y \in C(\overline{\Omega})$ that vanish at the boundary Γ.

The duality theory of these spaces, operators, and semigroups is as follows: $L_1(\Omega)^* = L^\infty(\Omega), L^1(\Omega)^{\odot} = C(\overline{\Omega})$ for a variational boundary condition, $L^1(\Omega)^{\odot} = C_0(\Omega)$ for the Dirichlet boundary condition, $S_1(t; A, \beta)^{\odot} = S_C(t; A', \beta')$. The dual $C(\overline{\Omega})^*$ can be identified with the space $\Sigma(\overline{\Omega})$ consisting of all finite Borel measures μ defined in $\overline{\Omega}$ acting on elements $y(\cdot) \in C(\overline{\Omega})$ in the form $\langle \mu, y\rangle = \int_\Omega y(x)\mu(dx)$ and endowed with the total variation norm $\|\mu\| = \int_\Omega |\mu(dx)|$. The dual $C_0(\overline{\Omega})^*$ can be identified with the subspace $\Sigma_0(\overline{\Omega})$ of $\Sigma(\overline{\Omega})$ consisting of all μ vanishing on Γ. For the variational boundary condition β we have $C(\overline{\Omega})^{\odot} = L^1(\Omega)$, and $C_0(\overline{\Omega})^{\odot} = L^1(\Omega)$ for the Dirichlet boundary condition; in both cases, $S_C(t; A, \beta)^{\odot} = S_1(t; A', \beta')$.

We consider a control system

$$y'(t) = Ay(t) + f(t, y(t), u(t)), \qquad y(0) = \zeta \tag{7.2}$$

in an arbitrary Banach space E satisfying

(i) $S(t)E \subseteq D(A)$ and $AS(t)$ is continuous in the uniform norm of operators in $t > 0$.

(ii) E and $E^{\odot}$ are separable and E is $\odot$-*reflexive* with respect to $S(\cdot)$, that is, $E^{\odot\odot} = E$.

Assumption (ii) implies $S^{\odot\odot}(t) = S(t), A^{\odot\odot} = A$. It follows from the duality theory that $L^1(\Omega)$ is $^{\odot}$-reflexive with respect to $A_1(\beta)$, and that $C(\overline{\Omega}), C_0(\overline{\Omega})$ (the latter for Dirichlet boundary condition) are $^{\odot}$-reflexive with respect to $A_C(\beta)$. Moreover, both spaces are separable, and (i) is satisfied by the semigroups $S_1(t; A, \beta)$ and $S_C(t; A', \beta')$; thus the theory is applicable to this example. The space of relaxed controls is $V_r(0, T; U)$ as defined in §2 and $f(t, y, u)$ satisfies Assumption (II$^{\odot}$), which is the same as (II), but where (b) is required only for $y^* \in E^{\odot}$. The relaxed control system is defined as follows: $\mathbf{f}(t, y)\mu$ is the unique element of $(E^{\odot})^* \supseteq E$ that satisfies

$$\langle y^*, \mathbf{f}(t, y, \mu)\rangle = \int_U \langle y^*, f(t, y, u)\rangle \mu(du) \tag{7.3}$$

for every $y^* \in E^{\odot}$. As a result of the definition, the function $t \to \langle y^*, \mathbf{f}(t, y(t))\mu(t)\rangle$ is measurable for every $y^* \in E^{\odot}$ (that is, $t \to \mathbf{f}(t, y(t))\mu(t)$ is $E^{\odot}$-weakly measurable). The relaxed control system is

$$y'(t) = Ay(t) + \mathbf{f}(t, y(t))\mu(t), \qquad y(0) = \zeta. \tag{7.4}$$

Solutions of (7.4) take values in $E^{\odot\odot} = E$, but $\mathbf{f}(t, y(t))\mu(t)$ takes values in $(E^{\odot})^* \supseteq E^{\odot\odot} = E$. The corresponding integral equation is

$$y(t) = S(t)\zeta + \int_0^t S^{\odot}(t - \sigma)^* \mathbf{f}(t, y(\sigma))\mu(\sigma)d\sigma \tag{7.5}$$

if $\zeta \in E$; we may take the initial condition ζ in $(E^{\odot})^*$, in which case $S(t)\zeta$ is replaced by $S^{\odot}(t)^*\zeta$.

We note that, under (i) and (ii) we have

$$S(t)^*E^* \subseteq E^{\odot}, \qquad S^{\odot}(t)^*(E^{\odot})^* \subseteq E. \tag{7.6}$$

In fact, since $S(t)E \subseteq D(A)$ and $AS(t)$ is bounded for $t > 0$ we obtain, taking adjoints and using commutativity of A and $S(t)$, that $S(t)^*E^* \subseteq D(A^*) \subseteq E^{\odot}$. The same argument, this time applied in $E^{\odot}$ to the semigroup $S^{\odot}(\cdot)$, proves that $S^{\odot}(t)^*(E^{\odot})^* \subseteq E^{\odot\odot} = E$.

LEMMA 7.1. *Let $u(\cdot)$ be a $E^{\odot}$- weakly measurable $(E^{\odot})^*$-valued bounded function defined in $0 \le t \le T$. Then the function*

$$\sigma \to S^{\odot}(t - \sigma)^* u(\sigma) \tag{7.7}$$

is strongly measurable in $0 \le \sigma \le t$. (b) The E-valued function

$$y(t) = (\Lambda u)(t) = \int_0^t S^{\odot}(t - \sigma)^* u(\sigma)d\sigma \tag{7.8}$$

is continuous in $0 \le t \le T$.

For a proof, see [21, §6]. The theory of (7.4) is essentially the same as that of (2.15); note that, in the integral equation (7.5) that defines the solutions of (7.4), the integral is a continuous E-valued function by virtue of Corollary 7.2.

We limit ourselves to an obvious analog of Theorem 5.2. The assumption on $f(t,y,u)$ is (III$^{\odot}$), which is (III) in §5 with the difference that the element y^* now belongs to $E^{\odot}$ instead of E^*. The function f_0 satisfies (III$_0$).

THEOREM 7.2. *Let E be $^{\odot}$-reflexive with respect to a strongly continuous semigroup $S(t)$ satisfying* (i), *and assume that E and $E^{\odot}$ are separable. Assume that* (III) *and* (III$_0$) *hold and that there exists a minimizing sequence $\{\mu_n(\cdot)\}, \mu_n(\cdot) \in V_r(0,t_n;U)$ of relaxed controls with $\{t^n\}$ bounded and $\{y(\cdot,\mu_n)\}$ uniformly bounded in $0 \leq t \leq t_n$. Then there exists a solution $\overline{\mu}(\cdot)$ of the relaxed control problem.*

As Theorem 5.2, Theorem 7.2 depends on compactness of the integral operator on the right side of the integral equation (7.5) defining solutions of the equation (7.4). In this case, the operator is Λ in (7.8), and the compactness result is Lemma 7.3 below. In it, we denote by $L^1_w(0,T;(E^{\odot})^*)$ the space of all $E^{\odot}$-measurable $(E^{\odot})^*$-valued functions with integrable norm endowed with the L^1 norm (since $E^{\odot}$ is separable, the norm is measurable; see [14]).

LEMMA 7.3. *Assume $S(t)$ is compact for every $t > 0$. Then the operator $\Lambda : L^1_w(0,T;(E^{\odot})^*) \to C(0,T;E)$ satisfies the conclusions of Lemma* 5.1.

The proof is essentially the same as that of Lemma 5.1.

8. The maximum principle. Control systems of the form (2.15) or (7.4) essentially fit the model in [20] (where controls are only required to be weakly measurable), but a few modifications of the theory are necessary. To apply the nonlinear programming theory in [20] we equip the relaxed control space $V_r(0,T;U)$ with the metric

$$d(\mu(\cdot),\nu(\cdot)) = \lambda\{t \in [0,T]; \mu(t) \neq \nu(t)\}, \tag{8.1}$$

where λ is the outer measure generated by the Lebesgue measure in the real line and define as equivalent elements of $V_r(0,T;U)$ that lie at distance zero. This, however, requires explanation since $V_r(0,T;U)$ is already equipped with a different equivalence relation inherited from the space $L^{\infty}_w(0,T;\Sigma_{rba}(U))$. The precise definitions follow.

Let $F(0,T;X)$ be the space of all functions $f(\cdot), g(\cdot)$ defined in the interval $0 \leq t \leq T$ with values in an arbitrary set X. Denote by λ an arbitrary outer measure in $0 \leq t \leq T$. Two functions $f(\cdot), g(\cdot) \in F(0,T;X)$ are declared equivalent if $f(t) = g(t)$ except in a set e with $\lambda(e) = 0$; we denote by $F(0,T;X)_d$ the space of the equivalence classes corresponding to this equivalence relation. Since we have $\{t; f(t) \neq h(t)\} \subseteq \{t; f(t) \neq g(t)\} \cup \{t; g(t) \neq h(t)\}$ it follows that (8.1) depends only on the equivalence classes of $f(\cdot)$ and $g(\cdot)$ and defines a distance. The space $F(0,T;X)_d$ is a complete metric space equipped with d; Ekeland's original completeness proof [13] uses only the countable subadditivity property of the Lebesgue measure and thus generalizes to the present setting. In what follows, λ will be the outer measure generated by Lebesgue measure.

Given a Banach space E, $L^{\infty}_w(0,T;E^*)_0$ is the space of all E-weakly measurable E^*-valued functions $g(\cdot)$ with $|\langle y, g(t)\rangle| \leq C\|y\|$ almost everywhere in $0 \leq t \leq T$, *without* any equivalence relation. $L^{\infty}_w(0,T;\Sigma_{rba}(U))_d$ is the quotient of $L^{\infty}_w(0,T;\Sigma_{rba}(U))_0$ by the equivalence relation associated with the distance (8.1); two elements $\mu(\cdot), \nu(\cdot)$ of $L^{\infty}_w(0,T;\Sigma_{rba}(U))_0$ are equivalent if and only if $\{t; \mu(t) \neq \nu(t)\}$ has outer measure zero. This equivalence relation is more demanding than that of the space $L^{\infty}_w(0,T;\Sigma_{rba}(U))$, which is as follows: $\mu(\cdot)$ and $\nu(\cdot)$ are equivalent if and only if $\langle y, \mu(t)\rangle = \langle y, \nu(t)\rangle$ outside of a null set (depending on y) for every $y \in BC(U)$. The two equivalent relations are not the same. The space $V(0,T;U)_0$ is the subspace of $L^{\infty}_w(0,T;\Sigma_{rba}(U))_0$ defined by (2.7)–(2.9), and, for the present purposes, the space of (relaxed) controls is $V_r(0,T;U)_d$, quotient of $V(0,T;U)_0$ by the equivalence relation associated with the distance (8.1). Obviously,

(possibly different) controls in $V_r(0,T;U)_d$ that are equivalent in the equivalence relation of $L_w^\infty(0,T;\Sigma_{rba}(U))$ will produce the same trajectory.

LEMMA 8.1. *$V_r(0,T;U)_d$ is complete under d.*

Proof. $V_r(0,T;U)_d$ is a subspace of $F(0,T;\Sigma_{rba}(U))_d$, thus we only have to show that if $\{\mu_n(\cdot)\}$ is a sequence in $V_r(0,T;U)_d$ and $f(\cdot) \in F(0,T;\Sigma_{rba}(U))_d$ is such that $d(\mu_n,f) = \lambda\{t;\mu_n(t) \neq f(t)\} \to 0$, then $f(\cdot) \in V_r(0,T;U)_d$. This is plain, since $d(\mu_n,f) \to 0$ implies that $f(\cdot)$ is $BC(U)$-weakly measurable and the three conditions (2.6), (2.7), (2.8) defining an element of $V_r(0,T;U)$ are preserved through d-convergence.

Remark 8.2. If $f(\cdot), g(\cdot) \in L_w^\infty(0,T;E^*)$ with E separable, the set $\{t;f(t) \neq g(t)\}$, being the union of all sets $\{t;\langle e_n,f(t)\rangle \neq \langle e_n,g(t)\rangle\}(\{e_n\}$ a sequence dense in $E)$ is measurable. However, this is essentially irrelevant here since $E = BC(U)$ is not separable unless U is compact. Hence we need to use an outer measure in (8.1).

Pontryagin's maximum principle for the relaxed control system (2.15) will be obtained by application of the theory of the nonlinear programming problem

$$\text{minimize } f_0(\mu) \tag{8.2}$$

$$\text{subject to } f(\mu) \in Y, \tag{8.3}$$

where $f : V \to E$ (V a complete metric space, E a Banach space) and $F_0 : V \to \mathbb{R}$. For fixed terminal time $\bar{t}$, we choose $V = V_r(0,\bar{t};U)_d$ and

$$f(\mu) = y(\bar{t},\mu), \qquad f_0(\mu) = y_0(\bar{t},\mu), \tag{8.4}$$

where $y(t,\mu)$ is the trajectory of (2.14) corresponding to a control $\mu \in V_r(0,T;U)_d$ and $y_0(t,u)$ is the cost functional. The optimal relaxed control $\overline{\mu}$ is assumed to exist or is constructed by means of one of the existence theorems in §5 and 6. The first difficulty is that $y(t,\mu)$ (thus $f(\mu)$, a fortiori $f_0(\mu)$) may not be defined in $0 \le t \le \bar{t}$ if $\mu \neq \overline{\mu}$. Without aiming for maximum generality, we place conditions on E, U, and f that will legitimize all computations that follow. We assume that E is reflexive and separable and that U is normal; $f(t,y,u)$ satisfies (VII) below.

(VII) (a) $f(t,y,u)$ is continuous in $[0,T] \times E \times U$ uniformly with respect to u, and for every $c > 0$ there exists $\alpha(\cdot) = \alpha(c,\cdot) \in L^1(0,T)$ such that

$$\|f(t,y,u)\|_E \le \alpha(t) \qquad (0 \le t \le T, \|y\| \le c, u \in U).$$

(b) $f(t,y,u)$ has a Fréchet derivative $\partial_y f(t,y,u)$ with respect to y in $[0,T] \times E \times U$ uniformly with respect to u, i.e.,

$$f(t,y+h,u) = f(t,y,u) + \partial_y f(t,y,u)h + \rho(t,y,h,u), \tag{8.5}$$

where for each t,y we have $\|\rho(t,y,h,u)\|/\|h\| \to 0$ as $h \to 0$ uniformly with respect to $u \in U$. (c) $\partial_y f(t,y,u)$ is strongly continuous in $[0,T] \times E \times U$ uniformly with respect to u and for every $c > 0$ there exists $\beta(\cdot) = \beta(c,\cdot)$ such that

$$\|\partial_y f(t,y,u)\|_{L(E,E)} \le \beta(t) \qquad (0 \le t \le T, \|y\| \le B, u \in U). \tag{8.6}$$

These properties of $f(t,y,u)$ imply the following properties of the function $\mathbf{f}(t,y)\mu$ defined in §2: (a') $\mathbf{f}(t,y)\mu$ is continuous in $[0,T] \times E \times \Sigma_{rba}(U)$ (in fact, Lipschitz continuous with respect to μ by linearity) and

$$\|\mathbf{f}(t,y)\mu\| \le \alpha(t)\|\mu\|_{\Sigma_{rba}(U)} \qquad (0 \le t \le T, \|y\| \le B, \mu \in \Sigma_{rba}(U)). \tag{8.7}$$

Moreover, the function $t \to \mathbf{f}(t, y(t))\mu(t)$ is strongly measurable for every $y(\cdot) \in C(0, T; E)$ and $\mu(\cdot) \in L^\infty_w(0, T; \Sigma_{rba}(U))$. ($b'$) $\mathbf{f}(t, y)\mu$ has a Fréchet derivative $\partial_y \mathbf{f}(t, y)\mu$ with respect to y given by

$$\langle y^*, (\partial_y \mathbf{f}(t, y)\mu)h\rangle = \int_U \langle y^*, \partial_y f(t, y, u)h\rangle \mu(du) \qquad (y^* \in E^*). \tag{8.8}$$

Moreover, $\partial_y \mathbf{f}(t, y)\mu$ is strongly continuous in $[0, T] \times E \times \Sigma_{rba}(U)$ and satisfies

$$\|\partial_y \mathbf{f}(t, y)\mu\|_{L(E,E)} \le \beta(t)\|\mu\|_{\Sigma_{rba}(U)} \qquad (0 \le t \le T, \|y\| \le c, \mu \in \Sigma_{rba}(U)). \tag{8.9}$$

Existence of the Fréchet derivative and the mean value theorem imply the local Lipschitz condition

$$\begin{gathered}\|\mathbf{f}(t, y')\mu - \mathbf{f}(t, y)\mu\| \le \beta(t)\|y' - y\|\|\mu\|_{\Sigma_{rba}(U)} \\ (0 \le t \le T, \|y\| \le c, \mu \in \Sigma_{rba}(U))\end{gathered} \tag{8.10}$$

and the following result.

LEMMA 8.3. *Let $\overline{\mu}(\cdot) \in L^\infty_w(0, \bar{t}; \Sigma_{rba}(U))_0$ be such that the trajectory $y(t, \overline{\mu})$ exists in $0 \le t \le \bar{t}$. Then there exists $\rho > 0$ such that if $\mu(\cdot) \in L^\infty_w(0, \bar{t}; \Sigma_{rba}(U))_0, d(\mu, \overline{\mu}) \le \rho$ then the trajectory $y(t, \mu)$ exists in the same interval. If $d(\nu, \overline{\mu}) \le \rho$ as well, we have*

$$\|y(t, \mu) - y(t, \nu)\| \le C \int_{[\{t \in [0,\bar{t}]\mu(t) \ne \nu(t)\}]} \alpha(\sigma)d\sigma. \tag{8.11}$$

The notation in (8.11) is: given a set e, $[e]$ is a *measurable envelope* of e (that is, a measurable set with $e \subseteq [e], \lambda(e) = \text{meas}[e]$). The proof of Lemma 8.3 is similar to that of Lemma 5.1 in [20] (with the only minor difference that the set $\{t; \mu(t) \ne \nu(t)\}$ is measurable in [20]) thus we only sketch it. Local existence of $y(t, \mu)$ is guaranteed by (8.7) and (8.10). Let $[0, t_\mu]$ be the maximal interval where $y(t, \mu)$ exists and satisfies $\|y(t, \mu) - y(t, \overline{\mu}\| \le 1$. We have

$$\begin{aligned} y(t, \mu) - y(t, \overline{\mu}) &\\ &= \int_0^t S(t - \sigma)\{\mathbf{f}(\sigma, y(\sigma, \mu))\mu(\sigma) - \mathbf{f}(\sigma, y(\sigma, \mu))\overline{\mu}(\sigma)\}d\sigma \\ &+ \int_0^t S(t - \sigma)\{\mathbf{f}(\sigma, y(\sigma, \mu))\overline{\mu}(\sigma) - \mathbf{f}(\sigma, y(\sigma, \overline{\mu}))\overline{\mu}(\sigma)\}d\sigma. \end{aligned}$$

Estimating,

$$\|y(t, \mu) - y(t, \overline{\mu})\| \le C \int_{[\{t \in [0,\bar{t}]\mu(t) \ne \overline{\mu}(t)\}]} \alpha(\sigma)d\sigma + C \int_0^t \beta(\sigma)\|y(\sigma, \mu) - y(\sigma, \overline{\mu})\|d\sigma.$$

Using Gronwall's inequality, (8.11) for μ and $\overline{\mu}$ (with another constant) results in $[0, t_\mu]$. Taking $d(\mu, \overline{\mu})$ sufficiently small, $\|y(t, \mu) - y(t, \overline{\mu})\| < 1$ in $[0, t_\mu]$, which contradicts the maximality of this interval unless $t_\mu = \bar{t}$. A similar argument deals with the pair μ, ν.

The companion of assumption (VI) for f is as follows.

(VI$_0$) (a) $f_0(t, y, u)$ is continuous in $[0, T] \times E \times U$ uniformly with respect to u, and for every $c > 0$ there exists $\alpha_0(\cdot) = \alpha_0(c, \cdot) \in L^1(0, T)$ such that

$$\|f(t, y, u)\|_E \le \alpha_0(t) \qquad (0 \le t \le T, \|y\| \le c, u \in U).$$

(b) $f_0(t,y,u)$ has a Fréchet derivative $\partial_y f_0(t,y,u)$ with respect to y in $[0,T]\times E\times U$ uniformly with respect to u. (c) $\partial_y f_0(t,y,u)$ is continuous in $[0,T]\times E\times U$ uniformly with respect to u and for every $c>0$ there exists $\beta_0(\cdot)=\beta_0(c,\cdot)$ such that

$$\|\partial_y f_0(t,y,u)\|_{E^*}\le \beta_0(t) \qquad (0\le t\le T, \|y\|\le c, u\in U). \tag{8.12}$$

Assumption (VI_0) implies the following properties of $\mathbf{f}_0$: (a′) $\mathbf{f}_0(t,y)\mu$ is continuous in $[0,T]\times E\times \Sigma_{rba}(U)$ (Lipschitz continuous with respect to μ) and

$$|\mathbf{f}_0(t,y)\mu|\le \alpha_0(t)\|\mu\|_{\Sigma_{rba}(U)} \qquad (0\le t\le T, \|y\|\le C, \mu\in\Sigma_{rba}(U)); \tag{8.13}$$

moreover, the function $t\to \mathbf{f}_0(t,y(t))\mu(t)$ is measurable for every $y(\cdot)\in C(0,T;E)$ and $\mu\in L^\infty_w(0,T;\Sigma_{rba}(U))_0$. (b′) $\mathbf{f}_0(t,y)\mu$ has a Fréchet derivative $\partial_y\mathbf{f}_0(t,y)\mu$ given by

$$(\partial_y\mathbf{f}_0(t,y)\mu)h=\int_U \partial_y f_0(t,y,u)h\mu(du). \tag{8.14}$$

Moreover, $\partial_y\mathbf{f}_0(t,y)\mu$ is continuous in $[0,T]\times E\times\Sigma_{rba}(U)$ and

$$\|\partial_y\mathbf{f}_0(t,y)\mu\|_{E^*}\le\beta_0(t)\|\mu\|_{\Sigma_{rba}(U)} \qquad (0\le t\le T,\|y\|\le C,\mu\in\Sigma_{rba}(U)). \tag{8.15}$$

As a very particular consequence of the properties of $\mathbf{f}_0$, we obtain using Lemma 8.3 that the function $\mu\to y_0(\bar t,\mu)$ is continuous in $B(\overline{\mu},\rho)$, the ball of center $\overline{\mu}$ and radius ρ in $V_r(0,\bar t;U)_d$.

The abstract nonlinear programming theory in [20] will be applied to the functions (8.4) but in the closed ball $V=B(\overline{\mu},\rho)\subseteq V_r(0,T;U)_d$, where $\overline{\mu}$ is the optimal control and ρ is the constant in Lemma 8.3. Inequality (8.11) and the above comments on f_0 guarantee that f, f_0 are continuous, which is much more than necessary. The theory in [20] provides a Kuhn–Tucker multiplier $(z_0,z)\in\mathbb{R}\times E^*$ with $z_0\ge 0$ and such that

$$z_0\xi_0(\bar t,s,\nu,\overline{\mu})+\langle z,\xi(\bar t,s,\nu,\overline{\mu})\rangle\ge 0 \tag{8.16}$$

for all *spike variations* (ξ_0,ξ) defined as follows:

$$\xi(\bar t,s,\nu,\overline{\mu})=\lim_{h\to 0+}\frac{y(\bar t,\overline{\mu}_{h,s,\nu})-y(\bar t,\overline{\mu})}{h},$$

$$\xi_0(\bar t,s,\nu,\overline{\mu})=\lim_{h\to 0+}\frac{y_0(\bar t,\overline{\mu}_{h,s,\nu})-y_0(\bar t,\overline{\mu})}{h},$$

where $\mu_{h.s.\nu}$ denotes the *spike perturbation* of the relaxed control μ, depending on the parameters $h,s,\nu(h\ge 0, 0\le s\le\bar t,\nu\in\Sigma_{rba}(U))$ and defined by $\mu_{h,s,\nu}(t)=\nu$ $(s-h<t\le\bar t), \mu_{h,s,\nu}(t)=\mu(t)$ elsewhere. Calculation of the spike variations and subsequent computations are performed in essentially the same way as with ordinary controls [20], thus we omit the details and limit ourselves to stating the final result.

THEOREM 8.4. (Pontryagin's maximum principle). *Let $\overline{\mu}(\cdot)$ be a solution of the relaxed optimal control problem in $0\le t\le\bar t$. Then there exists $(z_0,z)\in\mathbb{R}\times E^*$ such that $z_0\le 0$ and*

$$\begin{aligned} &z_0\mathbf{f}_0(t,y(t,\overline{\mu}))\overline{\mu}(t)+\langle z(t,\overline{\mu}),\mathbf{f}(t,y(t,\overline{\mu}))\overline{\mu}(t)\rangle\\ &\quad=\max_{\nu\in\Sigma_{rba}(U)}\{z_0\mathbf{f}_0(t,y(t,\overline{\mu}))\nu+\langle z(t,\mu),\mathbf{f}(t,y(t,\overline{\mu}))\nu\rangle\}\end{aligned} \tag{8.17}$$

almost everywhere in $0 \leq t \leq \bar{t}$, where $z(t)$ is the solution of the final value problem

$$\begin{aligned} z'(t) &= -\{A^* + \partial_y \mathbf{f}(t, y(t,\overline{\mu}))\overline{\mu}(t))^*\} \\ &\quad -z_0 \partial_y \mathbf{f}_0(t, y(t,\overline{\mu}))\overline{\mu}(t), \qquad z(\bar{t}) = z \end{aligned} \tag{8.18}$$

in $0 \leq t \leq \bar{t}$.

As typical in infinite-dimensional problems, without further assumptions the multiplier (z_0, z) may be zero. We use the nontriviality condition in [20, Cor. 2.14], as follows.

LEMMA 8.5. *Assume that, for every sequence* $\{y^n\} \subseteq Y$ *such that* $y^n \to \overline{y} = y(\bar{t}, \overline{\mu})$ *and every sequence* $\{\mu_n\} \in V_r(0, \bar{t}; U)_d$ *with* $\mu_n \to \overline{\mu}$ *there exists a compact set* Q *such that*

$$\bigcap_{n=1}^{\infty} \{R_n(\bar{t}) - K_Y(y^n) + Q\}$$

contains an interior point, where $K_Y(y^n)$ *is the tangent cone to* Y *at* y^n *and* $R_n(\bar{t})$ *is the reachable space of the system*

$$\begin{aligned} z(t) &= \{A + \partial_y \mathbf{f}(t, y(t, \mu_n))\mu_n(t)\}z(t) \\ &\quad + \mathbf{f}_0(t, y(t, \mu_n))\nu(t) - \mathbf{f}_0(t, y(t, \mu_n))\mu_n(t), \qquad z(0) = 0, \end{aligned} \tag{8.19}$$

$\nu(\cdot) \in V_r(0, \bar{t}; U)_d$.

The time optimal problem needs a special treatment but the final result may be included in Theorem 8.4 and Lemma 8.5; the result guarantees a multiplier $(z_0, z) = (0, z)$ with $z \neq 0$.

The conditions of Lemma 8.5 are always satisfied if the target set Y is "large" (for instance, a ball). For small target sets (say, $Y = \{y\}$) they are satisfied by some hyperbolic systems: see [17]. They are also satisfied automatically when E is finite dimensional; we may take $Q =$ unit ball of E. For comments on this condition for abstract parabolic equations see [20, §6]. See also [20] for additional information on the vector z.

The maximum principle for relaxed controls can also be established for the control systems in nonreflexive spaces treated in §7. Under the assumption that E is $^{\odot}$- reflexive with respect to the semigroup $S(t)$ the control system (7.4), with $\mathbf{f}$ taking values in $(E^{\odot})^*$ and $\mu(\cdot) \in L^{\infty}_w(0, T; \Sigma_{rba}(U))$ and $L^1(0, T; BC(U))$-weakly measurable is precisely of the form considered in [20] in a somewhat different context, and the results there on the maximum principle apply without changes.

The maximum principle (8.17) can be easily translated in the language of differential inclusions using the equivalence results in §3. For a direct treatment of the maximum principle for finite-dimensional differential inclusions without using relaxed controls see [5] and [6].

Acknowledgment. I am grateful to Professor W. Rueβ for much useful information on the space $L^{\infty}_w(0, T; E^*)$ and on Theorems 2.1 and 2.2.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] N. U. AHMED, *Properties of relaxed trajectories for a class of nonlinear evolution equations in a Banach space*, SIAM J. Control Optimiz., 21 (1983), pp. 953–967.

[3] N. U. AHMED, *Existence of optimal controls for a class of systems governed by differential inclusions in Banach space*, J. Optimiz. Theory Appl., 50 (1986), pp. 213–237.

[4] J. M. Ball, *A version of the fundamental theorem for Young measures,* Lecture Notes in Physics 344, Springer-Verlag, New York, 1990, pp. 206–215.
[5] V. I. Blagodatskikh, *The maximum principle for differential inclusions*, Proc. Steklov Inst. Mat., (1986), pp. 23–43.
[6] V. I. Blagodatskikh and A. F. Filippov, *Differential inclusions and optimal control,* Proc. Steklov Inst. Mat., (1986), pp. 199–259.
[7] A. G. Chentsov, *Finitely additive measures and minimum problems* Kibernetika 3 (1988), pp. 67–70. [In Russian].
[8] F. Clarke, *Optimization and Nonsmooth Analysis,* Wiley-Interscience, New York, 1983.
[9] J. Dieudonné, *Sur le théorème de Lebesgue*–Nikodym (III), Ann. Université Grenoble, 23 (1947–48), pp. 25–53.
[10] ———, *Sur le théorème de Lebesgue*–Nikodym (IV), J Indian Math. Soc., 22 (1951), pp. 77–86.
[11] N. Dunford and B. J. Pettis, *Linear operators on summable functions,* Trans. Amer. Math. Soc., 47 (1940), pp. 323–392.
[12] N. Dunford and J. T. Schwartz, *Linear Operators,* part 1, Interscience, New York, 1958.
[13] I. Ekeland, *Nonconvex minimization problems,* Bull. Amer. Math. Soc., 1 (1979), pp. 443–474.
[14] H. O. Fattorini, *The time optimal control problem in Banach spaces,* Appl. Math. Optimiz. 1 (1974), pp. 163–188.
[15] H. O. Fattorini, *The Cauchy Problem,* Cambridge University Press, Cambridge, 1983.
[16] ———, *Second Order Linear Differential Equations in Banach Spaces,* North-Holland Mathematics Studies/Notas de Matemática 108, Amsterdam, 1985.
[17] ———, *A unified theory of necessary conditions for nonlinear nonconvex control systems,* Appl. Math. Optimiz., 15 (1987), pp. 141–185.
[18] ———, *Optimal control of nonlinear systems: convergence of suboptimal controls,* I, Operator Methods for Optimal Control Problems, S. J. Lee, ed., Lecture Notes in Pure and Applied Mathematics, 108, Marcel Dekker, New York, 1987, pp. 159–199.
[19] ———, *Relaxed controls in infinite dimensional systems,* Estimation and control of distributed parameter systems, W. Desch, F. Kappel, and K. Kunisch, eds., Int. Series Numer. Math. 100, Birkhäuser, Basel, 1991, pp. 115–128.
[20] ———, *Optimal control problems for distributed parameter systems in Banach spaces,* Appl. Math. Optim., to appear.
[21] ———, *Relaxation theorems, differential inclusions and Filippov's theorem for relaxed controls in semilinear infinte dimensional systems,* J. Differential Equations, to appear.
[22] ———, *Relaxation in infinite dimensional control systems,* L. M. Fertschrift, K. D. Elworthy, W. N. Everitt, and E. B. Lee, eds., Lecture Notes in Pure and Applied Math., Marcel Dekker, New York, to appear.
[23] H. O. Fattorini and H. Frankowska, *Necessary conditions for infinite dimensional control problems,* Math. Control, Signals, and Systems, 4 (1991), pp. 41–67.
[24] H. O. Fattorini and S. S. Sritharan, *Existence of optimal controls for viscous flow problems,* Proc. Royal Soc. London A (1992), pp. 91–102.
[25] H. O. Fattorini and S. S. Sritharan, *Optimal chattering controls for viscous flow,* to appear.
[26] A. F. Filippov, *On certain questions in the theory of optimal control,* Vestnik Moskov. Univ. Ser. Mat. Mech. Astronom., 2 (1959), pp. 25–32. English translation: SIAM J. Control, 1 (1962), pp. 76–84.
[27] H. Frankowska, *Some inverse mapping theorems,* Ann. Inst. Henri Poincaré, 7 (1990), pp. 183–234.
[28] H. Frankowska, *A priori estimates for operational differential inclusions,* J. Differential Equations, 84, (1990), pp. 100–128.
[29] R. V. Gamkrelidze, *On sliding optimal states,* Dokl. Acad. Nauk SSSR, 143 (1962), pp. 1243–1245.
[30] E. Hille and R. S. Phillips, *Functional Analysis and Semi-Groups,* Amer. Math. Soc., Providence, RI, 1957.
[31] A. Ionescu Tulcea and C. Ionescu Tulcea, *Topics in the Theory of Lifting,* Springer-Verlag, Berlin, 1969.
[32] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non-Linéaires,* Dunod, Paris 1969.
[33] N. Papageorgiou, *Properties of the relaxed trajectories of evolution equations and optimal control,* SIAM J. Control Optimiz., 27 (1989), pp. 267–288.
[34] J. Warga, *Relaxed variational problems,* J. Math. Anal. Appl., 4 (1962), pp. 111–128.
[35] ———, *Necessary conditions for minimum in relaxed variational problems,* J. Math. Anal. Appl, 4 (1962), pp. 129–145.
[36] ———, *Optimal control of differential and functional equations,* Academic Press, New York, 1971.
[37] L. C. Young, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations,* Comptes Rendus Soc. Sc. Lett. Varsovie, 30 (1937), pp. 212–234.
[38] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory,* W. B. Saunders, Philadelphia, PA, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 332–357, March 1994

002

REGULAR UNITARY DILATION OF COMMUTING CONTRACTIONS AND MARKOVIAN REPRESENTATION OF GAUSSIAN STATIONARY PROCESSES ON Z^{2*}

PHILIPPE LOUBATON†

Abstract. In this paper, we show that the concept of regular unitary dilation of a pair of commuting contractions is intimately related to Markovian-like subspaces with respect to a pair of commuting unitary operators. Starting from this observation, a two-dimensional Markovian-like representation problem previously introduced by Attasi is studied geometrically. By using the geometric properties of the state spaces and some elementary properties of two-variable Hardy spaces, the spectral domain description of regular and coregular minimal representations are given.

Key words. regular unitary dilation of commuting contractions, perpendicular intersection, Attasi model, Markovian representation, wide-sense stationary random processes on Z^2, Hardy spaces, inner functions

AMS subject classifications. 60G25, 47A20

1. Introduction.

It is well-known that there exists a deep connection between the notion of Markovian space with respect to a unitary operator and the minimal unitary dilation of a contractive operator. Recently [4], the links between these objects were exploited in order to get new results concerning the Markovian representation problem of vector-valued wide-sense stationary random processes. In this paper we establish that the concept of regular minimal unitary dilation of a pair of commuting contractions is itself intimately related to Markovian-like subspaces with respect to a pair of commuting unitary operators; applications to a Markovian-like representation problem of two-parameter wide-sense stationary random processes are given.

Let us recall that if K is a Hilbert space and if U is a unitary operator, then a (closed) subspace X of K is said to be Markovian with respect to U if the spaces $X_- = \bigvee_{n\leq 0} U^n X$ and $X_+ = \bigvee_{n\geq 0} U^n X$ are conditionally orthogonal given X (we say that two subspaces A and B are conditionally orthogonal given C, written $A \perp B|C$, if $A \bigvee C \ominus C \perp B \bigvee C \ominus C$). As it is well known, [4, Lemma 1-1], this condition is equivalent to saying that $(U_{|X_\infty}, X_\infty)$ (where $X_\infty = \bigvee_{n\in Z} U^n X$) is the regular unitary dilation of $(E^X U_{|X}, X)$, i.e., $(E^X U_{|X})^n = (E^X U^n_{|X})$ for each $n \geq 0$. The operator $T = E^X U_{|X}$ is called the Markovian transition operator associated to X. The space W_f (respectively, W_b) given by $W_f = UX_- \ominus X_-$ (respectively, $W_b = X_+ \ominus UX_+$) is a wandering subspace for U called the forward (respectively, backward) innovation space of X; moreover, $W_f = \overline{(U-T)X}$ (respectively, $W_b = \overline{(I - UT^*)X)}$, and $X_\infty = U^*(W_b)_- \oplus X \oplus (W_f)_+$. Conversely, the fact that every contraction T defined on the Hilbert space X admits a minimal unitary dilation U [25] implies that X can be considered as a Markovian subspace with respect to U and that T coincides with its Markovian transition operator.

Let us now consider a pair of commuting unitary operators (U_1, U_2) defined on the Hilbert space K. A subspace X of K is said to be a forward-forward Markovian space (FFMS) with respect to (U_1, U_2) if the following conditional orthogonality relations hold:

$$X_{-,\infty} \perp X_{+,0} | X, \tag{1}$$

$$X_{\infty,-} \perp X_{0,+} | X, \tag{2}$$

* Received by the editors October 10, 1990; accepted for publication (in revised form) September 9, 1992.
† Ecole Nationale Superieure des Telecommunications, Departement Signal, 46 rue Barrault 75634 Paris, Cedex 13, France.

where $X_{-,\infty}$ (respectively, $X_{\infty,-}$) and $X_{+,0}$ (respectively, $X_{0,+}$) are defined by $X_{-,\infty} = \bigvee_{m\leq 0}\bigvee_{n\in Z} U_1^m U_2^n X$ (respectively, $X_{\infty,-} = \bigvee_{m\in Z}\bigvee_{n\leq 0} U_1^m U_2^n X$) and by $X_{+,0} = \bigvee_{m\geq 0} U_1^m X$ (respectively, $X_{0,+} = \bigvee_{n\geq 0} U_2^n X$). X is said to be a backward-forward (BF), forward-backward (FB), backward-backward (BB) Markovian space with respect to (U_1, U_2) if X is a FFMS with respect to $(U_1^*, U_2), (U_1, U_2^*), (U_1^*, U_2^*)$ respectively; finally, X is said to be a Markovian space (MS) with respect to (U_1, U_2) if it is both a FFMS and a BBMS with respect to (U_1, U_2).

The above definitions do not correspond to those that are usually used in the literature concerning the Markov random fields (MRF). In fact, the qualificative Markovian used in (1) and (2) is rather excessive, in the sense that (1) and (2) are only connected to a restrictive class of Markov random fields. The relations (1) and (2) occur in the study of the so-called Attasi model [1] introduced in the frame of the realization theory of two-dimensional linear systems. A multivariate two-parameter wide-sense stationary process $(X_{m,n})_{(m,n)\in Z^2}$ is said to be a (stochastic) Attasi model if it is given by the following (forward-forward) state-space equation:

$$X_{m+1,n+1} = F_1 X_{m,n+1} + F_2 X_{m+1,n} - F_1 F_2 X_{m,n} + L\nu_{m,n}, \tag{3}$$

where $(\nu_{m,n})_{(m,n)\in Z^2}$ is a white noise sequence (i.e., $E(\nu_{m,n}\nu^*_{m',n'}) = I\delta_{m-m',n-n'}$, where E denotes the mathematical expectation in this context), and where F_1 and F_2 are two commuting stable matrices. Such a process is characterized by the fact [11] that

$$X_{m+1,n}/sp(X_{m-k,n-l}|k \geq 0, l \in Z) = X_{m+1,n}/X_{m,n}, \tag{4}$$

$$X_{m,n+1}/sp(X_{m-k,n-l}|k \in Z, l \geq 0) = X_{m,n+1}/X_{m,n} \tag{5}$$

(where sp(.) stands for space generated by), which is clearly equivalent to (1) and (2) in the case where X is the space generated by the components of $X_{0,0}$ and where U_1 and U_2 coincide with the horizontal and vertical shift operators associated to the stationary random process $(X_{m,n})_{(m,n)\in Z^2}$. The Attasi models also satisfy

$$\begin{aligned} &X_{m+1,n+1}/sp(X_{m-k,n-l}|k \geq 0 \text{ or } l \geq 0) \\ &\quad = X_{m+1,n+1}/sp(X_{m,n+1}, X_{m+1,n}, X_{m,n}). \end{aligned} \tag{6}$$

This property characterizes the quarter-plane Markov random fields introduced by Pickard [18] in the scalar-valued case, so that the Attasi models belong to the class of the quarter-plane MRF. But, the Pickard's random fields are considerably more general. In particular, any process $(X_{m,n})_{(m,n)\in Z^2}$ given by a state-space equation

$$X_{m+1,n+1} = A_1 X_{m,n+1} + A_2 X_{m+1,n} + A_3 X_{m,n} + B\nu_{m,n}, \tag{7}$$

where $(\nu_{m,n})_{(m,n)\in Z^2}$ is a white noise, and where (A_1, A_2, A_3) are matrices for which $\det(I - z_1^{-1}A_1 - z_2^{-1}A_2 - z_1^{-1}z_2^{-1}A_3) \neq 0$ for $|z_1| \geq 1$ and $|z_2| \geq 1$ is a quarter-plane MRF. Therefore, in view of possible applications to two-dimensional stochastic realization, it would be more interesting to study these kinds of models. However, by contrast with the case of Attasi models, the deterministic realization problem of two-dimensional linear systems by means of the deterministic counterpart of the quarter-plane MRF (i.e., the so-called Fornasini–Marchesini model [5]) does not give rise to a satisfying theory. In particular, the state-space and the corresponding matrices $(A_i)_{i=1,3}, B$ cannot be clearly extracted from the data of the realization problem. Hence, we believe that the stochastic realization problem in the class of quarter-plane MRF would give rise to considerably less

satisfying results than the realization problem by the Attasi models. In any case, as in the deterministic case, there may exists a strong connexion between the two realization problems; therefore, it is certainly useful to get a better understanding of the properties of the Attasi models in order to study the realization problem by means of the quarter-plane MRF. Finally, we mention that although quite restrictive, the Attasi models have shown to be useful in certain important applications such as two-dimensional harmonic retrieval, for example, [13].

In §2 of this paper we show that for every FFMS X with respect to a pair of commuting unitary operators (U_1, U_2) defined on the Hilbert space K, the contractions T_1 and T_2 given by $T_1 = E^X U_{1|X}$ and $T_2 = E^X U_{2|X}$ (called the horizontal and vertical Markovian transition operators of X) commute and that (U_1, U_2, K) is a regular unitary dilation of (T_1, T_2, X) [7], [25], i.e.,

$$T_1^m T_2^n = E^X U_1^m U_{2|X}^n, \tag{8}$$

$$T_2^{*n} T_1^m = E^X U_1^m U_{2|X}^{*n} \tag{9}$$

for each $(m, n) \in N^2$ (T_2^* denotes the adjoint of T_2). Moreover, X is a MS with respect to (U_1, U_2) if and only if T_1 and T_2 doubly commute (i.e., $T_1 T_2^* = T_2^* T_1$). Conversely, let (T_1, T_2) be two commuting contractions defined on X; then, it is well known [7], [25] that (T_1, T_2, X) admits a unique minimal regular unitary dilation (U_1, U_2, K) (minimal in the sense that $K = \bigvee_{(m,n)\in Z^2} U_1^m U_2^n X$) if and only if the operator $I - T_1^* T_1 - T_2^* T_2 + T_1^* T_2^* T_2 T_1$ is positive. In this case, a remarkable fact is that X is a FFMS with respect to (U_1, U_2) (cf. Theorem 2.1); if, moreover, T_1 and T_2 doubly commute, then K is an MS with respect to (U_1, U_2). This result was given in [15], but with an incorrect proof. We take advantage of this property in order to study the structure of the regular minimal unitary dilation of a pair (T_1, T_2) of commuting contractions defined on a Hilbert space X. We demonstrate in a very easy way some known results due to Halperin [7]; we present some unknown properties which are of special interest for the study of the Markovian-like representation problem considered in §3 and which make more comprehensible some of the results of Slocinski [23] devoted to the case of a pair of doubly commuting contractions.

The last section of this paper is devoted to the Markovian-like representation problem of two-parameter wide-sense stationary random processes studied by Attasi [1]. By contrast with [1], our formulation is purely geometrical; it can be seen as a generalization of the attractive approach introduced by Lindquist–Picci and Ruckebusch (see [14] and [19], respectively, for a survey of the works of these authors). Let K be a Hilbert space, and let (U_1, U_2) be a pair of commuting unitary operators defined on K. Then, if Y is a subspace of K, the problem we treat consists in the characterization of the FFMS (respectively, the MS) with respect to (U_1, U_2) containing Y; such spaces will be called forward-forward Markovian representations (FFMR in short) (respectively, Markovian representations (MR)) of Y. Although it is possible to describe geometrically the set of all (FF)MR of Y (Theorems 3.1, 3.2), the representation problem considered here leads to considerably less powerful results than those obtained by Lindquist–Picci and Ruckebusch. This is essentially due to the fact that the existence of nontrivial (FF)MR is not guaranteed in the general case (see Propositions 3.2 and 3.3); in particular, there do not exist any "canonical" (FF)MR similar to the filter and the cofilter of the one-parameter theory (i.e., the spaces $\overline{E}^{Y_-}(Y_+)$ and $\overline{E}^{Y_+}(Y_-)$). Therefore, we study a very restrictive situation; namely, we consider the case where Y is a one-dimensional subspace for which the spectral measure of the two-parameter stationary sequence $y_{m,n} = U_1^m U_2^n y$ (where y is an element of Y) has properties

that guarantee the existence of nontrivial MR. Some results concerning the structure of the minimal representations (in the sense that they do not contain other representation as proper subspace) are obtained by using the "spectral" domain description of the regular and co-regular FFMR and MR; their derivation uses extensively some of the ideas of Lindquist and Picci. Finally, let us mention that their generalization to the case where $\dim Y$ is greater than 1 seems not to be trivial.

At this point, it is appropriate to introduce some notations and definitions. If K is a Hilbert space, we say briefly that H is a subspace of K if H is a closed vector subspace of K; in this case, E^H denotes the orthogonal projection operator onto H, and if G is a subspace of K, $\overline{E}^H(G)$ represents the closure of $E^H(G)$; if $z \in K$, the vector $E^H(z)$ will also be denoted z/H. If U is a unitary operator defined on K, then a subspace Δ is said to be wandering for U if $U^k\Delta \perp \Delta$ for $k \neq 0$. The notion of wandering subspace for a pair of commuting unitary operators is defined similarly. For a fixed pair (U_1, U_2) of commuting unitary operators defined on K, we use the following system of notation (no ambiguity will occur in the paper). Let X be a subspace of K; then, we denote by $X_{0,\infty}, X_{0,+}, X_{0,-}$ the spaces $\bigvee_{n\in Z} U_2^n X, \bigvee_{n\geq 0} U_2^n X, \bigvee_{n\leq 0} U_2^n X$, respectively, and by $X_{\infty,\infty}, X_{\infty,+}, X_{\infty,-}$ the spaces $\bigvee_{m\in Z}\bigvee_{n\in Z} U_1^m U_2^n X, \bigvee_{m\in Z}\bigvee_{n\geq 0} U_1^m U_2^n X, \bigvee_{m\in Z}\bigvee_{n\leq 0} U_1^m U_2^n X$. The spaces $X_{\infty,0}, X_{+,0}, X_{-,0}, X_{+,\infty}, X_{-,\infty}$ are defined similarly by exchanging the role of U_1 and U_2. Moreover, we put $X_{+,+} = \bigvee_{m\geq 0}\bigvee_{n\geq 0} U_1^m U_2^n X$; the definition of the spaces $X_{+,-}, X_{-,+}$ and $X_{-,-}$ is similar.

We finish by giving an important result due to Kallianpur and Mandrekar ([10]; see also [22]), which concerns the existence of a Wold-type decomposition for a pair of commuting isometries. For this purpose, let us recall that two subspaces H_1 and H_2 are said to intersect perpendicularly if $\overline{E}^{H_1}(H_2) = H_1 \cap H_2$, or equivalently if $E^{H_1}E^{H_2} = E^{H_2}E^{H_1}$.

PROPOSITION 1.1. *Let (U_1, U_2) be a pair of commuting unitary operators defined on the Hilbert space K, and let X be a subspace of K. Then, we have the following equivalence* ([10, *Thm.* 4.2]):

(i) *$X_{-,\infty}$ and $X_{\infty,-}$ intersect perpendicularly in $X_{-,-}$, that is,*

$$E^{X_{-,\infty}}E^{X_{\infty,-}} = E^{X_{\infty,-}}E^{X_{-,\infty}} = E^{X_{-,-}}; \tag{10}$$

(ii) *$U^*_{1|X_{-,-}}$ and $U^*_{2|X_{-,-}}$ are doubly commuting isometries.*

In this case, the wandering subspace Δ for (U_1, U_2) defined by

$$\Delta = (U_1 X_{-,\infty} \ominus X_{-,\infty}) \cap (U_2 X_{\infty,-} \ominus X_{\infty,-}) \tag{11}$$

is also given by

$$\Delta = (U_1U_2X_{-,-} \ominus U_1X_{-,-}) \cap (U_1U_2X_{-,-} \ominus U_2X_{-,-}) \tag{12}$$

or by

$$\begin{aligned}\Delta &= U_1(U_2X_{-,-} \ominus X_{-,-}) \ominus (U_2X_{-,-} \ominus X_{-,-})\\ &= U_2(U_1X_{-,-} \ominus X_{-,-}) \ominus (U_1X_{-,-} \ominus X_{-,-}).\end{aligned} \tag{13}$$

Moreover, $X_{-,-}$ has the following four-fold Wold-type decomposition:

$$\begin{aligned}X_{-,-} = U_1^*U_2^*\Delta_{-,-} \oplus \oplus_{k=1,\infty}U_1^{*k}(U_1X_{-,-\infty} \ominus X_{-,-\infty}) \oplus\\ \oplus_{l=1,\infty}U_2^{*l}(U_2X_{-\infty,-} \ominus X_{-\infty,-}) \oplus X_{-\infty,-\infty},\end{aligned} \tag{14}$$

where $X_{-,-\infty} = \cap_{n\in Z}U_2^nX_{-,-}$, $X_{-\infty,-} = \cap_{m\in Z}U_1^mX_{-,-}$, and $X_{-\infty,-\infty} = \cap_{(m,n)\in Z^2}$ $U_1^mU_2^nX_{-,-}$.

2. Minimal regular unitary dilation of a pair of commuting contractions and forward-forward Markovian spaces. In this section we make the connections between the regular minimal unitary dilation of a pair of commuting contractions and the notion of forward-forward Markovian space with respect to a pair of commuting unitary operators. We have the following result.

THEOREM 2.1. *Let (U_1, U_2) be a pair of commuting unitary operators defined on the Hilbert space K; let us assume that there exists a subspace X of K such that $K = X_{\infty,\infty}$. If X is a* FFMS *(respectively, a* MS*) with respect to (U_1, U_2), then the contractions $T_1 = E^X U_{1|X}$ and $T_2 = E^X U_{2|X}$ commute (respectively, doubly commute), and (U_1, U_2, K) is the minimal regular unitary dilation of $(E^X U_{1|X}, E^X U_{2|X}, X)$.*

Conversely, if (T_1, T_2) are two commuting contractions defined on the Hilbert space X for which there exists a regular minimal unitary dilation (U_1, U_2) defined on the space $K = X_{\infty,\infty}$, then X is a forward-forward Markovian space with respect to (U_1, U_2). In this case, T_1 and T_2 coincide with the Markovian transition operators associated to X. Moreover, if (T_1, T_2) doubly commute, then X is an MS *with respect to (U_1, U_2).*

Proof. Let us assume that X is a FFMS with respect to (U_1, U_2). Then, $X_{-,0} \perp X_{+,0} | X$, so that X is a Markovian subspace with respect to U_1, the Markovian transition of which is $T_1 = E^X U_{1|X}$. Therefore, $T_1^m x = E^X U_1^m x$ for all $m \in N$, and for all $x \in X$; but, by (1), $E^{X_{-,\infty}} U_1^m x$ belongs to X; this implies that $E^{X_{-,\infty}} U_1^m x = T_1^m x$; similarly, if we put $T_2 x = E^X U_{2|X}$, then, $E^{X_{\infty,-}} U_2^n x = T_2^n x$ for all $n \in N$ and for all $x \in X$. On the other hand, for each $(m, n) \in N^2$ and for each element x of X, $E^X U_1^m U_2^n x = E^X E^{X_{-,\infty}} U_1^m U_2^n x$; as $U_2 X_{-,\infty} = X_{-,\infty}$, $E^{X_{-,\infty}} U_1^m U_2^n x = U_2^n E^{X_{-,\infty}} U_1^m x = U_2^n T_1^m x$; therefore, $E^X U_1^m U_2^n x = T_2^n T_1^m x$. Similarly, $E^X U_1^m U_2^n x = T_1^m T_2^n x$; from this, it follows that $E^X U_{1|X}$ and $E^X U_{2|X}$ are two commuting contractions. Moreover, $E^X U_1^m U_2^{-n} x = E^X E^{X_{-,\infty}} U_1^m U_2^{-n} x = E^X U_2^{-n} E^{X_{-,\infty}} U_1^m x = T_2^{*n} T_1^m x$. Hence, (U_1, U_2, K) is the minimal regular unitary dilation of (T_1, T_2, X). Finally, if X is a MS with respect to (U_1, U_2), then all the preceding considerations hold if U_2 is replaced by U_2^*; from this, it is easily deduced that $T_1 = E^X U_{1|X}$ and $T_2^* = E^X U_{2|X}^*$ are commuting contractions; therefore, (T_1, T_2) are doubly commuting contractions.

Conversely, suppose that (U_1, U_2, K) is the minimal regular unitary dilation of (T_1, T_2, X). Let us show that (1) holds. For this purpose, it is sufficient to establish that

$$E^{X_{-,\infty}} U_1^m x = T_1^m x \tag{15}$$

for each integer $m \geq 0$ and for each element x of X. For all $k \in N$, for all $l \in Z$, for all $y \in X$, $\langle U_1^m x - T_1^m x, U_1^{-k} U_2^l y \rangle = \langle U_1^{m+k} U_2^{-l} x, y \rangle - \langle U_1^k U_2^{-l} T_1^m x, y \rangle = \langle E^X U_1^{m+k} U_2^{-l} x, y \rangle - \langle E^X U_1^k U_2^{-l} T_1^m x, y \rangle$. But, $E^X U_1^{m+k} U_2^{-l} x = E^X U_1^k U_2^{-l} T_1^m x = T_2^{*l} T_1^{m+k} x$ because (U_1, U_2, K) is the minimal regular unitary dilation of (T_1, T_2, X). Thus, $\langle U_1^m x - T_1^m x, U_1^{-k} U_2^l y \rangle = 0$. Consequently, $U_1^m x - T_1^m x$ is orthogonal to $X_{-,\infty}$; this implies (15). The proof of (2) is analogous. If T_1 and T_2 doubly commute, the fact that $X_{+,\infty} \perp X_{-,0} | X$ (respectively, $X_{\infty,+} \perp X_{0,-} | X$) is shown similarly by exchanging (T_1, T_2) by (T_1^*, T_2) (respectively, by (T_1, T_2^*)). □

Let us illustrate the converse of the theorem in the case where X is a finite dimensional subspace of centered complex-valued square integrable random variables. Let $X_{0,0} = (x_1, \ldots, x_N)^T$ be a basis of X, and let us denote by P the covariance matrix $E(X_{0,0} X_{0,0}^*)$. Let T_1 and T_2 be two commuting contractive operators defined on X admitting a regular unitary dilation, and let F_1 and F_2 be the two matrices defined by the fact that

$$(T_i x_1, \ldots, T_i x_n)^T = F_i X_{0,0} \tag{16}$$

for $i = 1, 2$. Clearly, the matrices of T_1 and T_2 in the basis $X_{0,0}$ coincide with F_1^T and F_2^T,

respectively. Then, it is easily seen that $(T_i^* x_1, \ldots, T_i^* x_n)^T = PF_i^* P^{-1} X_{0,0}$ from which it follows that the condition $I - T_1^* T_1 - T_2^* T_2 + T_1^* T_2^* T_2 T_1$ positive is equivalent to the positivity of the matrix $P - F_1 P F_1^* - F_2 P F_2^* + F_1 F_2 P F_2^* F_1^*$. Let (U_1, U_2) be the minimal regular unitary dilation of (T_1, T_2), and let us put $X_{m,n} = U_1^m U_2^n X_{0,0}$. Then, Theorem 2.1 states that $X_{m+1,n}/sp(X_{m-k,n-l}/k \geq 0, l \in Z) = X_{m+1,n}/X_{m,n} = F_1 X_{m,n}$, and that $X_{m,n+1}/sp(X_{m-k,n-l}/k \in Z, l \geq 0) = X_{m,n+1}/X_{m,n} = F_2 X_{m,n}$. This property is known [11] to be equivalent to the fact that $(X_{m,n})_{(m,n)\in Z^2}$ is an Attasi model given by a state-space equation $X_{m+1,n+1} = F_1 X_{m,n+1} + F_2 X_{m+1,n} - F_1 F_2 X_{m,n} + L\nu_{m,n}$ where ν is a white noise sequence. In particular, the positive matrix $P - F_1 P F_1^* - F_2 P F_2^* + F_1 F_2 P F_2^* F_1^*$ is equal to LL^*.

Throughout this section, (T_1, T_2) will denote a fixed pair of commuting contractions defined on the Hilbert space X, which admits a minimal regular unitary dilation (U_1, U_2). We are going to present some useful properties of (U_1, U_2). Although the structure of (U_1, U_2) was studied by Halperin [7], most of the results to be presented in this section are new. We begin by giving some properties of the spaces $X_{-,\infty}, X_{+,\infty}, X_{\infty,-}$ and $X_{\infty,+}$. First, it follows from (1) that the space $X_{0,\infty}$ is Markovian with respect to U_1, i.e., that $X_{-,\infty} \perp X_{+,\infty} | X_{0,\infty}$. In order to establish this property, it is sufficient to show that for all $x \in X, E^{X_{-,\infty}} U_1^k U_2^l x$ belongs to $X_{0,\infty}$ for all $k \geq 0$, for all $l \in Z$; but, this derives from the fact that $E^{X_{-,\infty}} U_1^k U_2^l x = U_2^l E^{X_{-,\infty}} U_1^k x = U_2^l T_1^k x$. Therefore, the spaces $X_{-,\infty}$ and $X_{+,\infty}$ intersect perpendicularly in $X_{0,\infty}$. Likewise, $X_{\infty,0}$ is Markovian with respect to U_2, or equivalently, $X_{\infty,-}$ and $X_{\infty,+}$ intersect perpendicularly in $X_{\infty,0}$. In the sequel, we shall denote by S_1 and S_2 the Markovian transition operators of $X_{0,\infty}$ and $X_{\infty,0}$, i.e.,

$$S_1 = E^{X_{0,\infty}} U_{1|X_{0,\infty}}, \tag{17}$$

$$S_2 = E^{X_{\infty,0}} U_{2|X_{\infty,0}}. \tag{18}$$

On the other hand, the following relations hold.

PROPOSITION 2.1.

$$E^{X_{-,\infty}} E^{X_{\infty,-}} = E^{X_{\infty,-}} E^{X_{-,\infty}} = E^{X_{-,-}}, \tag{19}$$

$$E^{X_{-,\infty}} E^{X_{\infty,+}} = E^{X_{\infty,+}} E^{X_{-,\infty}} = E^{X_{-,+}}, \tag{20}$$

$$E^{X_{+,\infty}} E^{X_{\infty,-}} = E^{X_{\infty,-}} E^{X_{+,\infty}} = E^{X_{+,-}}, \tag{21}$$

Moreover,

$$X_{-,-} \perp X_{+,+} | X. \tag{22}$$

Proof. In order to demonstrate (19), it is sufficient to establish that $U_{1|X_{-,-}}^*$ and $U_{2|X_{-,-}}^*$ are doubly commuting isometries (see Proposition 1.1). The fact that these two isometries commute is obvious. Therefore, it remains to show that

$$U_1^* E^{X_{-,-}} U_2 (U_1^{-k} U_2^{-l} x) = E^{X_{-,-}} U_2 U_1^* (U_1^{-k} U_2^{-l} x) \tag{23}$$

for all $(k, l) \in N^2$, for all $x \in X$. Equation (23) is obvious if $l \geq 1$. On the other hand, $E^{X_{\infty,-}} U_2 U_1^{-k} x$ coincides with $U_1^{-k} E^{X_{\infty,-}} U_2 x$ because $U_1 X_{\infty,-} = X_{\infty,-}$. But, by (2), $E^{X_{\infty,-}} U_2 x = E^X U_2 x$, from which we deduce that $E^{X_{\infty,-}} U_2 U_1^{-k} x = U_1^{-k} E^X U_2 x$ belongs to $U_1^{-k} X$, and therefore to $X_{-,-}$. Consequently, $E^{X_{-,-}} U_2 U_1^{-k} x = E^{X_{\infty,-}} U_2 U_1^{-k} x =$

$U_1^{-k}E^X U_2 x$. Similarly, $E^{X_{-,\cdot}} U_2 U_1^* U_1^{-k} x = U_1^{-(k+1)} E^X U_2 x$, so that (23) holds for $l = 0$. We omit the proof of (20) and (21) which are based on similar arguments. Let us finally establish (22). For this purpose, we have to show that for all $x \in X$, for all $(k,l) \in N^2, E^{X_{-,\cdot}} U_1^k U_2^l x$ coincides with $E^X U_1^k U_2^l x$. By (19), $E^{X_{-,\cdot}} U_1^k U_2^l x = E^{X_{-,\infty}} E^{X_{\infty,\cdot}} U_1^k U_2^l x$. But, $E^{X_{\infty,\cdot}} U_1^k U_2^l x = U_1^k E^{X_{\infty,\cdot}} U_2^l x = U_1^k E^X U_2^l x = U_1^k T_2^l x$. Therefore, $E^{X_{-,\cdot}} U_1^k U_2^l x = E^{X_{-,\infty}} U_1^k T_2^l x$ which coincides with $E^X U_1^k T_2^l x = T_1^k T_2^l x = E^X U_1^k U_2^l x$. □

It turns out that the spaces $(X_{-,\infty}, X_{+,\infty}, X_{\infty,-})$ intersect perpendicularly two by two. Let us show that moreover,

$$X_{-,\infty} \cap X_{+,\infty} \cap X_{\infty,-} = X_{0,-}. \tag{24}$$

For this purpose, we have to establish that $E^{X_{\infty,\cdot}} z = E^{X_{0,\cdot}} z$ for each element z of $X_{-,\infty} \cap X_{+,\infty}$; as $X_{-,\infty} \cap X_{+,\infty} = X_{0,\infty}$, this is equivalent to $E^{X_{\infty,\cdot}} U_2^l x = E^{X_{0,\cdot}} U_2^l x$ for each $l \in Z$ and for each $x \in X$. If $l \leq 0$, it is obvious; if $l > 0, E^{X_{\infty,\cdot}} U_2^l x = T_2^l x = E^X U_2^l x = E^{X_{0,\cdot}} U_2^l x$. Similarly, $(X_{\infty,-}, X_{\infty,+}, X_{-,\infty})$ intersect perpendicularly two by two, and

$$X_{\infty,-} \cap X_{\infty,+} \cap X_{-,\infty} = X_{-,0}. \tag{25}$$

When T_1 and T_2 doubly commute, it is clear that the relations

$$E^{X_{+,\infty}} E^{X_{\infty,+}} = E^{X_{\infty,+}} E^{X_{+,\infty}} = E^{X_{+,+}}, \tag{26}$$

$$X_{+,-} \perp X_{-,+} | X \tag{27}$$

also hold. In this case, the four subspaces $X_{-,\infty}, X_{+,\infty}, X_{\infty,-}, X_{\infty,+}$ intersect perpendicularly two by two, and

$$X = X_{-,\infty} \cap X_{\infty,-} \cap X_{+,\infty} \cap X_{\infty,+}. \tag{28}$$

An interesting point is that the converse is also true. That is, if (26) holds, then T_1 and T_2 doubly commute. The proof is left to the reader. However, it is important to note that it is possible to exhibit examples of non doubly commuting contractions for which $E^{X_{+,\infty}} E^{X_{\infty,+}} = E^{X_{\infty,+}} E^{X_{+,\infty}}$ (see §3.2).

Let us finish this discussion concerning the properties of the spaces $X_{-,\infty}, X_{+,\infty}$, $X_{\infty,-}, X_{\infty,+}$ by an interpretation of the properties (19) to (21). The property (19) plays an important role in the prediction theory of two parameters stationary random processes. Let $(X_{m,n})_{(m,n)\in Z^2}$ be such a vector-valued process; let us denote by X the space generated by the components of $X_{0,0}$, and by (U_1, U_2) the horizontal and the vertical shift operators defined on the space generated by the components of the variables $X_{m,n}, (m,n) \in Z^2$. Then, if we assume for ease of exposition that $X_{-\infty,\infty} = \cap_{m\in Z} U_1^m X_{-,\infty}$ and $X_{\infty,-\infty} = \cap_{n\in Z} U_2^n X_{\infty,-}$ are reduced to $\{0\}$ (a condition that is similar to the concept of pure non determinism of the one-parameter case), it is easy to show that (19) holds if and only if $X_{m,n}$ has the following representation

$$X_{m,n} = \sum_{(k,l)\in N^2} c_{k,l} \alpha_{m-k,n-l}, \tag{29}$$

where α is a white noise sequence for which the space $sp(\alpha_{m,n})$ coincides with $sp(X_{m,n} - X_{m,n}/sp(X_{m-k,n-l}/(k,l) \in N^2, (k,l) \neq (0,0)))$ (see [10], [12], and [24]). Therefore, X

has a quarter-plane causal and causally invertible moving-average representation in terms of a white noise α. In particular, under purely technical assumptions, the quarter-plane causal Markov random fields introduced by Pickard satisfy condition (19). Obviously, a random field satisfying both conditions (19) to (21) must possess stronger properties. In particular, the above mentioned quarter-plane MRF do not satisfy (20) and (21) in the general case. Similarly, the condition (22) is very strong, so that it does not hold for the case of Pickard's MRF. Let us finally consider the case of an Attasi model $(X_{m,n})_{(m,n)\in Z^2}$ defined by (3). Then, it is easily seen that $X_{-\infty,\infty} = X_{\infty,-\infty} = \{0\}$ if and only if the matrices F_1 and F_2 are asymptotically stable. In this case, the representation (29) can be written as

$$X_{m,n} = \sum_{(k,l)\in N^2} F_1^k F_2^l \nu_{m-k-1,n-l-1}$$

so that $\alpha_{m,n} = \nu_{m-1,n-1}$ for each $(m,n) \in Z^2$. Moreover, for $(m,n) \in N^2, X_{m,n}/sp$ $(X_{-r,-s}/(r,s) \in N^2) = \sum_{k\geq m, l\geq n} F_1^k F_2^l \nu_{m-k-1,n-l-1}$. By using the fact that $F_1F_2 = F_2F_1$, we get immediately that

$$X_{m,n}/sp(X_{-r,-s}/(r,s) \in N^2) = F_1^m F_2^n X_{0,0}$$

which illustrates (22).

Now, we will study the properties of the spaces $(Z_i)_{i=1,2}$ defined by

$$Z_i = \overline{(U_i - T_i)X} \tag{30}$$

for $i = 1, 2$. Let us begin by some obvious properties of Z_1 (similar results hold for Z_2). As the space X is Markovian with respect to U_1, the space Z_1 coincides with the forward innovation space $U_1X_{-,0} \ominus X_{-,0}$; therefore, Z_1 is a wandering subspace for U_1. Moreover, the space $U_1X_{-,\infty} \ominus X_{-,\infty}$, which is generated by the vectors $U_2^n(U_1x - E^{X_{-,\infty}}U_1x)$ for $x \in X$ and $n \in Z$, coincides with $(Z_1)_{0,\infty} = \bigvee_{n\in Z} U_2^n Z_1$ because $E^{X_{-,\infty}}U_1U_2^n x = E^{X_{0,\infty}}U_1U_2^n x$. From this, it follows that $(Z_1)_{0,\infty}$ is equal to the forward innovation space of the Markovian space (with respect to U_1) $X_{0,\infty}$; in particular, $(Z_1)_{+,\infty} = (X_{-,\infty})^\perp$ (the orthogonal is taken in $X_{\infty,\infty}$). The properties of the Z_1 with respect to the unitary operator $U_{2|(Z_1)_{0,\infty}}$ will play an important role in the following.

THEOREM 2.2. *The space Z_1 is a Markovian space with respect to $U_{2|(Z_1)_{0,\infty}}$, i.e.,*

$$(Z_1)_{0,+} \perp (Z_1)_{0,-} | Z_1. \tag{31}$$

Moreover, Z_1 reduces the operator S_2 (i.e., Z_1 is invariant under S_2 and S_2^*), and $S_{2|Z_1}$ is the Markovian transition operator of Z_1.

The proof is given in the Appendix. Obviously, the space Z_2 has the same properties. Now, we study the forward and backward innovation spaces of Z_1 and Z_2.

THEOREM 2.3. *Let Δ be the space defined by*

$$\Delta = \overline{(U_1U_2 - U_1T_2 - U_2T_1 + T_1T_2)X}. \tag{32}$$

Then, Δ is equal to the wandering subspace for (U_1, U_2) defined by $(U_1X_{-,\infty} \ominus X_{-,\infty}) \cap (U_2X_{\infty,-} \ominus X_{\infty,-})$. Moreover, Δ coincides with both the forward innovation space of Z_1 and the forward innovation space of Z_2, i.e.,

$$\Delta = \overline{(U_2 - S_2)Z_1} = U_2(Z_1)_{0,-} \ominus (Z_1)_{0,-}, \tag{33}$$

$$\Delta = \overline{(U_1 - S_1)Z_2} = U_1(Z_2)_{-,0} \ominus (Z_2)_{-,0}. \tag{34}$$

Finally, the backward innovation space of Z_1 (i.e., the space $\overline{(I - U_2 S_2^*) Z_1}$) is equal to the wandering subspace for $(U_1, U_2)\Delta_{fb}$ given by

$$\Delta_{fb} = (U_1 X_{-,\infty} \ominus X_{-,\infty}) \cap (X_{\infty,+} \ominus U_2 X_{\infty,+}). \tag{35}$$

Similarly, the backward innovation space of Z_2 is equal to the wandering subspace Δ_{bf} given by

$$\Delta_{bf} = (X_{+,\infty} \ominus U_1 X_{+,\infty}) \cap (U_2 X_{\infty,-} \ominus X_{\infty,-}). \tag{36}$$

The proof is given in the Appendix. Motivated by the previous expressions of the spaces $\Delta, \Delta_{fb}, \Delta_{bf}$, we shall call Δ (respectively, Δ_{fb}, Δ_{bf}) the forward-forward (FF) (respectively, forward-backward, backward-forward) innovation space of X. Let us illustrate Theorems 2.2 and 2.3 in the previously introduced finite-dimensional case. In view of (32), the white noise sequence $(\nu_{m,n})_{(m,n)\in Z^2}$ is such that $\nu_{m,n} = U_1^m U_2^n \nu_{0,0}$ where $\nu_{0,0}$ is a basis of Δ. The fact that Z_1 is a Markovian space with respect to U_2 with transition operator $S_{2|Z_1}$ and forward innovation space Δ is equivalent to the fact that for each m fixed, the one-parameter N-variate random process $(X_{m+1,n} - F_1 X_{m,n})_{n\in Z}$ is a wide-sense Markov process given by the state-space equation

$$(X_{m+1,n+1} - F_1 X_{m,n+1}) = F_2 (X_{m+1,n} - F_1 X_{m,n}) + L\nu_{m,n}. \tag{37}$$

Let us now formulate some remarks concerning the spaces Δ_{fb} and Δ_{bf}. Although their properties present some similarities with those of Δ, Δ_{fb} and Δ_{bf} do not play the same role as Δ. In particular, the relation (32) has no counterpart in the case of Δ_{fb} and Δ_{bf}, except in the case where T_1 and T_2 doubly commute. Let us illustrate this in our finite-dimensional example. For ease of exposition, we suppose that the dimension of Z_1 coincides with the dimension of X, i.e., that $E[(X_{m+1,n} - F_1 X_{m,n})(X_{m+1,n} - F_1 X_{m,n})^T] = P - F_1 P F_1^T$ is positive-definite. Then, the backward state-space equation corresponding to (37) is given by

$$(X_{m+1,n} - F_1 X_{m,n}) = \tilde{F}_2 (X_{m+1,n+1} - F_1 X_{m,n+1}) + \tilde{L}\tilde{\nu}_{m,n}, \tag{38}$$

where $\tilde{\nu}_{m,n}$ represents a white noise sequence such that $sp(\tilde{\nu}_{0,0})$ coincides with the space Δ_{fb}, and where $\tilde{F}_2$ is the matrix given by

$$\tilde{F}_2 = (P - F_1 P F_1^T) F_2^T (P - F_1 P F_1^T)^{-1}. \tag{39}$$

Equation (38) can be written as

$$X_{m+1,n} = F_1 X_{m,n} + \tilde{F}_2 X_{m+1,n+1} - \tilde{F}_2 F_1 X_{m,n+1} + \tilde{L}\tilde{\nu}_{m,n}. \tag{40}$$

In general, the matrices F_1 and $\tilde{F}_2$ do not commute, so that the process $\tilde{X}_{m,n} = X_{m,-n}$ is not an Attasi model, or equivalently the conditional orthogonality relation $X_{0,-} \perp X_{\infty,+} | X$ does not hold. Therefore, the expression of $\tilde{\nu}_{m,n}$ in terms of $X_{m+1,n}, X_{m+1,n+1}, X_{m,n}, X_{m,n+1}$ is not similar to that of $\nu_{m,n}$ in terms of $X_{m+1,n+1}, X_{m+1,n}, X_{m,n+1}, X_{m,n}$, or equivalently, (32) has no counterpart in the case of Δ_{fb}. However, when T_1 and T_2 doubly commute, i.e., when F_1 and $P F_2^T P^{-1}$ commute, it is easily seen that $\tilde{F}_2$ coincides with $P F_2^T P^{-1}$, so that (40) is an Attasi-like equation.

3. Application to a Markovian-like representation problem. In this section, we deal with the Markovian-like representation problem corresponding to the previously introduced Markovian properties. Its formulation can be interpreted as the two-parameter counterpart of the problem introduced by Lindquist–Picci and Ruckebusch. Throughout this section, (U_1, U_2) represents a fixed once and for all pair of commuting unitary operators defined on the Hilbert space K. Then, a subspace X of K is said to be a forward-forward Markovian representation (FFMR) of Y (with respect to (U_1, U_2)) if Y is included in X and if X is a forward-forward Markovian space with respect to (U_1, U_2). Forward-backward, backward-forward, backward-backward Markovian representations are defined in a similar way. Finally, X will be said to be a Markovian representation of Y if X is both a FFMR and a BBMR of Y. Thus, by §2, a subspace X of K containing Y is a FFMR (respectively, a MR) of Y if and only if $(U_{1|X_{\infty,\infty}}, U_{2|X_{\infty,\infty}}, X_{\infty,\infty})$ is the regular minimal unitary dilation of a pair of commuting (respectively, doubly commuting) contractions defined on X. According to [14], we shall say that a (FF)MR X of Y is internal if X is contained in $Y_{\infty,\infty}$; such representations are clearly of special interest because they can be constructed from the "data" of the realization problem (i.e., the space $Y_{\infty,\infty}$).

Let us begin by the following useful remark which derives from the fact that if X is a FF Markovian space with respect to (U_1, U_2), then $X_{0,\infty}$ (respectively, $X_{\infty,0}$) is a Markovian space with respect to U_1 (respectively, U_2): if X is FFMR of Y, then the space $X_{0,\infty}$ (respectively, the space $X_{\infty,0}$) is a Markovian representation of the space $Y_{0,\infty}$ (respectively, of $Y_{\infty,0}$) with respect to U_1 (respectively, U_2) [14], in the sense that $X_{0,\infty}$ (respectively, $X_{\infty,0}$) is a Markovian space with respect to U_1 (respectively, U_2) containing $Y_{0,\infty}$ (respectively, $Y_{\infty,0}$); moreover, its Markovian transition opertor coincides with the contraction S_1 (respectively, S_2) defined by (17) (respectively, (18)). This elementary property will allow us to use the theory of [14] and [19] in order to get some results concerning the minimality of the (FF)MR of Y with respect to (U_1, U_2).

3.1. Geometrical considerations. In this paragraph no particular assumption is made on the space Y. We give a geometrical characterization of the set of all (FF)MR of Y. We begin by the FFMR.

THEOREM 3.1. *A subspace X of K is a* FFMR *of Y if and only if*

$$X = \overline{E}^{H_1 \cap H_2}(\overline{H}), \tag{41}$$

where $H_1, H_2, \overline{H}$ are three subspaces containing Y such that

$$U_i(\overline{H}) \subset \overline{H} \quad \text{for } i = 1, 2, \tag{42}$$

$$H_i \subset U_i H_i, U_j H_i = H_i, \qquad j \neq i, \quad i = 1, 2, \tag{43}$$

$$H_1 \text{ and } H_2 \text{ intersect perpendicularly, i.e., } E^{H_1} E^{H_2} = E^{H_2} E^{H_1}. \tag{44}$$

Proof. Let X be a FFMR of Y; then, if we put $H_1 = X_{-,\infty}, H_2 = X_{\infty,-}, \overline{H} = X_{+,+}$, it follows from Proposition 2.1 that $X = \overline{E}^{H_1 \cap H_2}(\overline{H})$, and that $H_1, H_2, \overline{H}$ satisfy all the requirements of the theorem. Conversely, let us suppose that X is a subspace given by (41), and let us show that $X_{-,\infty} \perp X_{+,0} | X$; as $X_{-,\infty}$ is included in H_1, it is clearly sufficient to show that $H_1 \perp X_{+,0} | X$. From $m \in N, U_1^m X = \overline{E}^{U_1^m H_1 \cap H_2}(U_1^m \overline{H})$ is included in

$\overline{E}^{U_1^m H_1 \cap H_2}(\overline{H})$; therefore, $\overline{E}^{H_1}(U_1^m X)$ is itself included in $\overline{E}^{H_1}(\overline{E}^{U_1^m H_1 \cap H_2}(\overline{H}))$. Let us show that this subspace is contained in X; for all $z \in X, E^{H_1}(E^{U_1^m H_1 \cap H_2}(z)) = E^{H_1} E^{U_1^m H_1} E^{H_2} z = E^{H_1} E^{H_2} z = E^{H_1 \cap H_2} z$ which belongs to X; this implies that $\overline{E}^{H_1}(\overline{E}^{U_1^m H_1 \cap H_2}(\overline{H}))$ is contained in X; thus, $\overline{E}^{H_1}(U_1^m X) \subset X$ for each $m \in N$, from which we deduce that $H_1 \perp X_{+,0} | X$. The fact that $X_{\infty,-} \perp X_{0,+} | X$ is shown similarly. □

The characterization of the set of all MR of Y appears as a generalization of the basic representation Theorem 3.1 of [14].

THEOREM 3.2. *A subspace X of K is a* MR *of Y with respect to (U_1, U_2) if and only if X can be written as*

$$X = H_1 \cap H_2 \cap \overline{H_1} \cap \overline{H_2}, \tag{45}$$

where $H_1, H_2, \overline{H_1}, \overline{H_2}$ are four subspaces containing Y such that

$$U_i(\overline{H_i}) \subset \overline{H_i} \quad \text{for } i = 1,2 \; U_j \overline{H_i} = \overline{H_i}, \qquad j \neq i, \quad i = 1,2, \tag{46}$$

$$H_i \subset U_i H_i, \quad U_j H_i = H_i, \qquad j \neq i, \quad i = 1,2, \tag{47}$$

$$H_1, H_2, \overline{H_1}, \overline{H_2} \text{ intersect perpendicularly two by two.} \tag{48}$$

Moreover, if X is an internal MR, then $(X_{-,\infty}, X_{\infty,-}, X_{+,\infty}, X_{\infty,+})$ is the unique 4*-uple of subspaces contained in $Y_{\infty,\infty}$ and satisfying* (45)–(48).

Proof. If X is a MR of Y, we have already shown that $X = X_{-,\infty} \cap X_{+,\infty} \cap X_{\infty,-} \cap X_{\infty,+}$ (see (28)) and that $H_1 = X_{-,\infty}, \overline{H_1} = X_{+,\infty}, H_2 = X_{\infty,-}$ and $\overline{H_2} = X_{\infty,+}$ satisfy (46) to (48). Conversely, if X is given by (45), X can be written as $X = \overline{E}^{H_1 \cap H_2}(\overline{H_1} \cap \overline{H_2})$; by Theorem 3.1, this implies that X is a FFMR of Y. But, X is also equal to $X = \overline{E}^{(\overline{H_1} \cap \overline{H_2})}(H_1 \cap H_2)$ which is a BBMR by a trivial modification of Theorem 3.1. This in turn establishes that X is a MR of Y. The proof of the last statement of the theorem is left to the reader. □

By Theorem 3.1, the spaces $E^{Y_{-,\infty}}(Y_{+,+})$ and $E^{Y_{\infty,-}}(Y_{+,+})$ are FFMR of Y (take $\overline{H} = Y_{+,+}, H_1 = Y_{-,\infty}, H_2 = Y_{\infty,\infty}$ and $\overline{H} = Y_{+,+}, H_1 = Y_{\infty,\infty}, H_2 = Y_{\infty,-}$, respectively). However, these two FFMR are not interesting because they represent infinite-dimensional subspaces, even if Y is a finite-dimensional subspace admitting finite-dimensional FFMR (for example, $E^{Y_{-,\infty}}(Y_{+,+})$ contains $Y_{0,+}$ which is an infinite-dimensional subspace, except in trivial cases). A natural question is whether there exist less special examples of FFMR in the general case. In particular, is it possible to exhibit minimal FFMR or MR (in the sense that they do not contain another representation as proper subspace) without making any assumptions on Y, as it is the case in the one-parameter theory (the filter and the co-filter [19] are such minimal representations)? The answer to this question is negative because there may not exist a nontrivial pair of subspaces (H_1, H_2) containing Y and satisfying (43) and (44) (by a nontrivial pair, we mean that H_1 and H_2 do not contain $Y_{\infty,\infty}$, i.e., in the case where $K = Y_{\infty,\infty}$ that $H_1 \neq Y_{\infty,\infty}$ and $H_2 \neq Y_{\infty,\infty}$). This claim will be demonstrated in §3.2 (see Proposition 3.2). Therefore, it is not possible to develop a Markovian representation theory without making any extra assumptions on the space Y. Consequently, we consider in §3.2 restrictive situations for which some positive results hold. But, before going further, we introduce some useful definitions.

DEFINITION 3.1. Let X be an FFMR of Y. Then, X is said to be forward-forward observable if $X \cap (Y_{+,+})^{\perp} = \{0\}$, or equivalently, if $X = \overline{E}^{X}(Y_{+,+})$.

This definition makes sense for the following reason: let X be a non FF observable FFMR. Then, $X^0 = \overline{E}^X(Y_{+,+})$ is an FF observable FFMR contained in X. In fact, as $\overline{H} = Y_{+,+}, H_1 = X_{-,\infty}, H_2 = X_{\infty,-}$ satisfy (42)–(44), the space $\overline{E}^{X_{-,\infty} \cap X_{\infty,-}}(Y_{+,+})$ is an FFMR of Y; but, $X_{-,\infty} \cap X_{\infty,-}$ and $X_{+,+}$ are conditionally orthogonal given X; therefore, this is also the case for $X_{-,\infty} \cap X_{\infty,-}$ and $Y_{+,+}$, from which we deduce that $\overline{E}^{X_{-,\infty} \cap X_{\infty,-}}(Y_{+,+}) = \overline{E}^X(Y_{+,+})$; consequently, X^0 is an FFMR. The fact that X^0 is FF observable is obvious. Let us define now the concept of regular and coregular representation; the terminology originates from [19].

DEFINITION 3.2. Let X be an (FF)MR of Y. Then, X is said to be regular if

$$\bigcap_{m \in Z} U_1^m X_{-,\infty} = \{0\}, \tag{49}$$

$$\bigcap_{n \in Z} U_2^n X_{\infty,-} = \{0\}. \tag{50}$$

X is said to be co-regular if

$$\bigcap_{m \in Z} U_1^m X_{+,\infty} = \{0\}, \tag{51}$$

$$\bigcap_{n \in Z} U_2^n X_{\infty,+} = \{0\}. \tag{52}$$

In the sequel, we shall be concerned with regular and coregular FFMR of Y. Therefore, it will be useful to know whether a regular FFMR X is also coregular.

LEMMA 3.1. *Let X be a regular* FFMR *of Y whose forward-forward innovation space Δ is finite-dimensional. Then, X is co-regular if and only if*

$$\dim\Delta = \dim\Delta_{fb} = \dim\Delta_{bf}, \tag{53}$$

where Δ_{fb} and Δ_{bf} are the innovation spaces defined in Theorem 2.3. *In this case, $X_{-,+} = U_1^*(\Delta_{fb})_{-,+}, X_{\infty,+} = (\Delta_{fb})_{\infty,+}$ and $X_{+,-} = U_2^*(\Delta_{bf})_{+,-}, X_{+,\infty} = (\Delta_{bf})_{+,\infty}$.*

Proof. See the Appendix.

Finally, we give the following useful lemma whose proof is omitted.

LEMMA 3.2. *Let Δ be a wandering subspace for (U_1, U_2) such that*

(i) *Δ is included in $Y_{\infty,\infty}$;*

(ii) *Y is included in $U_1^* U_2^* \Delta_{-,-}$.*

Then, the subspace $X = \overline{E}^{U_1^* U_2^* \Delta_{-,-}}(Y_{+,+})$ is an internal regular FF *observable* FFMR *of Y with forward-forward innovation space Δ.*

A fundamental example. In order to illustrate what precedes, we give an example which will be very useful in §3.2. Let us denote by T the unit circle of the complex plane, and by D (respectively, $\overline{D}$) the open (respectively, closed) unit disk. Then, we consider the case where K is equal to the space $L^2(T^2)$ of all square integrable functions (with respect to the Lebesgue measure) defined on T^2, and where U_1 and U_2 coincide with the multiplication operators σ_1 and σ_2 by z_1 and z_2, respectively (in the following, the generic element of T^2 will be denoted (z_1, z_2) or $(e^{i\omega_1}, e^{i\omega_2})$).

Before going further, we have to introduce some basic notations. First, H_+^2 and H_-^2 denote the following Hardy subspaces of $L^2(T)$:

$$H_+^2 = \left\{ \sum_{k=0}^{\infty} \phi_k z^k \right\}, \qquad H_-^2 = \left\{ \sum_{k=0}^{\infty} \phi_k z^{-k} \right\}.$$

A function θ defined on T is said to be all-pass if $|\theta(z)| = 1$ almost everywhere on T; if moreover $\theta \in H_+^2$ (respectively, H_-^2), then θ is said to be inner (respectively, $*$-inner). In the following, some Hardy spaces of functions defined on T^2 will play an important role. We denote by $H_{+,\infty}^2$ and $H_{-,\infty}^2$ the subspaces

$$H_{+,\infty}^2 = \left\{ \sum_{k=0}^{\infty} \sum_{l=-\infty}^{\infty} \phi_{k,l} z_1^k z_2^l \right\}, \qquad H_{-,\infty}^2 = \left\{ \sum_{k=0}^{\infty} \sum_{l=-\infty}^{\infty} \phi_{k,l} z_1^{-k} z_2^l \right\}.$$

Clearly, the elements of $H_{+,\infty}^2$ (respectively, of $H_{-,\infty}^2$) can be extended for almost all $z_2 \in T$ and for all $z_1 \in D$ (respectively, for all $z_1 \in (\overline{D})^c$). In the sequel, the elements of $H_{+,\infty}^2$ (respectively, of $H_{-,\infty}^2$) will be called 1-analytic (respectively, $1*$-analytic) functions. The subspaces $H_{\infty,+}^2$ and $H_{\infty,-}^2$ are deduced from the previous one by exchanging the role of z_1 and z_2. $H_{+,+}^2$ represents the space

$$H_{+,+}^2 = \left\{ \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} \phi_{k,l} z_1^k z_2^l \right\}$$

and $H_{-,+}^2, H_{+,-}^2, H_{-,-}^2$ are defined similarly. The subspace of $H_{+,\infty}^2$ (respectively, of $H_{\infty,+}^2$) made of the functions not depending on z_2 (respectively, on z_1) is denoted by $H_+^2(T_1)$ (respectively, $H_+^2(T_2)$); $H_-^2(T_1)$ and $H_-^2(T_2)$ are defined similarly.

Finally, an all-pass function θ defined on T^2 is said to be 1-inner (respectively, $1*$-inner) if θ belongs to $H_{+,\infty}^2$ (respectively, to $H_{-,\infty}^2$). 2-inner and $2*$-inner functions are defined similarly. If ϕ is an element of $H_{+,\infty}^2$ and if α is a 1-inner function, then we say that α is a 1-inner divisor of ϕ if ϕ can be written as $\phi = \phi'\alpha$ for some element ϕ' of $H_{+,\infty}^2$. If α does not depend on z_2 (i.e., if α is inner), α is said to be an inner divisor of ϕ. If θ is a 1-inner function, ϕ and θ are said to be 1-coprime if they have no 1-inner common divisor; when θ depends only on z_1, ϕ and θ are said to be 1-weak coprime if ϕ and θ have no common inner divisor. The above definitions can be extended to the case where $\phi \in H_{-,\infty}^2$ and where θ is a $1*$-inner function, $\phi \in H_{\infty,+}^2$ and where θ is a 2-inner function, etc.

Let ϕ be a nonzero function of $\sigma_1^* \sigma_2^* H_{-,-}^2$, i.e.,

$$\phi(z_1, z_2) = \sum_{k \geq 1, l \geq 1} \phi_{k,l} z_1^{-k} z_2^{-l}.$$

Let us denote by Y the one-dimensional subspace of $K = L^2(T^2)$ generated by ϕ, and by X the space defined by

$$X = \overline{E^{\sigma_1^* \sigma_2^* H_{-,-}^2}(\overline{\phi H_{+,+}^2})}. \tag{54}$$

Clearly, X coincides with the range of the Hankel operator $\mathcal{H}_\phi$ defined by

$$\mathcal{H}_\phi \begin{array}{ccc} H_{+,+}^2 & \longrightarrow & \sigma_1^* \sigma_2^* H_{-,-}^2, \\ f & \longrightarrow & E^{\sigma_1^* \sigma_2^* H_{-,-}^2}(\phi f). \end{array} \tag{55}$$

Then, by Theorem 3.1, X is a FFMR of Y with respect to (σ_1, σ_2) (take $\overline{H} = \overline{\phi H^2_{+,+}}$, $H_1 = \sigma_1^* H^2_{-,\infty}$, $H_2 = \sigma_2^* H^2_{\infty,-}$); let us also mention that by §2, we get that $T_1 = E^X \sigma_{1|X}$ and $T_2 = E^X \sigma_{2|X}$ are commuting contractions admitting $(\sigma_1, \sigma_2, L^2(T^2))$ as regular minimal unitary dilation (by using the fact that a nonzero function of $\sigma_1^* \sigma_2^* H^2_{-,-}$ cannot be zero on a set of positive measure [20] and that every doubly invariant subspace H under σ_1 and σ_2 can be written as $H = 1_B L^2(T^2)$ for some Borel subset B of T^2 [8], it is easily seen that $Y_{\infty,\infty} = \overline{\phi L^2(T^2)} = L^2(T^2) = X_{\infty,\infty}$). Moreover, X is FF observable, regular, and by Lemma 3.2, its forward-forward innovation space Δ coincides with the one-dimensional subspace generated by the constant functions.

Let us now describe the spaces Z_1 and Z_2; as their properties are similar, we only give the results that are relative to Z_1. Let us begin by introducing the following functions $\phi_{2,k}$ (for $k \geq 1$) and $\phi_{1,l}$ (for $l \geq 1$) defined on T by

$$\phi_{2,k}(z_2) = \int_{[-\pi,\pi]} \phi(e^{i\omega_1}, z_2) e^{ik\omega_1} d\omega_1, \tag{56}$$

$$\phi_{1,l}(z_1) = \int_{[-\pi,\pi]} \phi(z_1, e^{i\omega_2}) e^{il\omega_2} d\omega_2. \tag{57}$$

Then, by an easy calculation, we get that

$$Z_1 = \bigvee_{k \geq 1} \overline{E}^{\sigma_2^* H^2_-(T_2)} (\overline{\phi_{2,k} H^2_+(T_2)}). \tag{58}$$

Let us now investigate the conditions under which X is coregular. In view of Lemma 3.1, we have to characterize the innovation spaces Δ_{fb} and Δ_{bf}. By Theorems 2.2 and 2.3, Z_1 is Markovian with respect to $\sigma_{2|(Z_1)_{0,\infty}}$ and its forward and backward innovation spaces coincide with Δ and Δ_{fb}, respectively. In particular, this implies that the space $(Z_1)_{0,-}$ is given by $(Z_1)_{0,-} = Z_1 \ominus \sigma_2^*(\Delta_{fb})_{0,-}$. On the other hand, it is easily seen that $\cap_{n \in Z} \sigma_2^n (Z_1)_{0,-} = \{0\}$; as Δ coincides with the space generated by the constant functions, we get that $(Z_1)_{0,-} = \sigma_2^* H^2_-(T_2)$ and that $(Z_1)_{0,\infty} = L^2(T_2)$. By classical arguments of multiplicity theory, it turns out that $\dim \Delta_{fb} = 0$ or 1. Therefore, it appears that $\dim \Delta_{fb} = 1$ if and only if the space Z_1 generated by the range of the Hankel operators associated with the one-variable functions $\phi_{2,k}$ for $k \geq 1$ does not coincide with $\sigma_2^*(H^2_-(T^2))$. This condition on ϕ is similar to the concept of noncyclicity introduced in [3] (see also [2], [6]), in the sense that it can be interpreted formally as the noncyclicity of the infinite-dimensional row-vector valued-function $(\phi_{2,1}, \ldots, \phi_{2,n}, \ldots)$. In the following, we shall say that a function satisfying the above property is 2-weakly noncyclic; 1-weakly noncyclic functions are defined similarly by exchanging the role of z_1 and z_2.

Clearly, the FFMR X is coregular if and only if the function ϕ is both 1-weakly and 2-weakly noncyclic. It is therefore useful to study in more details the 1-weakly and 2-weakly noncyclic functions. The following result can be seen as a generalization of the well known characterization of one-variable noncyclic functions (see [3]).

PROPOSITION 3.1. *Let ϕ be an element of $\sigma_1^* \sigma_2^* H^2_{-,-}$. Then ϕ is 2-weakly noncyclic if and only if there exists a $*$-inner function ω_2 and a function $\psi \in \sigma_1^* H^2_{-,+}$ for which*

$$\phi(z_1, z_2) = \psi(z_1, z_2) \omega_2(z_2). \tag{59}$$

In this case, there exists a unique pair $(\phi_{-,+}, \theta_2)$ (where $\phi_{-,+} \in \sigma_1^* H^2_{-,+}$ and where θ_2 is a $*$-inner function) satisfying

$$\phi(z_1, z_2) = \phi_{-,+}(z_1, z_2) \theta_2(z_2), \tag{60}$$

(61) $\phi_{-,+}$ *and* θ_2^* $2-$ *weak coprime.*

Moreover, θ_2 *is uniquely defined by the fact that*

(62) $$\bigvee_{k\geq 1} \overline{E^{\sigma_2^* H_-^2(T_2)}(\overline{\phi_{2,k} H_+^2(T_2)})} = \sigma_2^*(H_-^2(T_2) \ominus \theta_2 H_-^2(T_2)).$$

Proof. See the Appendix.

3.2. Spectral considerations. In this section we restrict ourselves to the case where Y is a one-dimensional subspace; moreover, we will only be concerned with internal representations, so that the Hilbert space K on which are defined U_1 and U_2 will be supposed to be $Y_{\infty,\infty}$. In order to make more concrete what follows, we present our results in the framework of the two-parameter wide-sense stationary stochastic processes theory.

Let $y = (y_{m,n})_{(m,n)\in Z^2}$ be a centered two-parameter wide-sense stationary stochastic process, that is, a sequence of complex-valued centered square integrable random variables for which $E(y_{m+k,n+l} y_{k,l}^*)$ depends only on (m,n)(E denotes the mathematical expectation in this context). We suppose that K is the Hilbert space generated by the variables $y_{m,n}$ for $(m,n) \in Z^2$, endowed with the scalar product $\langle z_1, z_2\rangle = E(z_1, z_2^*)$. U_1 and U_2 are the horizontal and vertical shift operators associated to y, that is the commuting unitary operators defined on K by $U_1 y_{m,n} = y_{m+1,n}$ and $U_2 y_{m,n} = y_{m,n+1}$ for each $(m,n) \in Z^2$. Let us put $Y = sp(y_{0,0})$; then, we are concerned with the study of the (FF)MR of Y with respect to (U_1, U_2).

Before going further, let us recall some basic definitions related to the stochastic process y. If E_1 and E_2 are the spectral families of U_1 and U_2 defined on $[-\pi,\pi]$, the random spectral measure $\hat{y}$ of y if the K-valued orthogonally scattered measure whose differential element is given by $d\hat{y}(\omega_1,\omega_2) = dE_1(\omega_1) dE_2(\omega_2) y_{0,0}$. Then, it is well known that y has the following so-called spectral representation:

$$y_{m,n} = \int_{[-\pi,\pi]^2} e^{i(m\omega_1 + n\omega_2)} d\hat{y}(\omega_1,\omega_2).$$

The positive bounded measure μ defined on the Borel sets of $[-\pi,\pi]^2$ by the differential element $d\mu(\omega_1,\omega_2) = E|d\hat{y}(\omega_1,\omega_2)|^2$ is called the spectral measure of y. In what follows, we shall suppose that μ is absolutely continuous with respect to the Lebesgue measure, and that its spectral density $F(\omega_1,\omega_2)$ satisfies the Szegö condition

(63) $$\int_{[-\pi,\pi]^2} LogF(\omega_1,\omega_2) d\omega_1 d\omega_2 > -\infty.$$

In this case, the process y is purely nondeterministic with respect to the prediction problems associated to the lexicographical-like orderings of Z^2 (see [9], for example). In particular, the spaces $\cap_{m\in Z} U_1^m Y_{-,\infty} = \cap_{n\in Z} U_2^n Y_{\infty,-} = \{0\}$.

We begin by showing that there do not necessarily exist nontrivial FFMR of Y, which will demonstrate the claim of §3.1. This follows from the fact that the existence of a nontrivial pair of subspaces (H_1, H_2) containing Y and satisfying conditions (43) and (44) is not guaranteed (see Theorem 3.1). More precisely, we have the following result.

PROPOSITION 3.2. *There exists a nontrivial pair* (H_1, H_2) *containing* Y *and satisfying conditions* (43) *and* (44) *if and only if there exists a function* $\phi \in \sigma_1^* \sigma_2^* H_{-,-}^2$ *for which*

(64) $$F(\omega_1,\omega_2) = |\phi(e^{i\omega_1}, e^{i\omega_2})|^2.$$

In order to demonstrate this result, we give the following lemma whose proof is given in the Appendix.

LEMMA 3.3. *Let* (H_1, H_2) *be two subspaces of* $Y_{\infty,\infty}$ *satisfying* (43) *and* (44), *and for which*

$$\bigvee_{m\in Z} U_1^m H_1 = \bigvee_{n\in Z} U_2^n H_2 = Y_{\infty,\infty}. \tag{65}$$

Then, one of the two following conditions holds.

$$H_1 = Y_{\infty,\infty} \quad or \quad H_2 = Y_{\infty,\infty}, \tag{66}$$

$$\bigcap_{m\in Z} U_1^m H_1 = \bigcap_{n\in Z} U_2^n H_2 = \{0\}. \tag{67}$$

If (67) *holds, then*

$$H_1 \cap H_2 = \oplus_{(k,l)=1,\infty} U_1^{*k} U_2^{*l} \Delta, \tag{68}$$

where Δ *is the wandering subspace for* (U_1, U_2) *defined by*

$$\Delta = (U_1 H_1 \ominus H_1) \cap (U_2 H_2 \ominus H_2). \tag{69}$$

Now, we prove the proposition. Let us suppose the existence of a nontrivial FFMR of Y, i.e., the existence of a nontrivial pair of subspaces (H_1, H_2) containing Y and satisfying (43) and (44). Let us denote by X the FF-observable FFMR given by

$$X = \overline{E}^{H_1 \cap H_2}(Y_{+,+}).$$

As Y is included in H_1 and H_2, the pair (H_1, H_2) satisfies (65). By Lemma 3.3, we deduce that condition (67) holds. The wandering subspace Δ given by (69) is therefore not reduced to $\{0\}$; as the pair (U_1, U_2) has multiplicity one, Δ is one-dimensional. Let $\nu_{0,0}$ be a unit vector of Δ and put $\nu_{m,n} = U_1^m U_2^n \nu_{0,0}$; then, $\nu = (\nu_{m,n})_{(m,n)\in Z^2}$ is a white noise sequence satisfying $H_1 \cap H_2 = sp(\nu_{-k,-l}/k \geq 1, l \geq 1)$. This implies that the FFMR X is regular; moreover, as $y_{0,0}$ belongs to X, there exists a square summable sequence $(\phi_{k,l})_{k\geq 1, l\geq 1}$ such that

$$y_{0,0} = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \phi_{k,l} \nu_{-k,-l}.$$

Put $\phi(z_1, z_2) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \phi_{k,l} z_1^{-k} z_2^{-l}$. Then, ϕ belongs to $\sigma_1^* \sigma_2^* H^2_{-,-}$, and $F(\omega_1, \omega_2) = |\phi(e^{i\omega_1}, e^{i\omega_2})|^2$.

Conversely, let us suppose the existence of a function $\phi \in \sigma_1^* \sigma_2^* H^2_{-,-}$ satisfying (64). Then, the wide-sense stationary process ν defined by the fact that $d\hat{\nu}(\omega_1, \omega_2) = \phi^{-1}(e^{i\omega_1}, e^{i\omega_2}) d\hat{y}(\omega_1, \omega_2)$ is a white noise. Let us put $\Delta = sp(\nu_{0,0})$; then, it is clear that Δ is wandering for (U_1, U_2) and that $y_{0,0} \in U_1^* U_2^* \Delta_{-,-}$. From this, we deduce immediately that the spaces H_1 and H_2 defined by $H_1 = U_1^* \Delta_{-,\infty}$ and $H_2 = U_2^* \Delta_{\infty,-}$ contain Y and satisfy (43) and (44). □

Proposition 3.2 implies that Y do not necessarily admit nontrivial FFMR. In fact, by a trivial modification of exercise 3.4.5, [20, p. 56] there exist spectral densities satisfying the Szegö condition, but which are not factorable by an element of $\sigma_1^* \sigma_2^* H^2_{-,-}$. Obviously, it would be interesting to derive the necessary and sufficient conditions on F for the existence of a spectral factor belonging to $\sigma_1^* \sigma_2^* H^2_{-,-}$. According to [20], this point seems to be an open problem. Let us mention however two sufficient conditions.

- $F > 0$, bounded and lower semi-continuous ([20, p. 55]).
- $\int_{[-\pi,\pi]^2} LogF(\omega_1,\omega_2)e^{i(m\omega_1+n\omega_2)}d\omega_1 d\omega_2 = 0$ for $mn < 0$. ([17, Lemma A1], [24], [12]).

This last condition is equivalent to the equality of the two innovation processes corresponding to the prediction problems associated to the column-by-column, and row-by-row lexicographical ordering of Z^2, respectively [12]. In this case, $Y_{-,\infty}$ and $Y_{\infty,-}$ intersect perpendicularly [24], [12], so that the space $\overline{E}^{Y_{-,\infty}\cap Y_{\infty,-}}(Y_{+,+})$ is an FFMR which is easily seen to be minimal.

Let us formulate some remarks concerning Proposition 3.2 and Lemma 3.3. First, it follows from Lemma 3.3 that an FFMR is nontrivial if and only if it is regular. Second, paralleling the proof of Proposition 3.2, it is easily seen that it is possible to associate to each regular FFMR X a spectral factor $\phi \in \sigma_1^*\sigma_2^* H^2_{-,-}$ defined by the fact that $d\hat{y}(\omega_1,\omega_2) = \phi(e^{i\omega_1}, e^{i\omega_2})d\hat{\nu}(\omega_1,\omega_2)$, where ν is a white noise sequence for which $\mathrm{sp}(\nu_{0,0})$ coincides with the FF innovation space of X. In the following, ϕ and ν will be called the forward-forward spectral factor of X and the forward-forward innovation process of X, respectively.

Let us now investigate under which conditions there exist nontrivial MR of Y; here, we mean by non trivial an MR in which none of the four internal subspaces defined by (45) coincide with $Y_{\infty,\infty}$. For this purpose, we need to introduce the following useful notation: if $\nu = (\nu_{m,n})_{(m,n)\in Z^2}$ is a (scalar) white noise sequence for which $Y_{\infty,\infty} = sp(\nu_{m,n}/(m,n) \in Z^2)$ (the existence of such a white noise is guaranteed by the Szegö condition), then we denote by Q_ν the unitary operator defined by

$$(70) \qquad Q_\nu \begin{array}{ccc} L^2(T^2) & \longrightarrow & Y_{\infty,\infty}, \\ h & \longrightarrow & \int_{[-\pi,\pi]^2} h(e^{i\omega_1}, e^{i\omega_2})d\hat{\nu}(\omega_1,\omega_2). \end{array}$$

PROPOSITION 3.3. *There exists a nontrivial* MR *of Y if and only if the spectral density F is factorable by a* 1- *and* 2-*weakly noncyclic function ϕ belonging to $\sigma_1^*\sigma_2^* H^2_{-,-}$.*

Proof. Suppose that $X = H_1 \cap H_2 \cap \overline{H_1} \cap \overline{H_2}$ is a nontrivial MR of Y. Then, by Lemma 3.3 applied to the pairs $(H_1,H_2),(H_1,\overline{H_2}),(\overline{H_1},H_2)$, we get that $\cap_{m\in Z} U_1^m H_1 = \cap_{n\in Z} U_2^n H_2 = \cap_{m\in Z} U_1^m \overline{H_1} = \cap_{n\in Z} U_2^n \overline{H_2} = \{0\}$. This implies that X is a regular and coregular MR; therefore, the FF-observable FFMR $X^0 = \overline{E}^X(Y_{+,+})$ is also regular and coregular. Let ϕ be its associated FF spectral factor, and let us denote by ν its FF innovation process. Then, it is clear that $X^0 = Q_\nu(\mathrm{Range}\mathcal{H}_\phi)$, where $\mathcal{H}_\phi$ is the Hankel operator defined by (55). Therefore, in view of the considerations concerning the fundamental example of §3.1, the fact that X^0 is coregular implies that ϕ is 1- and 2-weakly noncyclic.

Conversely, let us assume the existence of a 1- and 2-weakly noncyclic element ϕ of $\sigma_1^*\sigma_2^* H^2_{-,-}$ for which (64) holds. Then, there exist two $*$-inner functions θ_1 and θ_2 such that $\phi\theta_1^* \in \sigma_2^* H^2_{+,-}$ and $\phi\theta_2^* \in \sigma_1^* H^2_{-,+}$. Let ν be the white noise sequence defined by the fact that $d\hat{y}(\omega_1,\omega_2) = \phi(e^{i\omega_1}, e^{i\omega_2})d\hat{\nu}(\omega_1,\omega_2)$; then it is easily seen that the space X defined by

$$(71) \qquad X = Q_\nu[\sigma_1^* H^2_{-,\infty} \cap \sigma_2^* H^2_{\infty,-} \cap \theta_1 H^2_{+,\infty} \cap \theta_2 H^2_{\infty,+}]$$

$$(72) \qquad = Q_\nu[\sigma_1^*(H^2_-(T_1) \ominus \theta_1 H^2_-(T_1)) \otimes \sigma_2^*(H^2_-(T_2) \ominus \theta_2 H^2_-(T_2))]$$

(where $\otimes$ represents the Hilbertian tensor product) is a nontrivial (or equivalently, regular and coregular) MR of Y. □

Proposition 3.3 shows that the Markovian representation problem studied in this paper makes sense when the spectral density F of y is supposed to admit 1- and 2-weakly noncyclic spectral factor belonging to $\sigma_1^*\sigma_2^* H^2_{-,-}$. Therefore, we shall suppose from now

on that y satisfies this condition. This occurs, for example, if F can be written as

$$F(\omega_1,\omega_2)=\left|\sum_{k=0}^{M}\sum_{l=0}^{N}b_{k,l}e^{-i(k\omega_1+l\omega_2)}\right|^2 F_1(\omega_1)F_2(\omega_2), \tag{73}$$

where F_1 and F_2 are one-variable noncyclic spectral densities. In particular, it is easily seen that the processes admitting finite-dimensional (FF)MR are precisely those whose spectral densities are given by (73) with F_1 and F_2 rational.

Before giving the main results of this subsection, we need to introduce the notion of structural function of a regular and coregular (FF)MR; the terminology originates from [14]. Let X be a regular and coregular FFMR, and let us denote by ϕ and ν its FF spectral factor and its FF innovation process. By Theorems 2.2 and 2.3, the space $(Z_1)_{0,-}$ can be written as $(Z_1)_{0,-}=Z_1\oplus U_2^*(\Delta_{fb})_{0,-}$. As X is regular and coregular, $(Z_1)_{0,-}=U_2^*\Delta_{0,-}$, and Δ_{fb} is one-dimensional. Let ν^{fb} be a white noise sequence for which $\Delta_{fb}=sp(\nu_{0,0}^{fb})$, that we shall call the forward-backward innovation process of X. Then, as $U_2^*(\Delta_{fb})_{0,-}$ is included in $(Z_1)_{0,-}=U_2^*\Delta_{0,-}$, there exists a (uniquely-defined) $*$-inner function π_2 for which

$$d\hat{\nu}^{fb}(\omega_1,\omega_2)=\pi_2(e^{i\omega_2})d\hat{\nu}(\omega_1,\omega_2). \tag{74}$$

Similarly, there exists a uniquely defined $*$-inner function π_1 for which

$$d\hat{\nu}^{bf}(\omega_1,\omega_2)=\pi_1(e^{i\omega_1})d\hat{\nu}(\omega_1,\omega_2), \tag{75}$$

where ν^{bf} represents the backward-forward innovation process of X. In the sequel, we shall call π_1 and π_2 the structural functions of the FFMR X. Let us derive some properties of the structural functions. First, $\phi\pi_1^*$ belongs to $\sigma_2^*H^2_{+,-}$ and $\phi\pi_2^*$ belongs to $\sigma_1^*H^2_{-,+}$. In fact, by Lemma 3.1 the space $X_{-,+}$ coincides with $U_1^*(\Delta_{fb})_{-,+}$; as $y_{0,0}$ belongs to $X_{-,+}$, there exists a function $\phi^{fb}\in\sigma_1^*H^2_{-,+}$ (that we shall call the forward-backward spectral factor of K) such that

$$d\hat{y}(\omega_1,\omega_2)=\phi^{fb}(e^{i\omega_1},e^{i\omega_2})d\hat{\nu}^{fb}(\omega_1,\omega_2). \tag{76}$$

The conclusion follows immediately from the fact that $\phi\pi_2^*=\phi^{fb}$. Similarly, $\phi\pi_1^*$ coincides with the backward-forward spectral factor ϕ^{bf} of X, so that $\phi\pi_1^*\in\sigma_2^*H^2_{+,-}$.

Let us discuss now on the functional models based on π_1 and π_2. First, it is clear that $X_{-,\infty}=Q_\nu(\sigma_1^*H^2_{-,\infty})$ and that $X_{\infty,-}=Q_\nu(\sigma_2^*H^2_{\infty,-})$. Next, by Lemma 3.1, we get that $X_{+,\infty}=Q_\nu(\pi_1H^2_{+,\infty})$ and that $X_{\infty,+}=Q_\nu(\pi_2H^2_{\infty,+})$. By Theorem 3.2, it follows that the space $\tilde{X}=X_{-,\infty}\cap X_{\infty,-}\cap X_{+,\infty}\cap X_{\infty,+}$ is a MR of Y given by

$$\tilde{X}=Q_\nu[\sigma_1^*H^2_{-,\infty}\cap\sigma_2^*H^2_{\infty,-}\cap\pi_1H^2_{+,\infty}\cap\pi_2H^2_{\infty,+}] \tag{77}$$

$$=Q_\nu[\sigma_1^*(H^2_-(T_1)\ominus\pi_1H^2_-(T_1))\otimes\sigma_2^*(H^2_-(T_2)\ominus\pi_2H^2_-(T_2))]. \tag{78}$$

But again by Theorem 3.2, X is a MR if and only if X coincides with $\tilde{X}$ so that (78) gives a functional model for the regular and coregular MR of Y. This discussion also shows that the structural functions of a FFMR X depends only on the four subspaces $X_{-,\infty},X_{\infty,-},X_{+,\infty},X_{\infty,+}$, so that π_1 and π_2 do not characterize X, except in the case where X is a MR. In particular, the derivation of a general functional model for FFMR seems to be a difficult problem. However, if X is FF observable, then it is clear that

$$X=Q_\nu(\text{Range}(\mathcal{H}_\phi)). \tag{79}$$

But, this functional model is not very informative because the structure of the range of a two-dimensional Hankel operator is more complicated than in the one-dimensional case. In fact, the space $\sigma_1^*\sigma_2^* H^2_{-,-} \ominus \text{Range}(\mathcal{H}_\phi)$ is invariant by σ_1^* and σ_2^*, but except in situations similar to that of Proposition 1.1, Beurling's theorem has no extension to the two variable functions case. Let us finally mention the fact that the second structural function π_2 of a FF observable FFMR coincide with the $*$-inner function θ_2 defined by (62); similarly, π_1 is equal to the $*$-inner function θ_1 defined as θ_2 by exchanging the role of z_1 and z_2. However, the converse is not true, i.e., if the structural functions of a certain FFMR coincide with θ_1 and θ_2, then it is not necessarily FF observable.

Now, we present the main result of this section (Theorem 3.3), i.e., an explicit spectral characterization of a wide class of minimal FFMR and MR. It is obtained under the following assumption $\mathcal{A}$ on y which is somewhat more restrictive than the existence of nontrivial MR.

$\mathcal{A}$. The process y admits at least one nontrivial (i.e., regular and coregular) MR whose structural functions are Blaschke products.

Theorem 3.3 to be presented below is based on the fact that a regular and coregular MR $\tilde{X}$ whose structural functions are Blaschke products is minimal if and only if the "one-parameter" Markovian representations $\tilde{X}_{0,\infty}$ and $\tilde{X}_{\infty,0}$ of $Y_{0,\infty}$ and of $Y_{\infty,0}$ respectively, are themselves minimal. This fundamental result does not come from purely geometrical considerations. In fact, it follows from the particular properties of the functional models associated to the minimal doubly invariant under U_2 (respectively, U_1) Markovian representations of $Y_{0,\infty}$ (respectively of $Y_{\infty,0}$) when assumption $\mathcal{A}$ holds. In order to present these properties, we begin by giving general results (i.e., not depending on assumption $\mathcal{A}$) concerning the spectral description of the doubly invariant under U_2 Markovian representations of $Y_{0,\infty}$ with respect to U_1. The results which are relative to the doubly invariant under U_1 Markovian representations of $Y_{\infty,0}$ with respect to U_2 are similar. It is worth mentioning that as $Y_{0,\infty}$ is infinite-dimensional, the spectral description of [14] cannot be directly used. However, the fact that $Y_{0,\infty}$ and the considered Markovian spaces are doubly invariant under U_2 makes possible to immediately generalize the spectral analysis of [14].

PROPOSITION 3.4. *Let us suppose that the spectral density of y satisfies the Szegö condition, and that $Y_{0,\infty}$ admits at least one regular and coregular doubly invariant under U_2 Markovian representation. Let Z be such a regular and coregular* MR *of $Y_{0,\infty}$. Then, there exists a unique (up to an all pass function of z_2) pair of spectral factors $(\phi, \overline{\phi})$ of F such that*

- $\phi \in \sigma_1^* H^2_{-,\infty}, \overline{\phi} \in H^2_{+,\infty}$;
- $\theta(z_1, z_2) = \phi(z_1, z_2)/\overline{\phi}(z_1, z_2)$ *is a $1*$-inner function, called the structural function of Z.*

The white noise sequences $\nu, \overline{\nu}$ defined by the fact that $d\hat{y}(\omega_1, \omega_2) = \phi(e^{i\omega_1}, e^{i\omega_2})$ $d\hat{\nu}(\omega_1, \omega_2) = \overline{\phi}(e^{i\omega_1}, e^{i\omega_2}) d\hat{\overline{\nu}}(\omega_1, \omega_2)$ are such that $Z_{-,0} = U_1^* \nu_{-,\infty}$ and $Z_{+,0} = \overline{\nu}_{+,\infty}$. Moreover, Z is given by

$$Z = Q_\nu[\sigma_1^*(H^2_{-,\infty} \ominus \theta H^2_{-,\infty})] = Q_{\overline{\nu}}[H^2_{+,\infty} \ominus \theta^* H^2_{+,\infty}]; \tag{80}$$

- *Z is observable (i.e., $Z = \overline{E}^Z(Y_{+,\infty})$) if and only if $\overline{\phi}$ and θ^* are 1-coprime;*
- *Z is constructible (i.e., $Z = \overline{E}^Z(Y_{-,\infty})$) if and only if ϕ and θ are 1-coprime;*
- *All minimal doubly invariant under U_2 Markovian representations of $Y_{0,\infty}$ with respect to U_1 are regular, coregular, and they have the same structural function.*

Taking Lemma A.1 (below) as a starting point, Proposition 3.4 can be proved along the lines of [14]. The details are left to the reader.

If y satisfies assumption $\mathcal{A}$, the following fundamental result holds.

LEMMA 3.4. *Let us suppose that assumption $\mathcal{A}$ holds. Then, the structural function of the minimal doubly invariant under U_2 (respectively, U_1) Markovian representations of $Y_{0,\infty}$ with respect to U_1 (respectively, of $Y_{\infty,0}$ with respect to U_2) is a one-variable function $q_1(z_1)$ (respectively, $q_2(z_2)$).*

We are now in position to give the main result of this subsection.

THEOREM 3.3. *Let $\tilde{X}$ be a regular and coregular* MR *of Y with respect to (U_1, U_2), whose structural functions π_1 and π_2 are Blaschke products. Then, $\tilde{X}$ is minimal if and only if the "one-parameter" Markovian representations $\tilde{X}_{0,\infty}$ and $\tilde{X}_{\infty,0}$ of $Y_{0,\infty}$ and of $Y_{\infty,0}$, respectively, are themselves minimal, or equivalently if and only if $\pi_1 = q_1$ and $\pi_2 = q_2$. In this case, the* FF *observable* FFMR *X given by*

$$X = \overline{E}^{\tilde{X}}(Y_{+,+}) \tag{81}$$

is a minimal FFMR.

The proofs of Lemma 3.4 and of Theorem 3.3 are given in the Appendix.

Let us discuss this result. First, it appears that the minimal MR can be characterized in a rather satisfying way. Next, Theorem 3.3 shows that all the minimal MR of Y whose structural functions are Blaschke products are unitarily equivalent. In particular, if there exist finite-dimensional MR of Y, then all the minimal finite-dimensional MR have the same dimension. The results concerning the FFMR are less powerful. Although Theorem 3.3 allows to exhibit a wide class of minimal FFMR (i.e., those that are contained in a minimal MR), the characterization problem of all minimal FFMR remains open. The difficulty of this question is undoubtly related to the fact that the functional model given by (79) is not very informative, due to the complicated structure of the ranges of two-dimensional Hankel operators. Let us also mention the fact that when there exist finite-dimensional (FF)MR, there may exist minimal FFMR whose dimension do not coincide (see below).

In order to illustrate Theorem 3.3, let us consider the following simple example.

Let us suppose that y is given by $d\hat{y}(\omega_1, \omega_2) = \phi(e^{i\omega_1}, e^{i\omega_2}) d\hat{\nu}(\omega_1, \omega_2)$, where ν is a white noise and where ϕ is the element of $\sigma_1^* \sigma_2^* H^2_{-,-}$ given by

$$\phi(z_1, z_2) = \frac{1}{(z_1 - a_1)(z_2 - a_2)} + \frac{1}{(z_1 - b_1)(z_2 - b_2)},$$
$$|a_i| < 1, |b_i| < 1, a_1 \neq a_2, b_1 \neq b_2, a_i \neq b_j, (i, j) = 1, 2.$$

Then, it is obvious that the two-dimensional space X defined by

$$X = Q_\nu(\text{Range}(\mathcal{H}_\phi)) = Q_\nu \left[sp \left(\frac{1}{(z_1 - a_1)(z_2 - a_2)}, \frac{1}{(z_1 - b_1)(z_2 - b_2)} \right) \right]$$

is a regular and coregular FF observable FFMR whose structural functions θ_1 and θ_2 are given by

$$\theta_i(z_i) = \frac{1 - a_i^* z_i}{z_i - a_i} \frac{1 - b_1^* z_i}{z_i - b_i}, \qquad i = 1, 2.$$

On the other hand, it is easily seen that θ_1 and θ_2 coincide with the structural functions q_1 and q_2 introduced in Lemma 3.4. From this, we deduce that the four dimensional space $\tilde{X} = X_{-,\infty} \cap X_{\infty,-} \cap X_{+,\infty} \cap X_{\infty,+} = Q_\nu[\sigma_1^*(H^2_-(T_1) \ominus \theta_1 H^2_-(T_1)) \otimes \sigma_2^*(H^2_-(T_2) \ominus \theta_2 H^2_-(T_2))]$ is a minimal MR, and that X, which coincides $X = \overline{E}^{\tilde{X}}(Y_{+,+})$, is a minimal FFMR of Y.

Let us now exhibit a minimal FFMR X' whose dimension is four. ϕ can be written as

$$\phi(z_1,z_2)=\frac{N(z_1,z_2)}{(z_1-a_1)(z_1-b_1)(z_2-a_2)(z_2-b_2)}.$$

Let $N'(z_1,z_2)$ be the polynomial defined by $N'(z_1,z_2)=z_1z_2(N(1/z_1^*,1/z_2^*))^*$, and put

$$\phi'(z_1,z_2)=\frac{N'(z_1,z_2)}{(z_1-a_1)(z_1-b_1)(z_2-a_2)(z_2-b_2)}.$$

Then, ϕ' belongs to $\sigma_1^*\sigma_2^*H^2_{-,-}$ and represents a spectral factor of the spectral density of y. Let us consider the white noise ν' defined by the fact that $d\hat{y}(\omega_1,\omega_2)=\phi'(e^{i\omega_1},e^{i\omega_2})\,d\hat{\nu}'(\omega_1,\omega_2)$. It is easily seen that the space $X'=Q_{\nu'}(\text{Range}(\mathcal{H}_{\phi'}))$ is a minimal FF observable FFMR of Y; but, one can check that there always exist $(a_i,b_i)_{i=1,2}$ for which $\text{Range}(\mathcal{H}_{\phi'})$ is the four dimensional space $Q_{\nu'}[\sigma_1^*(H^2_-(T_1)\ominus\theta_1H^2_-(T_1))\otimes\sigma_2^*(H^2_-(T_2)\ominus\theta_2H^2_-(T_2))]$. In this case, $\dim X'=4$, so that X and X' are two minimal FFMR whose dimensions are not equal.

More generally, a process y may admit a unique minimal dimension FFMR although there always exist more than one minimal FFMR in the geometric sense; in this particular case, the stochastic realization algorithm derived by Attasi allows the calculation of the corresponding spectral factor. When there exist more than one minimal dimension FFMR, the calculation of the parameters of a stochastic realization seems to be an open problem.

Appendix.

Proof of Theorem 2.2. As $(Z_1)_{0,-}$ is included in $X_{\infty,-}$, it is sufficient to establish that $X_{\infty,-}$ and $(Z_1)_{0,+}$ are conditionally orthogonal given Z_1 in order to demonstrate (31). For this purpose, we have to show that $E^{X_{\infty,-}}U_2^n(U_1x-T_1x)$ belongs to Z_1 for all $n\geq 0$, for all $x\in X$. $E^{X_{\infty,-}}U_2^n(U_1x-T_1x)=E^{X_{\infty,-}}U_2^nU_1x-E^{X_{\infty,-}}U_2^nT_1x=U_1E^{X_{\infty,-}}U_2^nx-E^{X_{\infty,-}}U_2^nT_1x$. But, $E^{X_{\infty,-}}U_2^nx=T_2^nx$ and $E^{X_{\infty,-}}U_2^nT_1x=T_2^nT_1x=T_1T_2^nx$. Therefore, $E^{X_{\infty,-}}U_2^n(U_1x-T_1x)=(U_1-T_1)T_2^nx$, and thus belongs to Z_1. The above equality for $n=1$ also shows that Z_1 is invariant under S_2, and that the Markovian transition operator of Z_1 coincides with $S_{2|Z_1}$.

In order to establish that Z_1 is invariant under S_2^*, let us begin by showing that S_2^*x coincides with $E^{X_{-,0}}U_2^*x$, for all $x\in X$. In fact, $S_2^*x=E^{X_{\infty,0}}U_2^*x=E^{X_{\infty,0}}E^{X_{-,\infty}}U_2^*x$. As the subspaces $X_{\infty,-}, X_{\infty,+}, X_{-,\infty}$ intersect perpendicularly two by two, and as $X_{\infty,0}=X_{\infty,-}\cap X_{\infty,+}$, S_2^*x is equal to $E^{X_{\infty,-}\cap X_{\infty,+}\cap X_{-,\infty}}U_2^*x$, which coincides with $E^{X_{-,0}}U_2^*x$ by (25). In particular, $U_1S_2^*x$ belongs to $U_1X_{-,0}$. By using the same kind of arguments, it can be shown that $S_2^*T_1x=E^{X_{-,0}}U_1S_2^*x$, for all $x\in X$. Therefore, $S_2^*(U_1x-T_1x)$ coincides with $U_1S_2^*x-E^{X_{-,0}}U_1S_2^*x$, from which we deduce that $E^{X_{-,0}}S_2^*(U_1x-T_1x)=0$. On the other hand, $S_2^*(U_1x-T_1x)$ belongs to $U_1X_{-,0}$, which implies that $S_2^*(U_1x-T_1x)$ is an element of $U_1X_{-,0}\ominus X_{-,0}$. But as this last subspace is equal to Z_1, we get that Z_1 is invariant under S_2^*.

Proof of Theorem 2.3. Let us begin by showing that the space Δ given by (32) coincides with the forward innovation space of Z_1. For all $x\in X$, $(U_1U_2-U_1T_2-U_2T_1+T_1T_2)x=U_2(U_1-T_1)x-(U_1-T_1)T_2x$. But, $U_1T_2x=S_2U_1x$ and $T_1T_2x=T_2T_1x$, so that $(U_1U_2-U_1T_2-U_2T_1+T_1T_2)x$ is equal to $(U_2-S_2)(U_1-T_1)x$. This establishes (33); (34) is shown similarly. Let us show now that Δ coincides with the wandering subspace $(U_1X_{-,\infty}\ominus X_{-,\infty})\cap(U_2X_{\infty,-}\ominus X_{\infty,-})$. For this purpose, we remark that as (19) holds, then, by Proposition 1.1, the above wandering subspace is equal to$U_2(U_1X_{-,-}\ominus X_{-,-})\ominus(U_1X_{-,-}\ominus X_{-,-})$. But, it is easily seen that $U_1X_{-,-}\ominus X_{-,-}=(Z_1)_{0,-}$, from which we deduce that $(U_1X_{-,\infty}\ominus X_{-,\infty})\cap(U_2X_{\infty,-}\ominus X_{\infty,-})=U_2(Z_1)_{0,-}\ominus(Z_1)_{0,-}=\Delta$. Equation

(35) (respectively, (36)) can be deduced in a similar way from (20) (respectively, (21)) and from Proposition 1.1 applied to the pair (U_1, U_2^*) (respectively, to the pair (U_1^*, U_2)).

Proof of Lemma 3.1. Let us begin by remarking that as $X_{0,\infty}$ is a Markovian space with respect to U_1, (49) and (51) hold if and only if [25] the operator S_1 belongs to the class $C_{0,0}$ (i.e., S_1^m and S_1^{*m} converge strongly to zero as $m \to \infty$). Equations (50) and (52) are characterized in a similar way. Let X be a regular FFMR; then, let us show that X is coregular if and only if $S_{1|Z_2}$ and $S_{2|Z_1}$ belong to the class $C_{0,0}$. The direct part is obvious. Conversely, let us assume that $S_{1|Z_2}$ and $S_{2|Z_1}$ belong to the class $C_{0,0}$. As X is regular, $\cap_{m\in Z} U_1^m X_{-,+} = \{0\}$, so that $X_{-,+} = \oplus_{k\geq 1} U_1^{*k}(U_1 X_{-,+} \ominus X_{-,+})$. But, it is easily seen that $U_1 X_{-,+} \ominus X_{-,+} = (Z_1)_{0,+}$. As $S_{2|Z_1}$ belong to the class $C_{0,0}$, $(Z_1)_{0,+} = (\Delta_{fb})_{0,+}$ [25], from which we get that $X_{-,+} = U_1^*(\Delta_{fb})_{-,+}$, and that $X_{\infty,+} = (\Delta_{fb})_{\infty,+}$; this implies (51). Similarly, the fact that $S_{1|Z_2}$ belongs to the class $C_{0,0}$ implies (52), that $X_{+,-} = U_2^*(\Delta_{bf})_{+,-}$, and that $X_{+,\infty} = (\Delta_{bf})_{+,\infty}$. The first statement of the lemma follows from the well known fact [25] that if $\dim\Delta < \infty$, then $S_{2|Z_1}$ (respectively, $S_{1|Z_2}$) belongs to the class $C_{0,0}$ if and only if $\dim\Delta = \dim\Delta_{fb}$ (respectively, $\dim\Delta = \dim\Delta_{bf}$).

Proof of Proposition 3.1. Let us suppose that (59) is satisfied. Then, $\phi_{2,k}(z_2) = \psi_{2,k}(z_2)\omega_2(z_2)$, for all $k \geq 1$, where $\psi_{2,k}$ is the element of $H_+^2(T_2)$ defined in the same fashion that $\phi_{2,k}$. From this, we deduce that the range of the Hankel operator associated to $\phi_{2,k}$ is included in $\sigma_2^*(H_-^2(T_2) \ominus \omega_2 H_-^2(T_2))$ for all $k \geq 1$, and that the space Z_1 given by (58) does not coincide with $\sigma_2^* H_-^2(T_2)$. Conversely, if ϕ is 2-weakly noncyclic, then the space $\sigma_2^*(\Delta_{fb})_{0,-}$ is a nonzero subspace of $\sigma_2^* H_-^2(T_2)$ invariant under σ_2^*. By Beurling's theorem, there exists a uniquely defined $*$-inner function θ_2 for which (62) holds. From this, we get that for all $k \geq 1$, $\phi_{2,k}(z_2)$ can be written as $\phi_{2,k}(z_2) = \overline{\phi}_k(z_2)\theta_2(z_2)$ for some element ϕ_k of $H_+^2(T_2)$. Let $\phi_{-,+}$ be the function of $\sigma_1^* H_{-,+}^2$ defined by

$$\phi_{-,+}(z_1, z_2) = \sum_{k\geq 1} \overline{\phi}_k(z_2) z_2^{-k}.$$

Then, $\phi(z_1, z_2) = \phi_{-,+}(z_1, z_2)\theta_2(z_2)$. Let us show that $\phi_{-,+}$ and θ_2^* are 2-weak coprime. For this purpose, let us assume that there exist two $*$-inner functions α and β, and an element $\psi(z_1, z_2)$ of $\sigma_1^* H_{-,+}^2$ such that $\phi_{-,+}(z_1, z_2) = \psi(z_1, z_2)\alpha^*(z_2)$ and $\theta_2(z_2) = \alpha(z_2)\beta(z_2)$. Then, $\phi(z_1, z_2)$ is equal to $\psi(z_1, z_2)\beta(z_2)$, from which we deduce that Z_1 is included in $\sigma_2^*(H_-^2(T_2) \ominus \beta H_-^2(T_2))$. But, in view of (62), this is possible if and only if $\theta_2 = \beta$. Therefore, $\phi_{-,+}$ and θ_2^* are 2-weak coprime. The fact that $(\phi_{-,+}, \theta_2)$ is the unique pair satisfying (60) and (61) is shown similarly.

Proof of Lemma 3.3. As the spectral density of y satisfies the Szegö condition, there exists a white noise sequence α for which $Y_{\infty,\infty} = Q_\alpha(L^2(T^2))$, so that $Y_{\infty,\infty}$ and $L^2(T^2)$ are unitarily equivalent. Therefore, it is sufficient to establish the lemma in the case where $(U_1, U_2, Y_{\infty,\infty})$ coincides with $(\sigma_1, \sigma_2, L^2(T^2))$. We begin by giving the following result which can be proved by using arguments similar to those of the proof of Theorem 1.2.1 of [16].

LEMMA A.1. *Let H_1 be a subspace of $L^2(T^2)$ satisfying*

- $\bigvee_{m\in Z} \sigma_1^m H_1 = L^2(T^2)$,
- $\sigma_1^* H_1 \subset H_1$,
- $\sigma_2 H_1 = H_1$.

Then, there exists an all-pass function q defined on T^2 and a uniquely defined Borel set B_1 of T such that

- $H_{1,-\infty} = \cap_{m\in Z} \sigma_1^m H_1 = 1_{T\times B_1^c} L^2(T^2)$,
- $H_1 \ominus H_{1,-\infty} = 1_{T\times B_1} q H_{-,\infty}^2$,

where 1_A stands for the characteristic function of the set A. If, moreover, $H_{1,-\infty} = \{0\}$, then q is uniquely defined up to an all-pass function of z_2.

Now, we prove Lemma 3.3. First, paralleling the arguments of §1 of [10], we get that the commutation property (44) implies that

$$H_1 \cap H_2 = \sigma_1^* \sigma_2^* \Delta_{-,-} \oplus (H_1 \ominus H_{1,-\infty}) \cap H_{2,-\infty} \oplus (H_2 \ominus H_{2,-\infty}) \cap H_{1,-\infty} \oplus H_{1,-\infty} \cap H_{2,-\infty},$$

where Δ is defined by (69) and where $H_{2,-\infty}$ denotes the space $\cap_{n\in Z}\sigma_2^n H_2$. Let us consider the case where Δ is not reduced to $\{0\}$. Then, by arguments of multiplicity theory, Δ is one-dimensional. As Δ is wandering for (σ_1, σ_2), Δ coincides with the space generated by a certain all-pass function θ defined on T^2. Therefore, the space $\sigma_1^*\sigma_2^*\Delta_{-,-}$ is equal to $\theta\sigma_1^*\sigma_2^* H^2_{-,-}$. Hence, the space $\Delta_{\infty,\infty}$ coincides with $L^2(T^2)$. But, again by using (44), it is easily seen that

$$\Delta_{\infty,\infty} = \left(\bigvee_{m\in Z} \sigma_1^m H_1 \ominus H_{1,-\infty} \right) \cap \left(\bigvee_{n\in Z} \sigma_2^n H_2 \ominus H_{2,-\infty} \right),$$

i.e., in view of (65),

$$\Delta_{\infty,\infty} = (L^2(T^2) \ominus H_{1,-\infty}) \cap (L^2(T^2) \ominus H_{2,-\infty}).$$

Consequently, if Δ is not reduced to $\{0\}$, (67) and (68) hold. If $\Delta = \{0\}$, it turns out that $(L^2(T^2) \ominus H_{1,-\infty}) \cap (L^2(T^2) \ominus H_{2,-\infty}) = \{0\}$. By Lemma A.1 applied to the pairs (H_1, σ_1) and (H_2, σ_2), there exist two Borel sets B_1 and B_2 of T such that $(L^2(T^2) \ominus H_{1,-\infty}) = 1_{T\times B_1} L^2(T^2)$ and $(L^2(T^2) \ominus H_{2,-\infty}) = 1_{B_2\times T} L^2(T^2)$; therefore,

$$(L^2(T^2) \ominus H_{1,-\infty}) \cap (L^2(T^2) \ominus H_{2,-\infty}) = 1_{B_2\times B_1} L^2(T^2).$$

Consequently, if $\Delta = \{0\}$, then one of the set B_1 or B_2 must be negligible with respect to the Lebesgue measure. This in turn shows that (66) holds.

Proof of Lemma 3.4. In order to establish Lemma 3.4, we begin by showing the following fundamental result.

LEMMA A.2. *Let $\phi(z_1, z_2)$ be a function of $H^2_{+,+}$, and let $\theta(z_1)$ be an inner Blaschke product. Then, ϕ and θ are 1-weak coprime if and only if they are 1-coprime.*

Proof. Let us suppose that ϕ and θ are 1-weak coprime, and let us assume the existence of a 1-inner function α satisfying

$$\phi(z_1, z_2) = \psi(z_1, z_2)\alpha(z_1, z_2), \theta(z_1) = \gamma(z_1, z_2)\alpha(z_1, z_2)$$

almost everywhere on T^2 for some $\psi \in H^2_{+,\infty}$ and for some 1-inner function γ. Then, it is clear that, for almost all z_2 on T, the above equality also holds for each z_1 in the open unit disk. Let a be a zero of θ; then, $|a| < 1$, and $\gamma(a, z_2)\alpha(a, z_2) = 0$ almost everywhere on T. Let us denote by E_a the Borel set of T defined by $E_a = \{z_2 \in T / \alpha(a, z_2) = 0\}$. Then, the function $\phi_a : z_2 \to \phi(a, z_2)$ is zero on E_a; as ϕ belongs to $H^2_{+,+}$, ϕ_a is an element of H^2_+; therefore, it cannot be zero on a set of positive measure [21], unless in the case where it is identically zero. If $\phi(a, z_2) = 0$ almost everywhere, then ϕ can be written as $\phi(z_1, z_2) = \phi'(z_1, z_2)(z_1 - a)/(1 - a^* z_1)$ for some function ϕ' of $H^2_{+,+}$, which contradicts the fact that ϕ and θ are 1-weak coprime. Consequently, E_a is a set of zero measure, and $\gamma(a, z_2) = 0$ almost everywhere; but, this implies that $\gamma(z_1, z_2) = \gamma'(z_1, z_2)(z_1 - a)/(1 - a^* z_1)$ for some 1-inner function γ'. Repeating this procedure for each zero of θ (by taking into account

their multiplicities), we show that $\gamma(z_1, z_2) = \theta(z_1)$ (up to an all-pass function of z_2 only). This in turn shows that ϕ and θ are 1-coprime. □

Now, we prove Lemma 3.4. First, it is easily seen that if assumption $\mathcal{A}$ holds, then there exists an FF observable regular and coregular FFMR X of Y whose structural functions are Blaschke products. Let us denote by ϕ and ν its FF spectral factor and its FF innovation process. As X is FF observable, its second structural function coincides with the function θ_2 defined by (62), and its first structural function is equal to the function θ_1 defined as θ_2 by exchanging the role of z_1 and z_2. Similarly, the forward-backward spectral factor ϕ^{fb} of X coincides with the function $\phi_{-,+}$ defined in Proposition 3.1. Therefore, by proposition 3.1, ϕ^{fb} and θ_2^* are 2-weak coprime; by a straightforward modification of Lemma A.2, they are in fact 2-coprime. Similarly, ϕ^{bf} and θ_1^* are 1-coprime. On the other hand, the structural function of the doubly invariant under U_2 Markovian representation $X_{0,\infty}$ of $Y_{0,\infty}$ coincides with θ_1; therefore, it follows from Proposition 3.4 that $X_{0,\infty}$ is observable. Similarly, $X_{\infty,0}$ is observable. Now, again by using Lemma A.2, we are going to show that the structural function of the minimal Markovian representation $\overline{E}^{X_{0,\infty}}(Y_{-,\infty})$ of $Y_{0,\infty}$ is a function of z_1 only; this will demonstrate the first part of Lemma 3.4. ϕ^{bf} can be written as $\phi^{bf}(z_1, z_2) = \phi(z_1, z_2)\theta_1^*(z_1)$. By a modification of Proposition 3.1, ϕ^{bf} can be written as $\phi^{bf}(z_1, z_2) = \phi_1(z_1, z_2)q_1^*(z_1)$, where $\phi_1 \in \sigma_1^*\sigma_2^* H^2_{-,-}$ and where q_1 is a $*$-inner function such that ϕ_1 and q_1 are 1-weak coprime; moreover, there exists a $*$-inner function α_1 for which $\phi(z_1, z_2) = \phi_1(z_1, z_2)\alpha_1(z_1), \theta_1(z_1) = q_1(z_1)\alpha_1(z_1)$. As ϕ_1 and q_1 are 1-weak coprime, they are 1-coprime by Lemma A.2. On the other hand, as q_1 is a divisor of θ_1, ϕ^{bf} and q_1 are also 1-coprime. Let us denote by ν_1 the white noise defined by

$$d\hat{\nu}_1(\omega_1, \omega_2) = \alpha_1(e^{i\omega_1})d\hat{\nu}(\omega_1, \omega_2).$$

Then, it is easily seen that the subspace $Q_{\nu_1}[\sigma_1^*(H^2_{-,\infty} \ominus q_1 H^2_{-,\infty})]$ is a Markovian representation of $Y_{0,\infty}$ included in $X_{0,\infty}$, whose forward and backward spectral factors are equal to ϕ_1 and ϕ_{bf}, respectively, and whose structural function is the one-variable function $q_1(z_1)$. By Proposition 3.4, it turns out that this Markovian representation is minimal, and that it coincides with the constructible projection $\overline{E}^{X_{0,\infty}}(Y_{-,\infty})$ of $X_{0,\infty}$. This establishes the first part of the lemma.

Similarly, there exists two $*$-inner functions q_2 and α_2, and an element ϕ_2 of $\sigma_1^*\sigma_2^* H^2_{-,-}$ such that $\phi(z_1, z_2) = \phi_2(z_1, z_2)\alpha_2(z_2), \theta_2(z_2) = q_2(z_2)\alpha_2(z_2), \phi^{fb}(z_1, z_2) = \phi_2(z_1, z_2)$ $q_2^*(z_2)$, where ϕ^{fb} and q_2^*, and ϕ_2 and q_2 are 2-coprime. As previously, if we denote by ν_2 the white noise defined by the fact that

$$d\hat{\nu}_2(\omega_1, \omega_2) = \alpha_2(e^{i\omega_2})d\hat{\nu}(\omega_1, \omega_2),$$

then the subspace $Q_{\nu_2}[\sigma_2^*(H^2_{\infty,-} \ominus q_2 H^2_{\infty,-})]$ is a minimal doubly invariant under U_1 Markovian representation of $Y_{\infty,0}$ which coincides with $\overline{E}^{X_{\infty,0}}(Y_{\infty,-})$. The second part of the lemma follows from the fact that its structural function is the one-variable function $q_2(z_2)$.

Proof of Theorem 3.3. Let us begin by showing that if $\tilde{X}$ is a MR of Y for which $\tilde{X}_{0,\infty}$ and $\tilde{X}_{\infty,0}$ are minimal Markovian representations of $Y_{0,\infty}$ and of $Y_{\infty,0}$, then $\tilde{X}$ is a minimal MR of Y. For this purpose, let suppose that $\tilde{X}$ contains another MR $\tilde{X}'$ of Y. Then, $\tilde{X}'_{0,\infty} \subset \tilde{X}_{0,\infty}$ and $\tilde{X}'_{\infty,0} \subset \tilde{X}_{\infty,0}$; by the minimality of $\tilde{X}_{0,\infty}$ and $\tilde{X}_{\infty,0}$, this implies that $\tilde{X}'_{0,\infty} = \tilde{X}_{0,\infty}$ and $\tilde{X}'_{\infty,0} = \tilde{X}_{\infty,0}$. The equality $\tilde{X}' = \tilde{X}$ follows from the fact that $\tilde{X}' = \tilde{X}'_{0,\infty} \cap \tilde{X}'_{\infty,0}$ and that $\tilde{X} = \tilde{X}_{0,\infty} \cap \tilde{X}_{\infty,0}$.

Conversely, let $\tilde{X}$ be a regular and coregular MR of Y whose structural functions are Blaschke products. We have to show that it is possible to construct a MR of Y, say $\tilde{X}'$,

included in $\tilde{X}$, and for which $\tilde{X}'_{0,\infty}$ and $\tilde{X}'_{\infty,0}$ are minimal Markovian representations of $Y_{0,\infty}$ and of $Y_{\infty,0}$. Let ϕ and ν be the FF spectral factor and the FF innovation process of $\tilde{X}$, respectively. Then, there is no restriction to assume that the structural functions of $\tilde{X}$ coincide with the functions θ_1 and θ_2 defined by (62). In this case, all the reduction procedure presented in the proof of Lemma 3.4 can be applied to $(\phi, \nu, \theta_1, \theta_2)$. Let ν' be the white noise sequence defined by

$$\begin{aligned} d\hat{\nu}'(\omega_1,\omega_2) &= \alpha_1(e^{i\omega_1})\alpha_2(e^{i\omega_2})d\hat{\nu}(\omega_1,\omega_2) \\ &= \alpha_1(e^{i\omega_1})d\hat{\nu}_2(\omega_1,\omega_2) = \alpha_2(e^{i\omega_2})d\hat{\nu}_1(\omega_1,\omega_2). \end{aligned}$$

Put $\phi'(z_1,z_2) = \phi(z_1,z_2)\alpha_1^*(z_1)\alpha_2^*(z_2)$. Then, $\phi'(z_1,z_2)$ coincides with $\phi_2(z_1,z_2)\alpha_1^*(z_1)$; as ϕ_2 is an element of $\sigma_1^*\sigma_2^*H^2_{-,-}$, it appears that ϕ' belongs to $\sigma_2^*H^2_{\infty,-}$. Similarly, the fact that $\phi'(z_1,z_2) = \phi_1(z_1,z_2)\alpha_2^*(z_2)$ implies that $\phi' \in \sigma_1^*H^2_{-,\infty}$; finally, ϕ' is an element of $\sigma_1^*\sigma_2^*H^2_{-,-}$. Moreover, it is easily seen that $\phi' q_1^* \in \sigma_2^*H^2_{+,-}$, $\phi' q_2^* \in \sigma_1^*H^2_{-,+}$, and that $d\hat{y}(\omega_1,\omega_2) = \phi'(e^{i\omega_1}, e^{i\omega_2})d\hat{\nu}'(\omega_1,\omega_2)$. From this, one can deduce that the space $\tilde{X}'$ given by

$$\tilde{X}' = Q_{\nu'}[\sigma_1^*(H^2_-(T_1) \ominus q_1H^2_-(T_1)) \otimes \sigma_2^*(H^2_-(T_2) \ominus q_2H^2_-(T_2))]$$

is a MR of Y. Moreover, $\tilde{X}'_{0,\infty} = Q_{\nu'}[\sigma_1^*(H^2_{-,\infty} \ominus q_1H^2_{-,\infty})]$; but, this space coincides with $Q_{\nu_1}[\sigma_1^*(H^2_{-,\infty} \ominus q_1H^2_{-,\infty})]$, from which we deduce that $\tilde{X}'_{0,\infty}$ is a minimal Markovian representation of $Y_{0,\infty}$ with respect to U_1. The minimality of $\tilde{X}'_{\infty,0}$ is shown similarly.

Finally, let us show the last assumption of the theorem. Let $\tilde{X}$ be a minimal MR of Y whose structural functions are Blaschke products, and put $X = \overline{E}^{\tilde{X}}(Y_{+,+})$. Let us suppose that X contains a FFMR X'. Then, $X'_{0,\infty} \subset X_{0,\infty}$, so that the minimality of $X_{0,\infty}$ implies that $X'_{0,\infty} = X_{0,\infty}$; similarly, $X'_{\infty,0} = X_{\infty,0}$. Hence, $X'_{-,\infty} = X_{-,\infty}$ and $X'_{\infty,-} = X_{\infty,-}$, and

$$\overline{E}^{X'_{-,\infty} \cap X'_{\infty,-}}(Y_{+,+}) = \overline{E}^{X_{-,\infty} \cap X_{\infty,-}}(Y_{+,+}).$$

But, the right-hand side of the above equality is equal to X, and the left-hand side is included in X'. As it has been supposed that $X' \subset X$, this implies that $X = X'$, and that X is a minimal FFMR.

Acknowledgments. The author thanks Prof. H. Korezlioglu for his constant support during this work, and the referees for suggestions that helped to improve the presentation of this paper.

REFERENCES

[1] S. Attasi, *Modelling and recursive estimation for doubled indexed sequence*, in System Identification: Advances and Studies, R. K. Mehra and D. G. Lainiotis, eds., Mathematics in Science and Engineering, vol. 126, 1976, pp. 289–348.

[2] P. Dewilde, *Input-output description of roomy systems*, SIAM J. Control Optm., 14 (1976), pp. 712–736.

[3] R. G. Douglas, H. S. Shapiro, and A. L. Shields, *Cyclic vectors and invariant subspaces for the backward shift operator*, Ann. Inst. Fourier, Grenoble, 20 (1971), pp. 37–76.

[4] C. Foias and A. E. Frazho, *A note on unitary dilation and state spaces*, Acta Sci. Math., 45 (1983), pp. 165–175.

[5] E. Fornasini and G. Marchesini, *State space realization theory of two-dimensional filters*, IEEE. Trans. Automat. Control, 21 (1976), pp. 484–491.

[6] P. A. Furhmann, *Linear Operators and Systems in Hilbert Spaces*, McGraw-Hill, New York, 1981.

[7] I. Halperin, *Intrinsic description of the Sz–Nagy–Brehmer dilation*, Studia Math., 22 (1963), pp. 211–219.

[8] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.
[9] H. HELSON AND D. LOWDENSLAGER, *Prediction theory and Fourier series in several variables* I, Acta Math., 106 (1959), pp. 165–202.
[10] G. KALLIANPUR AND V. MANDREKAR, *Non deterministic random fields and Wold and Halmos decompositions for commuting isometries*, in Prediction Theory and Harmonic Analysis, Pesi Masani Volume, V. Mandrekar and H. Salehi, eds., North-Holland, Amsterdam, 1983, pp. 165–190.
[11] H. KOREZLIOGLU, *Two-parameter discrete wide-sense Markov processes and their recursive linear filtering*, Proc. MTNS 79, Delft, The Netherlands, July 3–6, 1979, pp. 481–487.
[12] H. KOREZLIOGLU AND P. LOUBATON, *Spectral factorization of wide-sense stationary processes on* Z^2, J. Multivariate Anal., 19 (1986), pp. 26–47.
[13] F. LI AND R. J. VACCARO, *On frequency-wavenumber estimation by state space realization*, IEEE. Trans. Circuit Systems, 38 (1991), pp. 800–804.
[14] A. LINDQUIST AND G. PICCI, *Realization theory of multivariate stationary Gaussian processes*, SIAM J. Control Optm., 6 (1985), pp. 809–857.
[15] P. LOUBATON, *Some properties of the regular minimal unitary dilation of a pair of contractive operators. Application to the quarter-plane Markovian representation problem of stationary processes on* Z^2, in Progress in Systems and Control Theory 5, Proc. of MTNS-89, Vol III, 1990, pp. 131–140.
[16] ———, *A regularity criterion for lexicographical prediction of multivariate wide-sense stationary processes on* Z^2 *with non-full rank spectral densities*, J. Function. Anal., 104 (1992), pp. 198–228.
[17] J. J. MURRAY, *Spectral factorization and quarter-plane digital filters*, IEEE. Trans. Circuit Systems, 25 (1978), pp. 586–592.
[18] D. K. PICKARD, *Unilateral Markov fields*, Adv. Applied Problems, 12 (1980), pp. 655–671.
[19] G. RUCKEBUSCH, *Théorie géometrique de la représentation markovienne*, Ann. Inst. Henri Poincaré, 16 (1980), pp. 225–297.
[20] W. RUDIN, *Function Theory on Polydiscs*, Benjamin, Amsterdam, 1969.
[21] ———, *Real and Complex Analysis*, McGraw-Hill, New York, 1970.
[22] M. SLOCINSKI, *On the Wold-type decomposition of commuting isometries*, Ann. Polon. Math., 37 (1980), pp. 255–262.
[23] M. SLOCINSKI, *Models for doubly commuting contractions*, Ann. Polon. Math., 45 (1985), pp. 23–42.
[24] A. R. SOLTANI, *Extrapolation and moving average representation for stationary random fields and Beurling's theorem*, Ann. Problems, 12 (1984), pp. 120–132.
[25] B. SZ-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, Budapest, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 358–386, March 1994

003

LINEARIZED CONTROL SYSTEMS AND APPLICATIONS TO SMOOTH STABILIZATION*

JEAN-MICHEL CORON†

Abstract. For a control system $\dot{x} = f(x, u)$, the author proves that, for generic feedback laws u such that $f(x, u(x))$ does not vanish, the linearized control systems around the trajectories of $\dot{x} = f(x, u(x))$ have the same strong accessibility algebra as f. Applications are given to the smooth stabilization problem.

Key words. nonlinear control systems, linearized control system, accessibility algebra, asymptotic stabilization

AMS subject classifications. 93D15, 93C10

Introduction. In two previous papers ([C1], [C2]) we showed that in order to stabilize asymptotically a nonlinear control system, it is sometimes useful to generate control laws which may depend—smoothly—on time, state, or initial data such that the strong accessibility algebras of the linearized control systems, around the trajectories of the nonlinear system obtained by using these control laws, are as large as possible, i.e., are equal to the strong accessibility algebra of the nonlinear control system at each point and time. If a control law has such a property we will say that it *saturates* the nonlinear system or that it is *saturating*. Such control laws can be perturbed in a suitable way in order to allow interesting local modifications of the trajectories. For example, if a saturating feedback law stabilizes the nonlinear control system, but not asymptotically, and if the strong accessibility algebra of the nonlinear system is large enough (in particular if it is equal to the tangent space at each point) then we can perturb slightly the feedback law in such a way that the new feedback law stabilizes *asymptotically* the nonlinear system. This is the well-known Jurdjevic–Quinn theorem [JQ]. This is applied in [C1]: The main idea of [C1] is to prove, for nonlinear systems without drift, the existence of saturating periodic time-varying feedback laws which stabilize, but not asymptotically, the nonlinear system; the existence of such feedback laws implies that any nonlinear system without drift which satisfies the accessibility rank condition can be *asymptotically* stabilized by means of periodic time-varying feedback laws. In [C2] we use the fact that, given an embedded curve $\mathcal{C}$ in the state space, any saturating open loop control depending smoothly on the initial data and on time can be modified slightly in such a way that, if the nonlinear control system satisfies the strong accessibility rank condition and if the dimension of the state space is at least four, with the new control the curve at any time is still embedded. This embedding property allows us to transform, along the trajectories starting on Σ, the open loop control law into a time-varying feedback control. This is important for the stabilization problem (see [C2] for more details).

In [C2] we briefly sketched the main part of a proof (relying partly on [C1]) that generic control laws depending smoothly on time and on the initial data saturate the nonlinear control system (if the strong accessibility algebra has constant rank). We give here the details of this proof (and we will see that, in fact, the constant rank hypothesis is not needed). Moreover we obtain the same result for feedback laws: Generic feedback laws such that the closed loop control system has no singular points in a fixed open set saturate the system on this open set.

* Received by the editors March 2, 1992; accepted for publication (in revised form) August 4, 1992.
† Université Paris-Sud, Laboratoire d'Analyse Numérique, Bâtiment 425, 91405 Orsay, France.

Let us mention that our results are connected to the prior works [S1] and [G]. In [S1] Sontag showed that if a system is completely controllable then any two points can be joined by means of a control law such that the linearized control system around the associated trajectory is controllable. In [G, §2.3.8.E, Thm., p. 156] Gromov showed that generic underdetermined linear (partial) differential equations are algebraically solvable; saturation, when the strong accessibility rank condition is satisfied, implies (and is in fact equivalent in the analytic case) to the algebraic solvability of the linearized control systems (see [INS] or [G, §2.3.8.(B)]). In our situation the linear differential equations are not generic; only the controls are generic, but this will be sufficient to get the result.

Recently Sontag obtained, as a consequence of an important result on observability due to Sussmann [S3], the following result: If the nonlinear analytic system $\dot{x} = f(x, u)$ satisfies the usual strong accessibility rank condition, then for any generic control law u in $C^\infty([0, T])$ the linearized control systems around the trajectories of $\dot{x} = f(x, u(t))$ are controllable on $[0, T]$. The novelty of this result compared to [C2, §2] is that now the control laws do not depend on the initial data. However, the method sketched in [C2] and that we present here in detail allows us to get this result without using [S3]; it also allows two slight improvements: We can assume f to be only C^∞—we need in this case to modify the definition of the strong accessibility rank condition in a natural way—and the linearized control systems can be required to be controllable with impulsive controls for all time in $[0, T]$. Let us remark that controllability with impulsive controls of the linearized control systems is important to get the embedding property mentioned above and used in [C2].

We also give some direct applications of our results on the genericity of saturating feedback laws to the asymptotic stabilization problem. Finally, we give straightforward modifications of our main proof in order to obtain results on observability spaces and codistributions instead of accessibility algebras.

1. Definition and statements of the main theorem and corollaries. Throughout this paper "manifold" always means finite-dimensional Hausdorff, second countable manifold of class C^∞. Unless otherwise specified the manifolds have no boundary. For two manifolds V and W, and for p in $\mathbb{N} \cup \{\infty\}$, $C^p(V; W)$ denotes the set of maps from V into W which are of class C^p; for p in $\mathbb{N}$, this set is equipped with the (fine) Whitney C^p-topology (see, e.g., [GG, p. 42]) called, for simplicity, C^p-topology. On $C^\infty(V, W)$ we define a topology, called the C^∞-topology, in the following way. For an integer k, let $J^k(V, W)$ be the set of k-jets of C^∞-mappings from V into W. Let $(K_i, i \in \mathbb{N})$ be a sequence of compact subsets of V such that $K_0 = \emptyset$, $K_i \subset K_{i+1}$ for all integer i, and $\cup_{i \in \mathbb{N}} K_i = V$. For a sequence $k = (k_i; i \in \mathbb{N})$ of integers and for a sequence $U = (U_i; i \in \mathbb{N})$ where U_i is an open subset of $J^{k_i}(V, W)$ for all integer i, we consider the set $\mathcal{O}(k, U)$ of u in $C^\infty(V; W)$ such that $j^{k_i} u(V \setminus K_i) \subset U_i$ for all integers i. Our C^∞-topology is the topology whose basis is the family of set $\mathcal{O}(k, U)$ where k and U are as above. This topology is independent of the choice of $(K_i; i \in \mathbb{N})$ and is finer than the Whitney C^∞-topology if V is not compact; for example, $\{u \in C^\infty((0, +\infty); \mathbb{R}); \lim_{x \to 0} u^{(i)}(x) = 0 \text{ for all } i \in \mathbb{N}\}$ is an open set for our topology, but is not open for the Whitney C^∞-topology. Note also that $C^\infty(V; W)$ with our topology, as $C^\infty(V; W)$ with the Whitney C^∞-topology, is a Baire space (the proof is similar to the proof of [GG, Prop. II.3.3]). Let us mention that Theorem 1.3 following holds also for the Whitney C^∞-topology, but our C^∞-topology is slightly more convenient for the applications to the stabilization problem.

Let N be a manifold, TN its tangent bundle, m a positive integer, U an open subset of $\mathbb{R}^m$. We denote by $C^\infty_U(TN)$ the set of f in $C^\infty(N \times U; TN)$ such that

$$f(x, u) \in T_x N \quad \text{for all} \quad (x, u) \quad \text{in } N \times U. \tag{1.1}$$

For two elements f_1 and f_2 of $C^\infty_U(TN)$ we define the Lie bracket $[f_1, f_2] \in C^\infty_U(TN)$ by

$$[f_1, f_2](x,u) = [f_1(\cdot,u), f_2(\cdot,u)](x) \tag{1.2}$$

where, in the right-hand side of (1.2), $[\cdot,\cdot]$ denotes the usual Lie bracket of tangent vector fields on N.

Let f be in $C^\infty_U(TN)$. We define the strong jet accessibility algebra of f by the following definition.

DEFINITION 1.1. The strong jet accessibility algebra of f is the vectorial subspace $\mathcal{A}(=\mathcal{A}(f))$ of the vectorial space $C^\infty_U(TN)$ defined by

$$\mathcal{A} = \text{Span}\ \{\{\partial^{|\alpha|} f/\partial u^\alpha; \alpha \in \mathbb{N}^m, \alpha \neq 0\} \cup Br_2\{\partial^{|\alpha|} f/\partial u^\alpha; \alpha \in \mathbb{N}^m\}\} \tag{1.3}$$

where, for a family $\mathcal{F} \subset C^\infty_U(TN)$, $Br_2(\mathcal{F})$ denotes the set of iterated Lie brackets of elements in $\mathcal{F}$ of length at least two. For example, $\partial f/\partial u^i$, $[f, \partial f/\partial u^i]$, and $\partial^2 f/\partial u^i \partial u^j$ are in $\mathcal{A}$. Let us remark that the strong jet accessibility algebra differs slightly from the classical strong accessibility algebra $\mathcal{L}_0$ (see, e.g., the definitions in [S1, p. 549] and [SJ, p. 101]). Note that

$$\{g(x,u); g \in \mathcal{A}\} \subset \{g(x); g \in \mathcal{L}_0\} = \mathcal{L}_0(x) \quad \forall x \in N, \quad \forall u \in U \tag{1.4}$$

and that the inclusions in (1.4) are equalities if, for example, f is a polynomial with respect to u (e.g., the classical affine case $f(x,u) = f_0(x) + \sum_{i=1}^m u_i f_i(x)$) or if N and f are analytic and U is connected. Note that $\{g(x,u); g \in \mathcal{A}\}$ depends only on the jet (of order ∞) of f at (x,u); this is the reason for our terminology.

For (x,u) in $N \times U$, let

$$a(x,u) = \{g(x,u); g \in \mathcal{A}\} \subset T_x N. \tag{1.5}$$

Let us remark that, if (1.4) is an equality for all u in U, then

$$a(x,u^1) = a(x,u^2) \quad \text{for all} \quad (u^1,u^2) \quad \text{in } U \times U. \tag{1.6}$$

Let x be in N and u be a smooth map, with values into U, defined on a neighborhood of x. Let $f_0(y) = f(y,u(y)) \in T_y N$ and, for $i \in [1,m]$, let $f_i(y) = \partial f/\partial u_i(y,u(y))$. We define $a_\ell(x;u) \subset T_x N$ by

$$a_\ell(x;u) = \text{Span}\{ad^k_{f_0}(f_i)(x), k \geq 0, i \in [1,m]\}, \tag{1.7}$$

with, as usual, $ad^0_{f_0}(f_i) = f_i$ and $ad^k_{f_0}(f_i) = [f_0, ad^{k-1}_{f_0}(f_i)]$. Let us remark that $a_\ell(x;u)$ can be interpreted in the following way. Let γ, defined on an open interval of $\mathbb{R}$ containing 0 with values in N, be such that

$$\dot\gamma(t) = f(\gamma(t), u(\gamma(t))), \tag{1.8}$$

$$\gamma(0) = x. \tag{1.9}$$

The linearized control system around γ is the time-varying linear system

$$\dot z = A(t)z + B(t)w, \tag{1.10}$$

with

$$A(t) = \frac{\partial f}{\partial x}(\gamma(t), u(\gamma(t))), \tag{1.11}$$

$$B(t)w = \sum_{i=1}^m w_i \frac{\partial f}{\partial u_i}(\gamma(t), u(\gamma(t))), \tag{1.12}$$

and where $w \in \mathbb{R}^m$ is the control and $z(t) \in T_{\gamma(t)}N$ is the state. Then we easily check that, with obvious notations,

$$a_\ell(x;u) = \text{Span}\left\{\left(\left(\frac{d}{dt} - A(t)\right)^i B(t)\right)_{t=0} w; w \in \mathbb{R}^m, i \geq 0\right\}. \tag{1.13}$$

The right-hand side of (1.13) is just the classical strong accessibility algebra, evaluated at $t = 0$, of the time-varying linear control system (1.10).

We introduce the following definition.

DEFINITION 1.2. A control $u \in C^\infty(N;U)$ saturates f at x if

$$a_\ell(x;u) = a(x,u(x)). \tag{1.14}$$

Moreover u saturates f on a subset $S \subset N$ if it saturates f at all points of S.

Let us remark that we always have

$$a_\ell(x;u) \subset a(x,u(x)) \tag{1.15}$$

and that, if $a(x,u(x)) = T_xN$, then (1.14) is equivalent to the controllability with "impulsive controls" at time 0 of (1.10) (see, e.g., [KAI; p. 614]).

Let Y be a manifold, let h be in $C^\infty(N;Y)$, and let

$$\Omega \subset \{u \in C^\infty(Y;U); h'(x)(f(x,u \circ h(x))) \neq 0 \forall x \in N\}. \tag{1.16}$$

Then we will provide proof in §2.

THEOREM 1.3. *Assume that, for the C^∞-topology,*

$$\Omega \text{ is open.} \tag{1.17}$$

Then the set of u in Ω such that $u \circ h$ saturates f on N is residual in Ω (for the C^∞-topology).

Let us recall that a residual set is the countable intersection of open dense subsets. Since, if (1.17) holds, Ω is a Baire space, the set of u in Ω such that $u \circ h$ saturates f on N is dense in Ω (for the C^∞-topology). Let us also remark that, if in Definition 1.2 we replace $a(x,u)$ by $\mathcal{L}_0(x)$, then Theorem 1.3 is wrong, e.g., $N = U = Y = \mathbb{R}, h(x) = x, f(x,u) = 1 + \exp(-1/u^2)$, and $\Omega = C^\infty(\mathbb{R},\mathbb{R})$: With $\mathcal{L}_0(x)$ instead of $a(x,u)$ in Definition 1.2 $u \in C^\infty(\mathbb{R},\mathbb{R})$ saturates f if and only if it does not vanish, and such maps are not dense in $C^\infty(\mathbb{R};\mathbb{R})$.

We may wonder if the set of u in $C^\infty(N;U)$, which saturates f on N, is residual in $C^\infty(N;U)$. The answer is no, in general. Let us give an example.

Example 1.4. Let $N = \mathbb{R}^3, U = \mathbb{R}^2$, and

$$f(x,u) = (u_1, u_2, x_1u_2 - x_2u_1). \tag{1.18}$$

We easily check that $a(x,u) = T_xN = \mathbb{R}^3$, for all x in $\mathbb{R}^3$ and all u in $\mathbb{R}$, and that u in $C^\infty(\mathbb{R}^3;\mathbb{R}^2)$ saturates f at x if and only if

$$u(x) \neq 0. \tag{1.19}$$

But the set of $C^\infty(\mathbb{R}^3;\mathbb{R}^2)$, which do not vanish on $\mathbb{R}^3$, is not dense in $C^\infty(\mathbb{R}^3;\mathbb{R}^2)$.

For $r > 0$, let

$$B'_r = \{x \in \mathbb{R}^n; 0 < |x| < r\}. \tag{1.20}$$

In §3 we will obtain the following corollary of Theorem 1.3.

COROLLARY 1.5. *Assume* $N = \mathbb{R}^n, 0 \in U$, *and* $f(0,0) = 0$. *Assume also that* $\dot{x} = f(x,u)$ *can be globally (respectively, locally) asymptotically stabilized by means of a continuous feedback law. Then it can be globally (respectively, locally) asymptotically stabilized by means of a continuous feedback law* $x \to u_0(x), C^\infty$ *on* $\mathbb{R}^n\backslash\{0\}$ *(respectively,* B'_r *for some* $r > 0$*) which saturates* f *on* $\mathbb{R}^n\backslash\{0\}$ *(respectively,* B'_r*); in particular if* $a(x,u) = \mathbb{R}^n$ *for all* (x,u) *in* $(\mathbb{R}^n\backslash\{0\}) \times U$ *then the linearized control systems around the trajectories of* $\dot{x} = f(x, u_0(x))$ *are controllable with impulsive controls at each time* t *such that* $x(t) \neq 0$ *(respectively,* $x(t) \in B'_r$*).*

Let us mention that this corollary is related to a previous result proved by Sontag in [S1]. There it is proved, in particular, that if $\dot{x} = f(x,u)$ is completely controllable and satisfies $a(x,u) = T_x(N)$ on $N \times U$, then any two points of N can be joined by a trajectory of $\dot{x} = f(x,u(t))$ such that the linearized control system around this trajectory is controllable. Our corollary can be viewed as a "stabilization" version of this controllability result.

Our next corollary of Theorem 1.3 concerns systems on $N = \mathbb{R}^n$ such that

$$0 \in U, \tag{1.21}$$

$$f(0,0) = 0, \tag{1.22}$$

and

$$f(x,u) \cdot \nabla V(x) \leq 0 \quad \text{for all } (x,u) \quad \text{in } \mathbb{R}^n \times U, \tag{1.23}$$

where $V \in C^\infty(\mathbb{R}^n; [0, +\infty))$ satisfies

$$V(x) = 0 \Longleftrightarrow x = 0 \tag{1.24}$$

and

$$\lim_{|x| \to +\infty} V(x) = +\infty. \tag{1.25}$$

Then we have the following statement which is proved in §3.

COROLLARY 1.6. *Assume that* (1.21), (1.22), (1.23), (1.24), *and* (1.25), *hold. Assume that*

$$f(x,0) \neq 0 \quad \textit{for all } x \textit{ in } \mathbb{R}^n\backslash\{0\}, \tag{1.26}$$

and that, for all x *in* $\mathbb{R}^n\backslash\{0\}$, *there exists* g *in* $\mathcal{A} \cup \{f\}$ *such that*

$$g(x,0) \cdot \nabla V(x) \neq 0. \tag{1.27}$$

Then $\dot{x} = f(x,u)$ *can be globally asymptotically stabilized by means of a feedback law of class* C^∞.

Let us give an application of Corollary 1.6.

Example 1.7. Let $V \in C^\infty(\mathbb{R}^n; [0, +\infty))$, satisfying (1.24) and (1.25). Let X_0, X_1, X_2 be three vector fields on $\mathbb{R}^n$ of class C^∞. Assume that

$$L_{X_i} V = 0 \quad \forall i \in \{0,1\}. \tag{1.28}$$

Assume also that

$$X_0(0) = 0 \tag{1.29}$$

and that

$$L_{X_2}V(x) \neq 0 \quad \text{if} \quad x \in \mathbb{R}^n\backslash\{0\} \quad \text{and} \quad X_0(x) = 0. \tag{1.30}$$

Let $\mathcal{F}$ be the set of iterated Lie brackets of the vector fields X_0 and X_1. Assume that, for all x in $\mathbb{R}^n\backslash\{0\}$ such that $L_{X_0}V(x) = 0$, there exist X in $\mathcal{F}$ and k in $\mathbb{N}$ such that

$$L_{ad_X^k(X_2)}V(x) \neq 0. \tag{1.31}$$

Then $\dot{x} = X_0 + u_1X_1 + u_2X_2$ can be globally asymptotically stabilized by means of a feedback law of class C^∞. Indeed, let

$$f(x,u) = X_0(x) - L_{X_2}V(x)X_2(x) + u_1^2X_1(x) - u_2^2(L_{X_2}V(x))X_2(x); \tag{1.32}$$

then (1.22), (1.23), and (1.26) are satisfied. Let us remark that if X in $\mathcal{F}$, x in $\mathbb{R}^n$, and k in $\mathbb{N}$ satisfy $L_{ad_X^k(X_2)}V(x) \neq 0$ and $L_{ad_X^i(X_2)}V(x) = 0$ for all $i \in [0, k-1]$, then $L_{ad_X^{2k}((L_{X_2}V)X_2)}V(x) \neq 0$; then it follows easily from (1.31) that (1.27) holds. Note that a direct application to our situation of [JQ] would require that (1.31) hold with $X = X_0$.

Our next corollaries of Theorem 3.1 concern time-varying control systems (but also give nontrivial information for time-independent control systems). Therefore, now $f \in C^\infty(N \times I \times U; TN)$ with

$$f(x,t,u) \in T_xN \quad \text{for all } (x,t,u) \quad \text{in } N \times I \times U \tag{1.33}$$

where I is an open subset of $\mathbb{R}$. Associated with f is the time-independent system on $N \times I$

$$\Sigma : \dot{x} = f(x,\tau,u), \qquad \dot{\tau} = 1. \tag{1.34}$$

Let a^Σ and a_ℓ^Σ be the corresponding maps for system Σ. Let us remark that, for any (x,τ,u) in $N \times I \times U$,

$$a^\Sigma((x,\tau),u) \subset T_xN \times \{0\}, \tag{1.35}$$

where we have identified $T_{(x,\tau)}(N \times I)$ with $T_xN \times T_\tau I$. Let us define, for (x,τ,u) in $N \times I \times U$,

$$a(x,\tau,u) = \{X; (X,0) \in a^\Sigma((x,\tau),u)\} \tag{1.36}$$

and, for u in $C^\infty(N \times I; U)$,

$$a_\ell(x,\tau;u) = \{X; (X,0) \in a_\ell^\Sigma((x,\tau);u)\}. \tag{1.37}$$

For example, $\partial f/\partial u_i, \partial^2 f/\partial t \partial u_i + [f, \partial f/\partial u_i], \partial^2 f/\partial u_i \partial u_j$ evaluated at (x,τ,u) are in $a(x,\tau,u)$. Of course, if f does not depend on t (and u does not depend on t) then the new $a(x,t,u)$ (and the new $a_\ell(x,t;u)$) coincides with the previous $a(x,u)$ (and the previous $a_\ell(x;u)$). For $S \subset N \times I$ we will say that $u \in C^\infty(N \times I; U)$ saturates f on S if

$$a_\ell(x,t;u) = a(x,t,u(x,t)) \quad \text{for all} \quad (x,t) \quad \text{in } S. \tag{1.38}$$

Then our next corollary is the following.

COROLLARY 1.8. *The set of u in $C^\infty(N \times I; U)$ which saturate f on $N \times I$ is residual in $C^\infty(N \times I; U)$. The set of u in $C^\infty(I; U)$ which saturate f on $N \times I$ is residual in $C^\infty(I; U)$.*

Proof. Apply Theorem 1.3 to system Σ with $Y = N \times I$ (respectively, $Y = I$), $h(x,t) = (x,t)$ (respectively, $h(x,t) = t$), and $\Omega = C^\infty(Y; U)$ (let us note that the vector $h'(x,t)(f(x,t,u(h(x,t))), \partial/\partial t)$ never vanishes).

Let us remark that, in our previous work [C2, Thm. 2.1; Remark 2.2], we stated (with a sketch of proof) an open loop version of the first part of Corollary 1.8. This open loop version can be derived by applying Theorem 1.3 to the system on $N \times (0,T) \times \Lambda\, \dot{x} = g(x,u,\lambda), \dot{\tau} = 1, \dot{\lambda} = 0$ with $Y = (0,T) \times \Lambda$ and $h(x,\tau,\lambda) = (\tau,\lambda)$. Let us remark also that the second part of this corollary is strongly related to a result due to Sontag: In [S2] he proves that a consequence of [S3] is that, if $f \in C^\infty_U(TN)$ satisfies $a(x,u) = T_xN$ on $N \times U$ and is analytic, then, for $T > 0$, the set of u in $C^\infty([0,T]; U)$ such that all the trajectories of $\dot{x} = f(x,u(t))$ defined on $[0,T]$ have a controllable linearized control system is residual in $C^\infty([0,T]; U)$. This result also follows from Corollary 1.8 (take $I = (-T, 2T) \ldots$) even with controllability with impulsive controls of the linearized systems and f not necessarily analytic.

Of course, we have a similar corollary for periodic systems. More precisely, assume that $I = \mathbb{R}$ and that for some positive real number T

$$f(x, t+T, u) = f(x,t,u) \quad \text{for all} \quad (x,t,u) \quad \text{in} N \times \mathbb{R} \times U. \tag{1.39}$$

Denote by $C^\infty_T(N \times \mathbb{R}; U) \simeq C^\infty(N \times (\mathbb{R}/T\mathbb{Z}); U)$ (respectively, $C^\infty_T(\mathbb{R}; U) \simeq C^\infty(\mathbb{R}/T\mathbb{Z}; U)$) the set of u in $C^\infty(N \times \mathbb{R}; U)$ (respectively, $C^\infty(\mathbb{R}; U)$) which are T-periodic in time. Then we have the following corollary.

COROLLARY 1.9. *The set of u in $C^\infty_T(N \times \mathbb{R}; U)$ (respectively, $C^\infty_T(\mathbb{R}; U)$) which saturate f on $N \times \mathbb{R}$ is residual in $C^\infty_T(N \times \mathbb{R}; U)$ (respectively, $C^\infty_T(\mathbb{R}; U)$).*

Proof. Consider Σ has a system on $N \times (\mathbb{R}/T\mathbb{Z}) \ldots$.

A direct consequence of the first part of Corollary 1.9 is the following time-varying version of Corollary 1.5.

COROLLARY 1.10. *Assume that $f \in C^\infty(\mathbb{R}^n \times \mathbb{R} \times U; \mathbb{R}^n)$ satisfies* (1.39) *(with $N = \mathbb{R}^n$), $0 \in U$, $f = 0$ on $\{0\} \times \mathbb{R} \times \{0\}$, and that $\dot{x} = f(x,t,u)$ can be globally (respectively, locally) asymptotically stabilized by means of a continuous T-periodic time varying feedback law. Then $\dot{x} = f(x,t,u)$ can be globally (respectively, locally) asymptotically stabilized by means of a continuous T-periodic, time varying feedback law, of class C^∞ on $(\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$ (respectively, $\{x \in \mathbb{R}^n; 0 < |x| < r\} \times \mathbb{R}$ for some $r > 0$) which saturates f on $(\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$ (respectively, $\{x \in \mathbb{R}^n; 0 < |x| < r\} \times \mathbb{R}$).*

The first part of Corollary 1.9 allows us also to give a time-varying version of Corollary 1.6. Let, for $T > 0, V$ in $C^\infty(\mathbb{R}^n \times \mathbb{R}; [0, +\infty))$ be such that

$$V(x, t+T) = V(x,t) \quad \forall (x,t) \in \mathbb{R}^n \times \mathbb{R}, \tag{1.40}$$

$$V(x,t) = 0 \Longleftrightarrow x = 0, \tag{1.41}$$

and

$$\lim_{|x| \to +\infty} V(x,t) = +\infty \quad \forall t \in \mathbb{R}. \tag{1.42}$$

Assume $0 \in U$ and let $f \in C^\infty(\mathbb{R}^n \times \mathbb{R} \times U; \mathbb{R}^n)$, satisfying (1.39) be such that

$$f(0,t,0) = 0 \quad \forall t \in \mathbb{R} \tag{1.43}$$

and

$$\frac{\partial V}{\partial t}(x,t) + \sum_{i=1}^{n} f_i(x,t,u)\frac{\partial V}{\partial x_i}(x,t) \leq 0 \quad \forall (x,t,u) \in N \times \mathbb{R} \times U. \tag{1.44}$$

Then we have the following corollary.

COROLLARY 1.11. *Assume that, for all* (x,t) *in* $(\mathbb{R}^n \backslash \{0\}) \times [0,T]$ *such that*

$$\frac{\partial V}{\partial t}(x,t) + \sum_{i=1}^{n} f_i(x,t,0)\frac{\partial V}{\partial x_i}(x,t) = 0, \tag{1.45}$$

there exists h *in* $a(x,t,0)$ *such that*

$$\sum_{i=1}^{n} h_i \frac{\partial V}{\partial x_i}(x,t) < 0. \tag{1.46}$$

Then $\dot{x} = f(x,t,u)$ *can be globally asymptotically stabilized by means of a* T*-periodic, time varying feedback law of class* C^∞.

Remark 1.12. Assume that all the assumptions of Corollary 1.6 hold except (1.26). Then, applying Corollary 1.9 with $V(x,t) = V(x)$ and $f(x,t,u) = f(x,u)$, we get that, for all $T > 0, \dot{x} = f(x,u)$ can be globally asymptotically stabilized by means of a T-periodic, time varying feedback law of class C^∞. In general $\dot{x} = f(x,u)$ will not be locally asymptotically stabilizable by means of a continuous feedback law $u = u(x)$, e.g., $n = 2$, $m = 2$, $f(x,u) = (-u_2^2 u_1(u_1 x_1 + x_2), -u_2^2(u_1 x_1 + x_2))$, $V = x_1^2 + x_2^2$: The assumptions of Corollary 1.6 hold except (1.26), but f does not map a neighborhood of zero in $\mathbb{R}^2 \times \mathbb{R}^2$ onto a neighborhood of zero in $\mathbb{R}^2$ and therefore, by a theorem of Brockett [B], $\dot{x} = f(x,u)$ cannot be locally asymptotically stabilized by means of a continuous feedback law $(u = u(x))$.

In our next corollary we have again $0 \in U$ and $f \in C^\infty(\mathbb{R}^n \times \mathbb{R} \times U; \mathbb{R}^n)$ satisfying (1.39) (with $N = \mathbb{R}^n$) and (1.43). We assume that there exists $\varphi \in C^\infty(\mathbb{R} \times U; U)$ such that

$$\varphi(t+T,u) = \varphi(t,u) \quad \forall (t,u) \in \mathbb{R} \times U, \tag{1.47}$$

$$f(x,T-t,\varphi(T-t,u)) = -f(x,t,u) \quad \forall (x,t,u) \in \mathbb{R}^n \times \mathbb{R} \times U, \tag{1.48}$$

and

$$\varphi(t,0) = 0 \quad \forall t \in \mathbb{R}. \tag{1.49}$$

Let $\overline{x} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be defined by

$$\frac{\partial \overline{x}}{\partial t} = f(\overline{x},t,0), \tag{1.50}$$

$$\overline{x}(x,0) = x. \tag{1.51}$$

We assume also that f is such that $\overline{x}$ is defined on $\mathbb{R}^n \times \mathbb{R}$. Note that by (1.48) and (1.49) we have

$$\overline{x}(x,T) = x \quad \text{for all} \quad x \quad \text{in } \mathbb{R}^n. \tag{1.52}$$

Let $V \in C^\infty(\mathbb{R}^n; [0,+\infty))$ be such that

$$V(x) = 0 \Longleftrightarrow x = 0, \tag{1.53}$$

$$\lim_{|x|\to+\infty} V(x) = +\infty, \tag{1.54}$$

and let $W \in (\mathbb{R}^n \times \mathbb{R}; [0, +\infty))$ be defined by

$$\frac{\partial W}{\partial t} + \sum_{i=1}^{n} f_i(x,t,0)\frac{\partial W}{\partial x_i} = 0, \tag{1.55}$$

$$W(x,0) = V(x) \quad \forall x \in \mathbb{R}^n. \tag{1.56}$$

We have $W(x,t) = V(\overline{x}(x,t))$ and therefore by (1.43), (1.52), (1.53), and (1.54), W satisfies (1.40), (1.41), and (1.42). Then our next corollary follows.

COROLLARY 1.13. *Assume that for all* (x,t) *in* $(\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}$ *there exists* X *in* $a(x,t,0)$ *such that*

$$\sum_{i=1}^{n} X_i \frac{\partial W}{\partial x_i}(x,t) \neq 0. \tag{1.57}$$

Then $\dot{x} = f(x,t,u)$ *can be globally asymptotically stabilized by means of a* T*-periodic, time varying feedback law of class* C^∞.

Let us give an example. Assume

$$f(x,t,u) = f(x,u) \tag{1.58}$$

$$\varphi(t,u) = \varphi(u). \tag{1.59}$$

Assume that, for $g^1, \dots, g^p$ in $\mathcal{A}(f)$ (see (1.3)), the control system

$$\Sigma_e : \dot{x} = \sum_{i=1}^{p} v_i g^i(x,0) \tag{1.60}$$

is globally asymptotically stabilized by means of continuous feedback law; then the conclusion of Corollary 1.13 holds. Indeed let $\overline{v} = (\overline{v}_1, \dots, \overline{v}_p)$ be a continuous feedback law which globally asymptotically stabilizes Σ_e. By a generalization of Kurzweil [KUR] of the converse of a classical Lyapunov's theorem there exists $V \in C^\infty(\mathbb{R}^n; [0, +\infty))$ satisfying (1.53) and (1.54) such that

$$\forall x \in \mathbb{R}^n \setminus \{0\} \quad \exists i \in [1,p] \quad \text{such that } g^i(x,0) \cdot \nabla V(x) \neq 0. \tag{1.61}$$

Hence the assumptions (and therefore the conclusion) of Corollary 1.13 hold. Note that if $f(x,u) = \sum_{i=1}^{m} u_i f_i(x)$, this result has been already proved in [C1] (see [C1, Remark 5.1]). In this case Σ_e is called, in the literature, an extended system of $\dot{x} = \sum_{i=1}^{m} u_i f_i(x)$. Links between trajectories of the extended systems and the trajectories of $\dot{x} = \sum_{i=1}^{m} u_i f_i(x)$ have already been extensively studied (see, e.g., [HH], [KW], [LS], [SL], and the references therein). Note also that Corollary 1.13 implies [C1, Thm. 1.1], [C1, Remark 5.1], [C2, Prop. 1.2], and [CA, Thm. 1].

Our proof of Theorem 1.3 can also be used to obtain similar results for observability. Let φ be in $C^\infty(N; \mathbb{R}^q)$, where q is some positive integer. Let

$$D = \{\partial^{|\alpha|} f / \partial u^\alpha; \alpha \in \mathbb{N}^m\} \subset C^\infty_U(TN) \tag{1.62}$$

and let $\mathcal{O}$ be the observation space defined, as a subspace of $C^\infty(N \times U; \mathbb{R}^q)$, by

$$\mathcal{O} = \text{Span}\{L_{X_k} L_{X_{k-1}} \dots L_{X_1} \varphi; k \geq 0, X_i \in D \, \forall i \in [1,k]\} \tag{1.63}$$

where L_{X_i} denotes the usual Lie derivative, $u \in U$ being considered as a parameter, and where, by convention, if $k = 0, L_{X_k} L_{X_{k-1}} \dots L_{X_1} \varphi = \varphi$. We define the observability codistribution $d\mathcal{O}$ by—where again $u \in U$ is a parameter—

$$d\mathcal{O} = \{d\psi; \psi \in \mathcal{O}\} \subset C_U^\infty (T^*N)^q. \tag{1.64}$$

Let $o(x,u)$ and $do(x,u)$, the observation space and the observability codistribution at $(x,u) \in N \times U$, be defined by

$$o(x,u) = \{\psi(x,u); \psi \in \mathcal{O}\} \subset \mathbb{R}^q \tag{1.65}$$

and

$$do(x,u) = \{w(x,u); w \in d\mathcal{O}\} \subset (T_x^* N)^q. \tag{1.66}$$

For u in $C^\infty(N;U)$ we define, with $X(x) = f(x,u(x))$,

$$o_\ell(x;u) = \text{Span}\{(L_X^k \varphi)(x); k \geq 0\} \subset \mathbb{R}^q \tag{1.67}$$

and

$$do_\ell(x,u) = \text{Span}\{(dL_X^k \varphi)(x); k \geq 0\} \subset (T_x^* N)^q. \tag{1.68}$$

Similarly, if I is an open subset of $\mathbb{R}$ and u is in $C^\infty(N \times I; U)$, we define, for (x,t) in $N \times I$,

$$o_\ell(x,t;u) = \text{Span}\left\{\left(\left(\frac{\partial}{\partial t} + L_X\right)^k \varphi\right)(x); k \geq 0\right\} \subset \mathbb{R}^q \tag{1.69}$$

and

$$do_\ell(x,t;u) = \text{Span}\left\{\left(\left(\frac{\partial}{\partial t} + L_X\right)^k \varphi\right)(x); k \geq 0\right\} \subset (T_x^* N)^q \tag{1.70}$$

with $X(x,t) = f(x,u(x,t))$. We have, for all x in N and all u in $C^\infty(N;U)$,

$$o_\ell(x,u) \subset o(x,u(x)), \tag{1.71}$$

and for all (x,t) in $N \times I$ and all u in $C^\infty(N \times I; U)$,

$$o_\ell(x,t;u) \subset o(x,u(x,t)). \tag{1.72}$$

Moreover, if

$$h'(x)\left(\frac{\partial}{\partial u_i} f(x,u)\right) = 0 \quad \forall (x,u) \in N \times I \quad \text{and} \quad \forall i \in [1,m], \tag{1.73}$$

then for all x in N

$$do_\ell(x;u) \subset do(x,u(x)). \tag{1.74}$$

Similarly, for all (x,t) in $N \times I$ and all $\tilde{u}$ in $C^\infty(I;U)$,

$$do_\ell(x,t;u) \subset do(x,\tilde{u}(t)) \tag{1.75}$$

with $u(x,t) = \tilde{u}(t)$.

The counterpart of Theorem 1.3 to this situation is the following theorem.

THEOREM 1.14. *Assume that* (1.16) *and* (1.17) *hold. Then the set of* u *in* Ω *such that, for all* x *in* N,

$$o_\ell(x; u \circ h) = o(x, u \circ h(x)) \tag{1.76}$$

is residual in Ω. *If, moreover,* (1.73) *holds, then the set of* u *in* Ω *such that, for all* x *in* N,

$$do_\ell(x; u \circ h) = do(x, u \circ h(x)) \tag{1.77}$$

is residual in Ω.

We will give in §3 the modifications of the proof of Theorem 1.3 in order to get Theorem 1.14. As Corollaries 1.8 and 1.9 are corollaries of Theorem 1.3, we have the following corollary of Theorem 1.14.

COROLLARY 1.15. *The set of* u *in* $C^\infty(N \times I; U)$ *(respectively,* $C^\infty_T(N \times \mathbb{R}; U)$ *where* $T > 0$*), such that* (1.72) *is an equality for all* (x,t) *in* $N \times I$ *(respectively,* $N \times \mathbb{R}$*), is residual in* $C^\infty(N \times I; U)$ *(respectively,* $C^\infty_T(N \times \mathbb{R}; U)$*). The set of* $\tilde{u}$ *in* $C^\infty(I; U)$ *(respectively,* $C^\infty_T(\mathbb{R}; U)$*) such that, with* $u(x,t) = \tilde{u}(t)$, (1.72) *is an equality for all* (x,t) *in* $N \times I$ *(respectively,* $N \times \mathbb{R}$*), is residual in* $C^\infty(I; U)$ *(respectively,* $C^\infty_T(\mathbb{R}; U)$*). The set of* $\tilde{u}$ *in* $C^\infty(I; U)$ *(respectively,* $C^\infty_T(\mathbb{R}; U)$*), such that* (1.75) *is an equality for all* (x,t) *in* $N \times I$ *(with, again,* $u(x,t) = \tilde{u}(t)$*), is residual in* $C^\infty(I; U)$ *(respectively,* $C^\infty_T(\mathbb{R}; U)$*).*

Remark 1.16. The last statement of Corollary 1.15 is related to a result obtained, independently of us and with different methods, by Wang and Sontag in [WS] when f is analytic.

As an application of Theorem 1.14 and Corollary 1.15 let us give an improvement of Example 1.7. Again let $V \in C^\infty(\mathbb{R}^n; [0, +\infty))$, satisfying (1.24) and (1.25). Let X_0, X_1, X_2 be three vector fields on $\mathbb{R}^n$ of class C^∞. Assume (1.29) and

$$L_{X_i} V \leq 0 \quad \forall i \in \{0, 1\}. \tag{1.78}$$

Assume also that, for all x in $\mathbb{R}^n \backslash \{0\}$ such that $L_{X_2} V(x) = 0$, there exist a positive integer and k vector fields $Y_1, \ldots, Y_k$ in the Lie algebra generated by X_0 and X_1 such that

$$L_{Y_1} \ldots L_{Y_k} V(x) \neq 0 \quad \text{or} \quad L_{Y_1} \ldots L_{Y_k} L_{X_2} V(x) \neq 0. \tag{1.79}$$

Then we have the following corollary.

COROLLARY 1.17. *Under the above assumptions, for any positive real number* T, $\dot{x} = X_0(x) + u_1 X_1(x) + u_2 X_2(x)$ *can be globally asymptotically stabilized by means of a time varying,* T*-periodic feedback law of class* C^∞. *If, moreover,* X_0 *does not vanish on* $\{x \in \mathbb{R}^n \backslash \{0\}; L_{X_1} V(x) = L_{X_2} V(x) = 0\}$ *then* $\dot{x} = X_0(x) + u_1 X_1(x) + u_2 X_2(x)$ *can be globally asymptotically stabilized by means of a feedback law of class* C^∞.

Proof. Let us start with the second part of this corollary. We first note that we may assume that

$$X_0(x) \neq 0 \quad \forall x \in \mathbb{R}^n \backslash \{0\}. \tag{1.80}$$

Indeed, if $\overline{X}_0 = X_0 - (L_{X_1} V) X_1 - (L_{X_2} V) X_2$, then

$$\overline{X}_0(x) \neq 0 \quad \forall x \in \mathbb{R}^n \backslash \{0\} \tag{1.81}$$

and

$$(\overline{X}_0, X_1, X_2) \tag{1.82}$$

satisfies the assumptions of Corollary 1.17. Hence replacing, if necessary, X_0 by $\overline{X}_0$, we may assume (1.80). We now apply Theorem 1.14 with $N = \mathbb{R}^n\backslash\{0\}, m = 1, U = \mathbb{R}, f(x, u_1) = X_0 + u_1^2 X_1, Y = \mathbb{R}^n\backslash\{0\}, h(x) = x, q = 3, \varphi = (L_{X_0}V, L_{X_1}V, L_{X_2}V)$ and Ω an open neighborhood of $u_1^* \equiv 0$ small enough so that (1.16) holds (see (1.80) and any u_1 in Ω extended by 0 at 0 is of class C^∞ on $\mathbb{R}^n$. We get the existence of $\overline{u}_1$ in $C^\infty(\mathbb{R}^n; [0, +\infty))$ such that

$$\overline{u}_1(0) = 0 \tag{1.83}$$

and—see (1.76) and (1.79)—

$$\{x \in \mathbb{R}^n; L_X^{k+1}V(x) = L_X^k L_{X_1}V(x) = L_X^k L_{X_2}V(x) = 0, \forall k \geq 0\} \subset \{0\} \tag{1.84}$$

with $X = X_0 + \overline{u}_1^2 X_1$. Note also that, by (1.78),

$$L_X V \leq 0. \tag{1.85}$$

Finally, using the improvements of [JQ] given in [LA, Thm. 1] or [OS, Prop. 1], we get that $u = (\overline{u}_1^2 - L_{X_1}V, -L_{X_2}V)$ globally asymptotically stabilized $\dot{x} = X_0 + u_1 X_1 + u_2 X_2$. The first part of Corollary 1.17 can be obtained in a similar way by using Corollary 1.15 (and more precisely the result dealing with $C_T^\infty(N \times \mathbb{R}; U)$ and o) instead of Theorem 1.14.

2. Proof of Theorem 1.3. For a subset S of N, let us denote by $\Omega'(S)$ the set of u in Ω such that $u \circ h$ saturates f on S. If, for all compact subsets S of N, $\Omega'(S)$ is residual in Ω, then $\Omega'(N)$ will also be residual in Ω; indeed $N = \bigcup_{n\in\mathbb{N}} S_n$ for some sequence $(S_n; n \in \mathbb{N})$ of compact subsets of N and we have $\Omega'(N) = \bigcap_{n\in\mathbb{N}} \Omega'(S_n)$. So it remains only to prove that, if

$$S \subset N \text{ is compact,} \tag{2.1}$$

then $\Omega'(S)$ is residual. Now we fix S satisfying (2.1) and, for simplicity, we will write Ω' for $\Omega'(S)$.

We equip N with a Riemannian metric. This allows us to define, for X in T_xN and $E \subset T_xN$,

$$d(X, E) = \text{Inf } \{|X - Z|; Z \in E\}, \tag{2.2}$$

where $|\cdot|$ is the norm on T_xN defined by the Riemannian metric. Let K be a compact set, let g be in $\mathcal{A}$, and let δ be a positive real number. For such K, g, and S we define

$$\Omega(K, g, \delta) = \{u \in \Omega; d(g(x, \overline{u}(x)), a_\ell(x; \overline{u})) < \delta; \forall x \in h^{-1}(K) \cap S\}, \tag{2.3}$$

where $\overline{u} = u \circ h$. Clearly, for a suitable sequence $((g_n, \delta_n); n \in \mathbb{N})$,

$$\Omega' = \bigcap_{n\in\mathbb{N}} \Omega(h(S), g_n, \delta_n).$$

Hence, if $\Omega(h(S), g, \delta)$ is open and dense in Ω, Ω' is residual in Ω. The fact that $\Omega(h(S), g, \delta)$ is open follows from the upper semicontinuity of the map $N \times C^\infty(N; U) \to [0, +\infty)$, $(x, u) \to d(g(x, u(x)), a_\ell(x; u))$ and from the compactness of S. It remains only to prove that

$$\Omega(h(S), g, \delta) \quad \text{is dense in } \Omega. \tag{2.4}$$

We equip Y with a Riemannian metric, and still denote by $|\cdot|$ the associated norm on T_yY. Let Q be a compact neighborhood of $h(S)$. For a positive real number β, let

$$\Omega_\beta = \{u \in \Omega; |h'(x)(f(x,\overline{u}(x)))| > 1/\beta \ \forall x \in S, \\ |u'(y)z| < \beta|z| \ \forall y \in Q, \forall z \in T_yY\setminus\{0\}\}, \tag{2.5}$$

and

$$\Omega_\beta(K,y,\delta) = \Omega_\beta \cap \Omega(K,g,\delta). \tag{2.6}$$

Clearly (2.4) will be proved if we check that, for all positive real number β,

$$\Omega_\beta(h(S),g,\delta) \text{ is dense in } \Omega_\beta. \tag{2.7}$$

We now consider β as a fixed positive real number. Then, from (2.5), we see that there exists a finite number of compact subsets $K_1, K_2, \dots, K_n$, each one included in a coordinate chart of Y, such that

$$h(S) = \bigcup_{i=1}^{n} K_i \tag{2.8}$$

and, for all u^0 in Ω_β, there exist n maps $\theta_1, \dots, \theta_n$ in $C^\infty(Y,\mathbb{R})$ such that, with $\overline{u}^0 = u^0 \circ h$,

$$(\theta_i \circ h)'(x)(f(x,\overline{u}_0(x))) \neq 0 \quad \forall x \in h^{-1}(K_i) \cap S. \tag{2.9}$$

Clearly,

$$\Omega_\beta(h(S),g,\delta) = \bigcap_{i=1}^{n} \Omega_\beta(K_i,g,\delta) \tag{2.10}$$

and

$$\Omega_\beta(K_i,g,\delta) \text{ is open} \quad \forall i \in [1,n]. \tag{2.11}$$

So (2.7) will be proved if we check that $\Omega_\beta(K_i,g,\delta)$ is dense in Ω_β for all i in $[1,n]$. We now fix i in $[1,n]$ and, for simplicity, we will omit this index: We will write $\Omega_\beta(K,g,\delta)$ for $\Omega_\beta(K_i,g,\delta)$ and θ_i for θ. Let u^0 be in Ω_β; we want to check that

$$u^0 \in \overline{\Omega_\beta(K,g,\delta)}. \tag{2.12}$$

Let K_1 be a compact neighborhood of K also included in a coordinate chart of Y. In order to prove (2.12) it suffices to check that, given an integer μ and a positive real number ε, there exists u in $\Omega(K,g,\delta)$ such that

$$\text{support } (u-u^0) \subset K_1, \tag{2.13}$$

$$|u-u^0|_{K_1,\mu} = \operatorname{Max}\{|\partial^\alpha(u-u^0)/\partial y^\alpha(x)|; \\ x \in K_1, |\alpha| \leq \mu - 1, \alpha \in \mathbb{N}^{\dim Y}\} < \varepsilon, \tag{2.14}$$

where, in (2.14), the derivatives are computed in a fixed coordinate chart containing K_1. We choose a function η in $C^\infty(Y;[0,1])$ such that

$$\eta = 1 \quad \text{on a neighborhood of } K \tag{2.15}$$

(2.16) $$\text{Support } (\eta) \subset K_1.$$

Let

(2.17) $$C^\infty_{2\pi}(\mathbb{R};\mathbb{R}^m) = \{b \in C^\infty(\mathbb{R};\mathbb{R}^m); b(t+2\pi) = b(t) \forall t \in \mathbb{R}\}.$$

Let $\theta(=\theta_i) \in C^\infty(Y;\mathbb{R}^m)$, satisfying (2.9) (with $K_i = K$). For b in $C^\infty_{2\pi}(\mathbb{R};\mathbb{R}^m)$, $\mu \in \mathbb{N}\backslash\{0\}$, and $\tau \in (0,1]$ we define u in $C^\infty(Y;\mathbb{R}^m)$—as in [C2, §2]—by

(2.18) $$u(y) = u^0(y) + \tau^\mu \eta(y) b(\theta(y)/\tau).$$

Then using (2.16) we have (2.13). Moreover, if τ is small enough (depending on b, μ, and ϵ), (2.14) is satisfied. We are going to check that for generic (and universal; they do not depend on $K, g, \delta, f \ldots$)b in $C^\infty_{2\pi}(\mathbb{R};\mathbb{R}^m)$ we have, if μ is large enough,

(2.19) $$u \in \Omega_\beta(K,g,\delta) \quad \text{for } \tau \text{ small enough.}$$

Let

(2.20) $$\overline{u} = u \circ h = (\overline{u}_1, \overline{u}_2, \ldots, \overline{u}_m), \qquad \overline{\theta} = \theta \circ h,$$

and

(2.21) $$Z(x) = f(x, \overline{u}(x)) \in T_x N.$$

Let us define a sequence $(X_j; j \geq 1)$ of tangent vector fields on N by

(2.22) $$X_1(x) = \tau^{\mu-1} \sum_{i=1}^m b_i^{(1)}(\overline{\theta}(x)/\tau) L_Z \overline{\theta}(x) \frac{\partial f}{\partial u_i}(x, \overline{u}(x)) \quad \forall x \in N,$$

and, for all $j \geq 2$,

(2.23) $$X_j(x) = [Z, X_{j-1}](x) + \sum_{i=1}^m L_{X_{j-1}} \overline{u}_i(x) \frac{\partial f}{\partial u_i}(x, \overline{u}(x)) \quad \forall x \in N.$$

Using (2.15), (2.18), (2.20), (2.21), (2.22), and (2.23), we easily have

(2.24) $$X_j(x) \in a_\ell(x; \overline{u}) \quad \forall x \in h^{-1}(K), \quad \forall j \geq 1.$$

In order to give a useful expression of X_j, $j \geq 1$, we introduce some combinatorial notations. Let $\mathcal{E}_k$ be the set of sequences $I = i_1 i_2 \ldots i_k$ of k elements of $\{0, 1, \ldots, m\}$; the length k of the sequence I will be denoted by $|I|$. For convenience we denote by $\mathcal{E}_0$ the set whose unique element is the empty sequence, denoted by $\emptyset$; we have $|\emptyset| = 0$. For $I = i_1 i_2 \ldots i_k$ and $J = j_1 j_2 \ldots j_{k'}$ we define $I * J \in \mathcal{E}_{k+k'}$ by

(2.25) $$I * J = i_1 i_2 \ldots i_k j_1 j_2 \ldots j_{k'}.$$

Let $\mathcal{E} = \bigcup_{k \geq 0} \mathcal{E}_k$ and $\mathcal{E}' = \mathcal{E}\backslash(\{I * 0; I \in \mathcal{E}\} \cup \{\emptyset\})$. For I in $\mathcal{E}'$ we define, by induction on $|I|$, an element f_I in $C^\infty_U(TN)$ by

(2.26) $$f_i = \frac{\partial f}{\partial u_i} \quad \forall i \in [1, m],$$

(2.27) $$f_{0*I} = [f, f_I] \quad \forall I \in \mathcal{E}',$$

and

$$f_{i*I} = \frac{\partial}{\partial u_i} f_I \quad \forall i \in [1,m], \quad \forall I \in \mathcal{E}'. \tag{2.28}$$

In a similar way as in [C1] we define a sequence $(c_r(I); r \geq 0, I \in \mathcal{E})$ of function in $C^\infty_{2\pi}(\mathbb{R};\mathbb{R})$ by

$$c_0(\emptyset) = 1, \quad c_0(I) = 0 \quad \text{if} \quad |I| \geq 1, \quad c_p(\emptyset) = 0 \quad \text{if} \quad p > 0, \tag{2.29}$$

$$c_p(i * I) = \dot{b}_i c_{p-1}(I) + \dot{c}_{p-1}(i * I) \quad \forall I \in \mathcal{E}, \quad \forall i \in [0,m], \quad \forall p \geq 1, \tag{2.30}$$

with the convention $\dot{b}_0 = 1$. For example, $c_1(i) = b_i^{(1)}$ for all $i \in [1,m]$, $c_1(I) = 0$ if $|I| \geq 2$, $c_2(i) = b_i^{(2)}$ for all $i \in [1,m]$, $c_2(0) = 0$, $c_2(i_1 i_2) = b_{i_1}^{(1)} b_{i_2}^{(1)}$ for all $i_1 \in [1,m]$, for all $i_2 \in [1,m]$. Note that

$$c_r(I) \text{ is a polynomial in the variables } b_i^{(j)}, j \leq r, i \in [1,m]. \tag{2.31}$$

For I in $\mathcal{E}$, let $\alpha(I)$ be the number of times the index 0 appear in I and let $\beta(I) = |I| - \alpha(I)$. Let also, for two integers i and j,

$$\Delta_{i,j} = \left\{ k = (k_1, \ldots, k_j) \in \mathbb{N}^j; 1 \leq k_1 \leq \ldots \leq k_j, \sum_{r=1}^{j} k_r = i \right\} \tag{2.32}$$

and, if $k \in \Delta_{i,j}$,

$$L_Z^k \overline{\theta} = L_Z^{k_1} \overline{\theta} \ldots L_Z^{k_j} \overline{\theta}. \tag{2.33}$$

Note that $\Delta_{i,j}$ is not empty if and only if $j \leq i$. By induction on $r \geq 1$ we can check that

$$\begin{aligned} X_r(x) = & \sum_{I \in \mathcal{E}', |I| \leq r} (L_Z \overline{\theta})^{r-\alpha(I)} \tau^{\alpha(I)+\mu\beta(I)-r} c_r(I)(\overline{\theta}(x)/\tau) f_I(x, \overline{u}(x)) \\ & + \sum_{I \in \mathcal{E}', |I| \leq r} \sum_{\alpha(I) < s < r} \tau^{\alpha(I)+\mu\beta(I)-s} X_{r,I,s}(x) \end{aligned} \tag{2.34}$$

with

$$X_{r,I,S}(x) = \sum_{k \in \Delta_{r-\alpha(I), s-\alpha(I)}} (L_Z^k \overline{\theta})(x) Q_{r,s,I,J}(\overline{\theta}(x)/\tau) f_I(x, \overline{u}(x))$$

where

$$Q_{r,s,I,J} \text{ is a polynomial in the variable } (b_i^{(j)}; j \geq 1, i \in [1,m]). \tag{2.35}$$

By definition of g and using Jacobi's identity we can see that, in the vectorial space $C^\infty_U(TN)$ and for some large enough integer ℓ',

$$g \in \text{Span } \{f_I; I \in \mathcal{E}', |I| \leq \ell'\}. \tag{2.36}$$

Let

$$\ell = \ell' \mu \tag{2.37}$$

and

$$q_0(\ell) = \text{cardinal of } \{I; I \in \mathcal{E}', |I| \leq \ell\} = m(m+1)^{\ell-1}. \tag{2.38}$$

Let us introduce an ordering on the I in $\mathcal{E}'$ with $0 < |I| \leq \ell$ (e.g., lexicographical) and, for $q > q_0(\ell)$, let us consider the nonsquare $q \times q_0(\ell)$ matrix with entries $(c_r(I); 1 \leq r \leq q, I \in \mathcal{E}'$ with $|I| \leq \ell)$. Let J_q be the space of jets

$$J_q = \{b_i^{(j)}; 0 \leq j \leq q, 1 \leq i \leq m\}. \tag{2.39}$$

In Appendix A we prove the following lemma.

LEMMA 2.1. *If q is large enough, then the set of $(b_i^{(j)}; 0 \leq j \leq q, 1 \leq i \leq m)$ in J_q such that*

$$\text{rank } (c_r(I); 1 \leq r \leq q, I \in \mathcal{E}', |I| \leq \ell) < q_0(\ell) \tag{2.40}$$

is of codimension at least 2 in J_q.

Let us remark that Appendix A provides an explicit value of q (which is not optimal and, of course, depends on m and ℓ) such that the conclusion of Lemma 2.1 holds. Applying Thom's transversality theorem (see, e.g., [GG, Chap. II, Thm. 4.9]) and using Lemma 2.1 we get, for $q_1(\ell)$ large enough, the existence of a b in $C^\infty_{2\pi}(\mathbb{R}; \mathbb{R}^m)$ such that

$$\text{rank } (c_r(I)(s); 1 \leq r \leq q_1(\ell), I \in \mathcal{E}', |I| \leq \ell) = q_0(\ell), \quad \forall s \in \mathbb{R}. \tag{2.41}$$

In fact, there exists a residual set $\mathcal{R}$ in $C^\infty_{2\pi}(\mathbb{R}; \mathbb{R}^m) \simeq C^\infty(\mathbb{R}/2\pi\mathbb{Z}; \mathbb{R}^m)$ such that (2.41) holds for all b in $\mathcal{R}$ and all $\ell \geq 1$. We choose a b such that (2.41) holds; we are going to check that (2.19) holds. By (2.36) we may assume

$$g = f_{I_0} \quad \text{where } I_0 \in \mathcal{E}' \quad \text{with } |I_0| \leq \ell'. \tag{2.42}$$

Let us denote by M various constants independent of τ in (0, 1] and x in $h^{-1}(K) \cap S$ (but M may depend on $K, S, \ell, h, f, b, u_0, \beta, \mu \ldots$). Let us remark that, for all τ in (0, 1],

$$|L_Z \overline{\theta}| \leq M \text{ on } h^{-1}(K) \cap S \tag{2.43}$$

and

$$|L_Z^j \overline{\theta}| \leq M\tau^{2-j} \quad \text{on } h^{-1}(K) \cap S \quad \forall j \in [2, q_1(\ell)]. \tag{2.44}$$

From these two inequalities, we get that, if $I \in \mathcal{E}'$, $\beta(I) \leq s < r \leq q_1(\ell)$, and $k \in \Delta_{r-\beta(I), s-\beta(I)}$, then

$$|L_Z^k \overline{\theta}| \leq M\tau^{s-r+1} \quad \text{on } h^{-1}(K) \cap S, \quad \forall \tau \in (0, 1]. \tag{2.45}$$

By (2.9) there exist $\gamma > 0$ and τ_0 in (0, 1] such that

$$|L_Z \overline{\theta}| \geq \gamma \quad \text{on } h^{-1}(K) \cap S, \quad \forall \tau \in (0, \tau_0). \tag{2.46}$$

From (2.34), (2.45), and (2.46) we get, on $h^{-1}(K) \cap S$ and for r in $[1, q_1(\ell)]$,

$$X_r = \sum_{I \in \mathcal{E}', |I| \leq r} (L_Z \overline{\theta})^{r-\alpha(I)} \tau^{\alpha(I)+\mu\beta(I)-r} (C_r(I)(\overline{\theta}/\tau) + \tau R(r, I, \tau)) f_I \tag{2.47}$$

where $R(r, I, \tau) \in C^0(h^{-1}(K) \cap S; \mathbb{R})$ satisfies for all $(r, I, \tau) \in [1, q_1(\ell)] \times \mathcal{E}' \times (0, \tau_0]$, with $|I| \leq q_1(\ell)$,

$$|R(r, I, \tau)| \leq M \quad \text{on } h^{-1}(K) \cap S. \tag{2.48}$$

From (2.41) and (2.48) we get, on $h^{-1}(K) \cap S$, for some $\tau_1 \in (0, \tau_0)$ and for all τ in $(0, \tau_1)$,

$$\sum_{r=1}^{q_1(\ell)} \gamma(r, \tau)(C_r(I)(\overline{\theta}/\tau) + \tau R(r, I, \tau)) = 0 \quad \forall I \in \mathcal{E}' \backslash \{I_0\} \quad \text{with } |I| \leq \ell \tag{2.49}$$

and

$$\sum_{r=1}^{q_1(\ell)} \gamma(r, \tau)(C_r(I_0)(\overline{\theta}/\tau) + \tau R(r, I_0, \tau)) = 1, \tag{2.50}$$

where $\gamma(r, \tau) \in C^0(h^{-1}(K) \cap S; \mathbb{R})$ satisfies

$$|\gamma(r, \tau)| \leq M \quad \text{on } h^{-1}(K) \cap S, \quad \forall (r, \tau) \in [1, q_1(\ell)] \times (0, \tau_1]. \tag{2.51}$$

From (2.47), (2.48), (2.49), (2.50), and (2.51) we get, on $h^{-1}(K) \cap S$,

$$\left| f_{I_0}(x, \overline{u}(x)) - \sum_{r=1}^{q_1(\ell)} (L_Z \overline{\theta})^{-r+\alpha(I_0)} \tau^{-\alpha(I_0)-\mu\beta(I_0)+r} \gamma(r, \tau)(x) X_r(x) \right| \leq M \sum_{I \in \mathcal{E}', \ell < |I| \leq r} \tau^{-\alpha(I_0)-\mu\beta(I_0)+\alpha(I)+\mu\beta(I)} \tag{2.52}$$

which proves (2.19) (and ends the proof of Theorem 1.3) since $g = f_{I_0}$ (see (2.42)) and, by (2.37),

$$\alpha(I_0) + \mu\beta(I_0) < \alpha(I) + \mu\beta(I), \quad \forall I \in \mathcal{E}' \quad \text{with } |I| > \ell. \tag{2.53}$$

(Note also that since Ω is open we have, using (2.16) and (2.18), $u \in \Omega_\beta$ if μ is large enough and, then, τ small enough.)

3. Proofs of corollaries and Theorem 1.14. In this section we give the proof of the corollaries of §1, which were not proved in that section, and the modifications of the proof of Theorem 1.3 in order to get Theorem 1.14.

3.1. Proof of Corollary 1.5. We prove the global statement only (the proof of the local statement is similar). Let $\overline{u} \in C^0(\mathbb{R}^n; U)$ be such that

$$\overline{u}(0) = 0 \text{ and } 0 \text{ is globally asymptotically stable point of } \dot{x} = f(x, \overline{u}(x)). \tag{3.1}$$

By a generalization of Kurzweil [KUR] of the converse of Lyapunov's second theorem, there exists V in $C^\infty(\mathbb{R}^n; [0, +\infty))$ such that

$$(V(x) = 0 \Longleftrightarrow x = 0), \lim_{|x| \to +\infty} V(x) = +\infty, \tag{3.2}$$

and

$$f(x, \overline{u}(x)) \cdot \nabla V(x) < 0 \quad \forall x \neq 0. \tag{3.3}$$

Let Ω be the set of u in $C^\infty(\mathbb{R}^n\backslash\{0\};U)$ such that

$$|u(x)| < |\overline{u}(x)| + |x| \quad \forall x \in \mathbb{R}^n\backslash\{0\}, \tag{3.4}$$

$$f(x,u(x)) \cdot \nabla V(x) < 0 \quad \forall x \in \mathbb{R}^n\backslash\{0\}. \tag{3.5}$$

This set Ω is open and, since $C^\infty(\mathbb{R}^n\backslash\{0\};U)$ is dense in $C^0(\mathbb{R}^n\backslash\{0\};U)$, it is nonempty by (3.3). We take $Y = N = \mathbb{R}^n\backslash\{0\}, h(x) = x$. By (3.5), (1.16) holds. Applying Theorem 1.3 we get that Ω contains at least a map u^0 which saturates f on $\mathbb{R}^n\backslash\{0\}$. This feedback law u^0 globally asymptotically stabilizes $\dot{x} = f(x,u)$, belongs to $C^\infty(\mathbb{R}^n\backslash\{0\};U)\cap C^0(\mathbb{R}^n;U)$, and satisfies $u^0(0) = 0$.

3.2. Proof of Corollary 1.6. Let Ω be an open neighborhood of 0 in $C^\infty(\mathbb{R}^n\backslash\{0\};U)$ such that, for all u in Ω,

$$f(x,u(x)) \neq 0 \quad \forall x \in \mathbb{R}^n\backslash\{0\}, \tag{3.6}$$

$$\forall x \in \mathbb{R}^n\backslash\{0\}, \quad \exists g \in \mathcal{A}\cup\{f\} \quad \text{such that } g(x,u(x)) \cdot \nabla V(x) \neq 0, \tag{3.7}$$

and

$$u \text{ extended by 0 on } \{0\} \text{ is of class } C^\infty \text{ on } \mathbb{R}^n. \tag{3.8}$$

The existence of such a Ω follows from (1.26) and (1.27). We take $Y = N = \mathbb{R}^n\backslash\{0\}, h(x) = x$. It follows from Theorem 1.3 that there exists u^0 such that

$$u^0 \in \Omega \tag{3.9}$$

$$u^0 \text{ saturates } f \text{ on } \mathbb{R}^n\backslash\{0\}. \tag{3.10}$$

We extend u^0 by 0 on $\{0\}$ and still denote by u^0 this extension. Let

$$\mathcal{O} = \{(x,v); v + u^0(x) \in U\} \tag{3.11}$$

and let $F \in C^\infty(\mathcal{O};\mathbb{R}^n)$ be defined by

$$F(x,v) = f(x,v+u^0(x)). \tag{3.12}$$

Also let

$$F_0(x) = F(x,0), \tag{3.13}$$

$$F_i(x) = \frac{\partial f}{\partial u_i}(x,u^0(x)), \qquad i \in [1,m]. \tag{3.14}$$

Then, using (1.23), (3.7), and (3.10),

$$F_0(x) \cdot \nabla V(x) \leq 0 \quad \forall x \in \mathbb{R}^n, \tag{3.15}$$

and for all x in $\mathbb{R}^n\backslash\{0\}$ such that (3.15) is an equality, there exists k in $\mathbb{N}$ and i in $[1,m]$ such that

$$L_{ad^k_{F_0}(Fi)}V(x) \neq 0. \tag{3.16}$$

The conclusion of Corollary 1.6 follows from Jurdjevic and Quinn [JQ] (see Appendix B).

3.3. Proof of Corollary 1.9. The proof is similar to the proof of Corollary 1.5. We omit it.

3.4. Proof of Corollary 1.11. The proof is similar to the proof of Corollary 1.6.; the only difference is that we use Corollary 1.9 instead of Theorem 1.3.

3.5. Proof of Corollary 1.13. Let Ω be an open neighborhood of 0 in $C^\infty((\mathbb{R}^n\backslash\{0\})\times(0,T/2);U)$ such that

(3.17) any u in Ω extended by 0 outside $(\mathbb{R}^n\backslash\{0\})\times(0,T/2)$ is C^∞ on $\mathbb{R}^n\times\mathbb{R}$.

For u in Ω we extend u to $\mathbb{R}^n\times\mathbb{R}$ by

$$u = 0 \text{ on } (\{0\}\times\mathbb{R})\cup(\mathbb{R}^n\times(T/2)\mathbb{Z}), \tag{3.18}$$

and, as in [C1],

$$u(x,t) = \varphi(T-t, u(x,T-t)) \quad \forall (x,t)\in\mathbb{R}^n\times(T/2,T), \tag{3.19}$$

$$u(x,t+T) = u(x,t) \quad \forall (x,t)\in\mathbb{R}^n\times\mathbb{R}. \tag{3.20}$$

We still denote by u this extension. Note that by (1.49), (3.17), (3.18), (3.19), and (3.20)

$$u\in C^\infty(\mathbb{R}^n\times\mathbb{R};U). \tag{3.21}$$

Hence we may consider Ω as a subset of $C^\infty(\mathbb{R}^n\times\mathbb{R};U)$. Now, as Pomet in [P], we define $W^u:\mathbb{R}^n\times\mathbb{R}\to[0,+\infty)$ by

(3.22)
$$\frac{\partial W^u}{\partial t}(x,t)+\sum_{i=1}^n f_i(x,t,u(x,t))\frac{\partial W^u}{\partial x_i}(x,t)=0, \quad W^u(x,0)=V(x) \quad \forall (x,t)\in\mathbb{R}^n\times\mathbb{R},$$

Then, using (1.47), (1.48), (3.19), and (3.20), we easily have

$$W^u(x,t+T)=W^u(x,t) \quad \forall (x,t)\in\mathbb{R}^n\times\mathbb{R}. \tag{3.23}$$

Moreover W^u satisfies (1.41) and (1.42), and, diminishing Ω if necessary, we get from (1.57) that, for all (x,t) in $\mathbb{R}^n\backslash\{0\}\times[0,T/2]$ and for all u in Ω, there exists X in $a(x,t,u(x,t))$ such that

$$\sum_{i=1}^n X_i\frac{\partial W^u}{\partial x_i}(x,t)\neq 0. \tag{3.24}$$

Now, using Corollary (1.8), we get the existence of a $\overline{u}$ in Ω such that

$$\overline{u} \text{ saturates } f \text{ on } (\mathbb{R}^n\backslash\{0\})\times(0,T/2). \tag{3.25}$$

Finally, we obtain the desired conclusion by using the version of [JQ] given in Appendix B: We take $N=(\mathbb{R}^n\backslash\{0\})\times(\mathbb{R}/T\mathbb{Z})$, $F((x,t),u)=(f(x,t,\overline{u}(x,t)+u),\partial/\partial t)$, $V((x,t))=W^{\overline{u}}(x,t)\ldots$; note that by (3.24) and (3.25)

$$Q\subset(\mathbb{R}^n\backslash\{0\})\times([T/2,T]/T\mathbb{Z}). \tag{3.26}$$

Remark 3.1. We may wonder if there exists $\overline{u} \in C^\infty(\mathbb{R}^n \times \mathbb{R}; U)$ satisfying (3.18), (3.19), and (3.20) which saturates f on all $(\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$. This is the case if

$$f(x,u) = \sum_{i=1}^{m} u_i f_i(x) \tag{3.27}$$

and

$$\varphi(u) = -u. \tag{3.28}$$

This can be seen in the following way. First, by Corollary 1.8, there exists $u_1 \in C^\infty(\mathbb{R}^n \times \mathbb{R}; U)$ which saturates f on $(\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$ and satisfies

$$u_1(0,t) = 0 \quad \text{for all } t \text{ in } \mathbb{R}. \tag{3.29}$$

Then, let $u_2(x,t) = d(tu_1(x,t^2))$ where $d \in C^\infty(\mathbb{R}^m; U)$ satisfies $d(u) = u$ for u small and $d(-u) = -d(u)$ for all u. Then u_2 satisfies (3.29), is odd and, as can be easily seen, also saturates f on $(\mathbb{R}^n \backslash \{0\}) \times \{0\}$. From u_2 we can construct u_3 in $C^\infty(\mathbb{R}^n \times \mathbb{R}; U)$ which satisfies (3.18), (3.19), and (3.20) and saturates f on $(\mathbb{R}^n \backslash \{0\}) \times (T/2)\mathbb{Z}$. Now applying Corollary 1.8 on $(\mathbb{R}^n \backslash \{0\}) \times (0,T)$ we get $u_4 \in C^\infty(\mathbb{R}^n \times [0,T]; U)$ which saturates f on $(\mathbb{R}^n \backslash \{0\}) \times (0, T/2)$ and satisfies, for all x in $\mathbb{N}$ and for all (α, β) in $\mathbb{N} \times \mathbb{N}^n$,

$$\frac{\partial^\alpha}{\partial t^\alpha} \frac{\partial^{|\beta|} u_4}{\partial x^\beta} = \frac{\partial^\alpha}{\partial t^\alpha} \frac{\partial^\beta u_3}{\partial x^\beta} \quad \text{on } (\{0\} \times [0,T]) \cup ((\mathbb{R}^n \backslash \{0\}) \times \{0, T/2\}); \tag{3.30}$$

(note that the set of u in $C^\infty(\mathbb{R}^n \times [0,T]; U)$ which satisfy (3.30) is open for the $C^\infty-$topology of $C^\infty((\mathbb{R}^n \backslash \{0\}) \times (0,T/2); U)$). We extend u_4 to $\mathbb{R}^n \times \mathbb{R}$ by requiring (3.18), (3.19), and (3.20). Then this extended u_4 saturates f on $(\mathbb{R}^n \backslash \{0\}) \times (\mathbb{R} \backslash (T/2)\mathbb{Z})$ and, by (3.30), saturates f on $(\mathbb{R}^n \backslash \{0\}) \times (T/2)\mathbb{Z}$. Therefore u_4 saturates f on $(\mathbb{R}^n \backslash \{0\}) \times \mathbb{R}$.

3.6. Proof of Theorem 1.14. To get the first (respectively, second) part of Theorem 1.14 we just use §2 with the following modifications:

- In the definition of $\Omega'(S)$ "$u \circ h$ saturates f on S" is replaced by "u satisfies (1.76) (respectively, (1.77)) for all x in S";
- in (2.2) E is now a subset of $\mathbb{R}^q$ (respectively, a subset of $(T^*_x N)^q$) and $|\ |$ is a norm on $\mathbb{R}^q$ (respectively, a norm on $(T^*_x N)^q$ which depends continuously on x in N);
- in the definition of $\Omega(K, g, \delta), g$ is now in $\mathcal{O}$) (respectively $d\mathcal{O}$) and $a_\ell(x; \overline{u})$ is replaced by $o_\ell(x; \overline{u})$ (respectively, $do_\ell(x; \overline{u})$);
- X_i is now in $C^\infty(N; \mathbb{R}^q)$ (respectively, $C^\infty(T^*N)^q$); X_1 is defined by replacing in (2.22) $\partial f / \partial u_i(x, \overline{u}(x))$ by φ (respectively, $d\varphi$),
- f_I is now in $C^\infty(N \times U; \mathbb{R}^q)$ (respectively, $C^\infty_U(T^*N)^q$) and is defined by

$$f_i(x,u) = \varphi(x) (\text{respectively, } f_i(x,u) = d\varphi(x)) \quad \forall (x,u) \in N \times U, \quad \forall i \in [1,m)$$
$$f_{0*I} = L_f f_I \quad \forall I \in \mathcal{E}'$$

and (2.28).

The remaining part of the proof is unchanged.

Appendix A. In this appendix we prove, in particular, Lemma 2.1. We consider ℓ as fixed and, in order to emphasize that $b_i^{(j)}$ are considered as independent real numbers, we will write b_i^j for $b_i^{(j)}$. Let J be the set of sequences $b = (b_i^j; 1 \le i \le m, j \ge 0)$ and, for an integer q, we denote by J_q the set of b in J such that

$$b_i^j = 0 \quad \text{for all } (i,j) \quad \text{in } [1,m] \times (q, +\infty). \tag{A.1}$$

We have $J_q \subset J_{q+1}$. We define also

$$J_q^{\perp} = \{b \in J; b_i^j = 0 \text{ for all } (i,j) \text{ in } [1,m] \times [1,q]\}. \tag{A.2}$$

Let B_q be the set of b in J_q such that

$$\operatorname{rank}\{c_r(I); 0 \leq r \leq q, I \in \mathcal{E}', 0 < |I| \leq \ell\} < q_0(\ell), \tag{A.3}$$

and let $G_q = J_q \backslash B_q$. Note that

$$G_q \subset G_{q+1}. \tag{A.4}$$

Our first statement is the following proposition.

PROPOSITION A.1. *For any integer q', there exists an integer q larger than q' such that*

$$J_{q'}^{\perp} \cap G_q \neq \phi. \tag{A.5}$$

Lemma 2.1 is a special case of the following corollary of Proposition A.1.

COROLLARY A.2. *Let s be an integer. Then, for q large enough, the codimension of B_q in J_q is at least s.*

Proof of Corollary A.2. Using Proposition A.1 we get the existence of $(s+1)$ integers $q_0, q_1, \ldots, q_s$ such that

$$0 = q_0 < q_1 < \ldots < q_s \tag{A.6}$$

and

$$J_{q_{i-1}}^{\perp} \cap G_{q_i} \neq \phi \quad \text{for all} \quad i \quad \text{in } [1,s]. \tag{A.7}$$

Let, for i in $[1,s]$, $n_i = m(q_i - q_{i-1})$ and let $x_i = (x_i^1, \ldots, x_i^{n_i}) \in \mathbb{R}^{n_i}$ be defined by

$$x_i^{k+(j-1)m} = b_k^{j+q_i}, \quad i \leq j \leq q_i - q_{i-1}, \quad k \in [1,m]. \tag{A.8}$$

Then, for all i in $[1,s]$, there exists a polynomial $P_i \in \mathbb{R}[x_1, \ldots, x_i]$ in the variables $(x_r^j; 1 \leq j \leq n_i, r \in [1,i])$ such that

$$P_i(x_1, \ldots, x_i) = 0 \Longleftrightarrow (b_r^j; r \in [1,m], 0 \leq j \leq q_i) \in B_{q_i}. \tag{A.9}$$

Let us remark that, if we multiply all the components of b by λ, then $c_r(I)$ is multiplied by $\lambda^{\beta}(I)$ where (see §2) $\beta(I)$ is the number of indices in I, counted according to their multiplicity, which are not zero. Hence we may impose that, for any integer i in $[1,s]$, there exists a positive integer m_i such that

$$P_i(\lambda x_1, \ldots, \lambda x_i) = \lambda^{m_i} P(x_1, \ldots, x_i) \quad \forall (\lambda, x_1, \ldots, x_i) \in \mathbb{R} \times \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_i}. \tag{A.10}$$

By (A.10) we get that, for any x_i in $\mathbb{R}^{n_i}$ such that

$$P_i(0, \ldots, 0, x_i) \neq 0, \tag{A.11}$$

and, for any $(x_1, \ldots, x_{i-1})$ in $\mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_{i-1}}$,

$$P_i(x_1, \ldots, x_{i-1}, \lambda x_i) \neq 0 \quad \text{for } \lambda \text{ large enough.} \tag{A.12}$$

Therefore using (A.7) and (A.9), we get that

$$\begin{aligned} &\forall i \in [1,s], \quad \forall (x_1,\dots,x_{i-1}) \in \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_{i-1}}, \\ &\exists x_i \in \mathbb{R}^{n_i} \quad \text{such that } P_i(x_1,\dots,x_i) \neq 0, \end{aligned} \tag{A.13}$$

where, by convention, the meaning of (A.13), when $i = 1$, is

$$\exists x_1 \in \mathbb{R}^{n_1} \quad \text{such that} \quad P_1(x_1) \neq 0. \tag{A.14}$$

Hence Corollary A.2 is a consequence of the following lemma.

LEMMA A.3. *Let* $(P_i; 1 \leq i \leq s)$ *be a sequence of polynomials such that*

$$P_i \in \mathbb{R}[x_1,\dots,x_i]. \tag{A.15}$$

Assume that (A.13) *holds. Then the codimension in* $\mathbb{R}^{n_1+\dots+n_s}$ *of* $\Sigma_s := \{x \in \mathbb{R}^{n_1+\dots+n_s};$ $P_i(x) = 0$, *for all* $i \in [1,s]\}$, *is at least* s.

Proof of Lemma A.3. We prove the lemma by induction on s. This lemma is clearly true for $s = 1$ (see (A.14)). Assume it holds for s; we are going to check that it holds for $s+1$. Let $\theta : \Sigma_{s+1} \to \Sigma_1$ be defined by

$$\theta(x_1,\dots,x_{s+1}) = x_1.$$

By (A.14)

$$\dim \Sigma_1 \leq n_1 - 1. \tag{A.16}$$

Let x_1 be in $\mathbb{R}^{n_1}$. Let us define a new sequence of polynomials $(\overline{P}_i; 1 \leq i \leq s)$

$$\overline{P}_i \in \mathbb{R}([x_2,\dots,x_{i+1}]) \tag{A.17}$$

$$\overline{P}_i(x_2,\dots,x_{i+1}) = P_{i+1}(x_1,x_2,\dots,x_{i+1}). \tag{A.18}$$

This sequence of s polynomials satisfies the hypothesis of Lemma A.3. Hence, by the induction assumption,

$$\dim \theta^{-1}(x_1) \leq n_2 + \dots + n_s - s \quad \text{for all} \quad x_1 \quad \text{in } \Sigma_1. \tag{A.19}$$

From (A.16) and (A.19), we get

$$\dim \Sigma_{s+1} \leq n_1 + n_2 + \dots + n_s - (s+1), \tag{A.20}$$

which ends the proof of Lemma A.3.

We turn to the proof of Proposition A.1. We first introduce some notations. Let $\tilde{J}$ be the set of sequences of real numbers $\tilde{b} = (\tilde{b}_i^j; i \in [0,m], j \geq 0)$. For $\tilde{b}$ in $\tilde{J}$, I in $\mathcal{E}, r$ in $\mathbb{N}$, we define $\tilde{c}_r(I) \in \mathbb{R}$ as we have defined $c_r(I)$ in §2. Note that now, we do not have, in general, $\tilde{b}_0^0 = 1, \tilde{b}_0^j = 0$ for $j > 0$. For example, $\tilde{c}_1(i) = \tilde{b}_i^1$, for all $i \in\in [0,m]$, and $\tilde{c}_2(ij) = \tilde{b}_1^j \tilde{b}_j^1$, for all $i \in [0,m]$, and for all $j \in [0,m]$. Let, for each integer q,

$$\tilde{J}_q = \{\tilde{b} \in \tilde{J}; \tilde{b}_i^j = 0 \, \forall j \geq q+1, \forall i \in [0,m]\}, \tag{A.21}$$

$$\tilde{B}_q = \{\tilde{b} \in \tilde{J}_q; \operatorname{rank}(c_r(I); 0 \leq r \leq q, 0 \leq |I| \leq \ell) < \tilde{q}_0(\ell)\}, \tag{A.22}$$

with $\tilde{q}_0(\ell) = (m+1)((m+1)^\ell - 1)/m = \#\{I \in \mathcal{E} \backslash \{\phi\}; |I| \leq \ell)$. Let

$$\tilde{G}_q = \tilde{J}_q \backslash \tilde{B}_q. \tag{A.23}$$

We have again

$$\tilde{G}_q \subset \tilde{G}_{q+1}. \tag{A.24}$$

For a sequence of real numbers $a = (a_i; i \geq 1)$, we define a map $F_a : \tilde{J} \to \tilde{J}$ by

$$(F_a(\tilde{b}))_i^0 = \tilde{b}_i^0 \quad \forall i \in [0, m] \tag{A.25}$$

$$(F_a(\tilde{b}))_i^j = D((F_a(\tilde{b}))_i^{j-1}) \quad \forall i \in [0, m], \quad \forall j \geq 1, \tag{A.26}$$

where D denotes the differentiation on the set of polynomials in the variables $(a_i; i \geq 1), (\tilde{b}_i^j; i \in [0, m], j \geq 0)$ such that $Da_i = a_{i+1}$, for all $i \geq 1$ and $D\tilde{b}_i^j = a_1\tilde{b}_i^{j+1}$, for all $i \in [0, m]$, for all $j \geq 0$. More precisely, we first define with this differentiation, (A.25), and (A.26) polynomials $(F_a(\tilde{b}))_i^j$ in the variables $(a_i; i \geq 1), (\tilde{b}_i^j; i \in [1, m], j \geq 0)$ and then obtain an element of $\tilde{J}$, still denoted $F_a(\tilde{b})$, by substituting in these polynomials the "values" of $(a_i; i \geq 1)$ and $(\tilde{b}_i^j; i \in [1, m], j \geq 0)$. For example, we have

$$(Fa(\tilde{b}))_i^2 = a_2\tilde{b}_i^1 + a_1^2\tilde{b}_i^2 \quad \forall i \in [1, m] \tag{A.27}$$

and

$$(F_a(\tilde{b}))_i^3 = a_3\tilde{b}_i^1 + 3a_1a_2\tilde{b}_i^2 + a_1^3\tilde{b}_i^3 \quad \forall i \in [1, m]. \tag{A.28}$$

Also let, for each integer q, $\pi_q : \tilde{J} \to \tilde{J}_q$ be defined by

$$(\pi_q(\tilde{b}))_i^j = \tilde{b}_i^j \quad \forall j \leq q, \quad \forall i \in [0, m], \tag{A.29}$$

$$(\pi q(\tilde{b}))_i^j = 0 \quad \forall j > q, \quad \forall i \in [0, m]. \tag{A.30}$$

Let us assume, for the moment, the following lemma.

LEMMA A.4. *If r is an integer such that*

$$a_r \neq 0, \tag{A.31}$$

then, for all integers q,

$$(\pi_{rq} \circ F_a)(\tilde{G}_q) \subset \tilde{G}_{rq}. \tag{A.32}$$

Let $\tilde{J}_q^{\perp} = \{\tilde{b} \in \tilde{J}; \tilde{b}_i^j = 0$ for all i in $[0, m]$ and all $j \geq q + 1\}$. As a corollary of Lemma A.4 and [C1, Lem. 4.1], we have the following proposition.

PROPOSITION A.5. *For any integer q', there exists an integer $q > q'$ such that*

$$J_{q'}^{\perp} \cap \tilde{G}_q \neq \phi. \tag{A.33}$$

Proof of Proposition A.5. Let $a = (a_i; i \geq 1)$ be such that

$$a_i = 0 \quad \text{for all} \quad i \quad \text{in } [1, q'], \tag{A.34}$$

$$a_{q'+1} \neq 0. \tag{A.35}$$

From (A.34) we easily get

$$F_a(\tilde{J}) \subset \tilde{J}_{q'}^{\perp}. \tag{A.36}$$

Applying Lemma A.4 with $r = q' + 1$, we obtain from (A.35)

$$(\pi_{(q'+1)s} \circ F_a)(\tilde{G}_s) \subset \tilde{G}_{(q'+1)s}. \tag{A.37}$$

By [C1, Lem. 4.1], we know that there exists an integer s such that

$$G_s \neq \phi. \tag{A.38}$$

Let us remark that the definition of $c_r(I)$ we give here is slightly different from the one we gave previously in [C1]. Our new $c_r(I)$ correspond to our old $c_r(I)$ computed with $b_i^j = \tilde{b}_i^{j+1}$. We choose s satisfying (A.38) and take $q = (q' + 1)s$. Then (A.33) follows from (A.36), (A.37), and (A.38).

The same proof as the proof of Corollary A.2 (just replace $\beta(I)$ by $|I|$) gives, as a corollary of Proposition A.5, the following.

COROLLARY A.6. *For any integer s, there exists an integer q such that the codimension of $\tilde{B}_q$ in $\tilde{J}_q$ is at least s.*

We now give the proof of Proposition A.1. We first define a sequence of rational functions $(\bar{a}_i; i \geq 1)$ in the variables $(\tilde{b}_0^i; i \geq 0)$ by

$$\bar{a}_1 = 1/\tilde{b}_0^1 \tag{A.39}$$

$$\bar{a}_i = (d\bar{a}_{i-1})/\tilde{b}_0^1 \quad \text{for } i \geq 2 \tag{A.40}$$

where d denotes the differentiation in the field of rational functions in the variables $(\tilde{b}_0^i; i \geq 0)$ such that $d\tilde{b}_0^i = \tilde{b}_0^{i+1}$ for all integer i. Let us remark that

$$\forall i \geq 1, \quad (\tilde{b}_0^1)^{2i-1}\bar{a}_i \text{ is a polynomial in the variables } (\tilde{b}_0^j; j \geq 0). \tag{A.41}$$

Let q' be an integer. By Corollary A.6, there exist an integer q and $\tilde{b}$ in $\tilde{J}_q$ such that

$$\tilde{b} \in \tilde{G}_q, \tag{A.42}$$

$$\tilde{b}_i^j = 0 \quad \text{for all} \quad (i,j) \quad \text{in } ([0,m] \times [0,q'])\backslash\{(0,1)\}, \tag{A.43}$$

$$\tilde{b}_0^1 = 1. \tag{A.44}$$

From (A.41) and (A.44), we see that the rational functions $(\bar{a}_i; i \geq 1)$ can be evaluated for this $\tilde{b}$. This leads to a sequence of real numbers $(a_i; i \geq 1)$. Let

$$\bar{b} = F_a(\tilde{b}). \tag{A.45}$$

From (A.25), (A.26), (A.39), (A.40), and (A.45), we obtain

$$\bar{b}_0^1 = 1, \tag{A.46}$$

$$\bar{b}_0^j = 0 \quad \text{if} \quad j \geq 2. \tag{A.47}$$

From (A.25), (A.26), (A.39), (A.40), (A.43), and (A.45), we get

$$\bar{b}_i^j \quad \text{for all} \quad (i,j) \in [1,m] \times [0,q']. \tag{A.48}$$

From Lemma A.4, (A.39), (A.42), and (A.45), we have

$$\pi_q(\bar{b}) \in \tilde{G}_q. \tag{A.49}$$

Finally, let $b \in J_q$ be defined by

$$b_i^j = (\pi_q(\bar{b}))_i^j \quad \forall i \in [1,m], \quad \forall j \geq 0. \tag{A.50}$$

Then, from (A.46), (A.47), (A.48), and (A.49), we have

$$b \in J_{q'}^{\perp} \cap G_q \tag{A.51}$$

which implies (A.5).

It remains to prove Lemma A.4. We first define a sequence $(A_{i,j}; i \in \mathbb{N}, j \in \mathbb{N})$ of polynomials in the variables $(a_i; i \geq 1)$ by

$$A_{1,j} = 0 \qquad j \geq 1 \tag{A.52}$$

$$A_{1,1} = a_1 \tag{A.53}$$

$$A_{i,j} = \dot{A}_{i-1,j} + a_1 A_{i-1,j-1} \quad \forall i \geq 2 \quad \forall j \geq 1 \tag{A.54}$$

where, by convention, $A_{0,j} = 0$ for all $j \geq 1$ and $\cdot$ denotes the differentiation on the set of polynomials in the variables $(a_i; i \geq 1)$ defined by $\dot{a}_i = a_{i+1}$ for all $i \geq 1$. For example,

$$A_{i,1} = a_i \quad \text{for all} \quad i \geq 1, \tag{A.55}$$

$$A_{i,i} = a_i^i \quad \text{for all} \quad i \geq 1. \tag{A.56}$$

For $i \geq 1$ and $j \geq 1$, let

$$\Delta_{i,j} = \left\{ s = (s_1, \ldots, s_j) \in \mathbb{N}^j; \quad 1 \leq s_1 \leq \ldots \leq s_j, \sum_{k=1}^{j} s_k = i \right\} \tag{A.57}$$

and

$$\Delta = \bigcup_{\substack{i \geq 1 \\ j \geq 1}} \Delta_{i,j}. \tag{A.58}$$

We easily check (by induction on i) that there exists a map $\alpha : \Delta \to \mathbb{N}$ such that

$$\alpha(s) > 0 \quad \forall s \in \Delta, \tag{A.59}$$

$$A_{i,j} = \sum_{s \in \Delta_{i,j}} \alpha(s) a_{s_1} \ldots a_{s_j}, \tag{A.60}$$

with the convention that, if $\Delta_{i,j} = \phi$, the right-hand side of (A.60) is 0.

We consider now a sequence of real numbers $(a_i; i \geq 1)$ and still denote by $A_{i,j}$ the value of the polynomial $A_{i,j}$ evaluated for this sequence of real numbers $(a_i; i \geq 1)$. Let $\tilde{b}$ in J and let $\bar{b} = F_a(\tilde{b})$. Let $\tilde{c} = (\tilde{c}_j(I); j \geq 0, 0 < |I| \leq \ell)$ and $\bar{c} = (\bar{c}_j(I); j \geq 0, 0 < |I| \leq \ell)$ be the associated sequences. We easily check (by induction on j) that

$$\bar{c}_j(I) = \sum_{1 \leq j' \leq j} A_{j,j'} \tilde{c}_{j'}(I). \tag{A.61}$$

Let r and q be two integers and let us assume that

$$a_r \neq 0 \tag{A.62}$$

and

$$\tilde{b} \in \tilde{G}_q. \tag{A.63}$$

We want to prove

$$\pi_{rq}(\bar{b}) \in \tilde{G}_{rq}. \tag{A.64}$$

By (A.24), without loss of generality, we may assume

$$a_i = 0 \quad \forall i \in [1, r-1]. \tag{A.65}$$

From (A.59), (A.60), (A.62), and (A.65), we obtain

$$A_{rj,j} \neq 0 \quad \forall j \geq 1 \tag{A.66}$$

and

$$A_{i,j} = 0 \quad \forall j \geq 1, \quad \forall i < rj. \tag{A.67}$$

Finally (A.64) follows (A.61), (A.63), (A.66), and (A.67).

Appendix B. In this appendix, we give a "nonaffine" version of the classical result of Jurdjevic and Quinn [JQ].

We first introduce some notations. Let N be a manifold, $\eta \in C^0(N; (0, +\infty))$, and

$$M = \{(x, u) \in N \times \mathbb{R}^m; |u| < \eta(x)\}.$$

Also let $F \in C^\infty(M; TN)$ be such that

$$F(x, u) \in T_x N \quad \forall (x, u) \in M. \tag{B.1}$$

Let, for $i \in [0, m]$, $F_i \in C^\infty(TN)$ be defined by

$$F_0(x) = F(x, 0) \quad \text{for all} \quad x \quad \text{in } N \tag{B.2}$$

$$F_i(x) = \frac{\partial F}{\partial u_i}(x, 0) \quad \text{for all} \quad x \quad \text{in } N, \quad \text{for all } i \text{ in } [1, m]. \tag{B.3}$$

We denote by φ the (maximal) solution of the flow associated to F_0, i.e.,

$$\frac{\partial \varphi}{\partial t} = F_0(\varphi) \tag{B.4}$$

$$\varphi(x, 0) = x. \tag{B.5}$$

Let $V \in C^\infty(N; \mathbb{R})$ and

$$Q = \{x \in N; L_{F_0} V(x) = 0, L_{ad^j_{F_0}(F_i)} V(x) = 0, \forall j \geq 0, \forall i \in [1, m]\}. \tag{B.6}$$

We assume

$$L_{F_0} V \leq 0 \quad \text{on } M, \tag{B.7}$$

$$\forall x \in N, \quad \exists t > 0 \quad \text{such that } \varphi(x, t) \notin Q. \tag{B.8}$$

For $\theta \in C^\infty(M;(0,+\infty))$ with

$$\theta(x)|(L_{F_1}V(x),\dots,L_{F_m}V(x))| < \eta(x), \tag{B.9}$$

we define $u^\theta \in C^1(M;\mathbb{R}^m)$ by

$$u_i^\theta(x) = -\theta(x)L_{F_i}V(x) \quad \text{for } i \in [1,m], \quad x \in N, \tag{B.10}$$

and denote by φ_θ the maximal solution of the flow associated with the vector fields $x \to F(x,u^\theta(x))$. Then we have the following lemma.

LEMMA B.1. *There exists* $\epsilon \in C^\infty(M;(0,+\infty))$ *such that if* $\theta \in C^1(M;(0,+\infty))$ *satisfies*

$$\theta \le \epsilon \quad \textit{on } N, \tag{B.11}$$

then (B.9) *holds and, with* $F_\theta(x) = F(x,u^\theta(x))$,

$$L_{F_\theta}V \le 0, \tag{B.12}$$

$$\forall x \in M, \quad \exists t > 0 \quad \textit{such that } V(\varphi_\theta(x,t)) < V(x). \tag{B.13}$$

Proof. We first note that there exist $C \in C^0(N;(0,+\infty))$ and ϵ_0 in $C^0(M;(0,+\infty))$ such that, if $\theta \in C^0(N;(0,+\infty))$ satisfies

$$\theta \le \epsilon_0, \tag{B.14}$$

then (B.9) holds and

$$L_{F_\theta}V \le L_{F_0}V - \sum_{i=1}^m \theta(L_{F_i}V)^2 + C\theta^2\sum_{i=1}^m (L_FV)^2. \tag{B.15}$$

Therefore, there exists $\epsilon \in C^0(N;(0,+\infty))$ such that, if $\theta \in C^0(N;(0,+\infty))$ satisfies (B.11), then (B.9) holds and

$$L_{F_\theta}V \le L_{F_0}V - \frac{\theta}{2}\sum_{i=1}^m (L_{f_i}V)^2, \tag{B.16}$$

which, by (B.7), gives (B.12). Now, assume moreover that θ is of class C^1 (so φ_θ is defined) and that for some $\overline{x}$ in M

$$V(\varphi_\theta(\overline{x},t)) = V(\overline{x}) \tag{B.17}$$

for all $t > 0$ such that $\varphi_\theta(\overline{x},t)$ is defined. Then, by (B.16)

$$L_{F_0}(\varphi_\theta(\overline{x},t)) = 0 \tag{B.18}$$

$$L_{F_i}V(\varphi_\theta(\overline{x},t)) = 0 \tag{B.19}$$

for all $t > 0$ such that $\varphi_\theta(\overline{x},t)$ is defined. In particular, by (B.10) and (B.19),

$$\varphi_\theta(\overline{x},t) = \varphi(\overline{x},t). \tag{B.20}$$

Let $\overline{x}(t) = \varphi(\overline{x}, t)$ and let I be the set of t in $[0, +\infty)$ such that $\overline{x}(t)$ is defined. By (B.18), (B.19), and (B.20)

$$L_{F_0} V(\overline{x}(t)) = 0 \quad \forall t \in I \tag{B.21}$$

$$\alpha_i(t) := L_{F_i} V(\overline{x}(t)) = 0 \quad \forall t \in T, \quad \forall i \in [1, m]. \tag{B.22}$$

Using (B.7) and (B.21), we get, for all X in $C^\infty(TN)$,

$$L_X L_{F_0} V(\overline{x}(t)) = 0 \quad \forall t \in I. \tag{B.23}$$

Using (B.22) and (B.23), we get

$$0 = \alpha_i^{(j)}(t) = L_{ad^j_{F_0}(F_i)} V(\overline{x}(t)) \quad \forall i \in [1, m], \quad \forall j \geq 0, \quad \forall t \in I. \tag{B.24}$$

Using (B.8), (B.21), and (B.24), we get a contradiction. Hence (B.13) holds.

Acknowledgment. We thank the Forschungsinstitut für Mathematik of Eidgenössische Technische Hochschule, Zürich, where part of this work has been carried out, for its hospitality.

REFERENCES

[B] R. W. Brocket, *Asymptotic stability and feedback stabilization,* in Differential Geometric Control Theory, R. W. Brockett, R. S. Milman, and H. J. Sussman, eds. Birkhäuser, Basel, Boston, 1983.

[C1] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift,* Math. Control Signals System, 5 (1992), pp. 295–312.

[C2] ———, *Links between local controllability and local continuous stabilization,* preprint, Université Paris-Sud and ETH Zürich, October 1991.

[CA] J.-M. Coron and B. D'Andréa-Novel, *Smooth stabilizing time-varying control laws for a class of nonlinear systems. Application to mobile robots,* preprint, ENSMP and Université Paris-Sud, Sept. 1991.

[G] M. Gromov, *Partial Differential Relations,* Ergebnisse Mathematik 3, Folge 9, Springer-Verlag, Berlin, 1986.

[GG] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities,* Graduate Texts in Mathematics 14, Springer-Verlag, New York, Heidelberg, Berlin, 1973.

[HH] G. W. Haynes and H. Hermes, *Nonlinear controllability via Lie theory,* SIAM J. Control Optim., 8 (1970), pp. 450–460.

[INS] A. Ilchmann, I. Nürnberger, and W. Schmale, *Time-varying polynomial matrix systems,* Internat. J. Control, 40 (1984), pp. 329–362.

[JQ] V. Jurdjevic and J. P. Quinn, *Controllability and stability,* J. Differential Equations, 28 (1978), pp. 381–389.

[KAI] T. Kailath, *Linear Systems,* Prentice Hall, London, 1980.

[KW] H. W. Knobloch and K. Wagner, *On local controllability of non-linear systems,* Dynamical Systems and Microphysics Control Theory and Mechanics, (1984), pp. 243–286.

[KUR] J. Kurzweil, *On the inversion of Lyapunov's second theorem on stability of motion,* Ann. Math. Soc. Trans. Ser. 2, 24 (1956), pp. 19–77.

[LA] K. K. Lee and A. Arapostahis, *Remarks on smooth feedback stabilization of nonlinear systems,* Syst. Control Lett. 10 (1988), pp. 41–44.

[LS] G. Lafferiere and H. J. Sussmann, *Motion planning for controllable systems without drift: A preliminary report,* Tech. Report Sycon 90-04, Rutgers University, June 1990.

[OS] R. Outbib and G. Sallet, *Stabilizability of the angular velocity of a rigid body revisited,* Systems Control Lett., 18 (1992), pp. 93–98.

[P] J.-B. Pomet, *Explicit design of time-varying stabilizing control laws for a class of controllable systems without drift,* Systems Control Lett. 18 (1992), pp. 147– 158.

[S1] E. D. Sontag, *Finite dimensional open-loop control generators for nonlinear systems,* Internat J. Control, 47 (1988), pp. 537–556.

[S2] ———, *Universal nonsingular controls,* Systems and Control Letters, 19 (1992), pp. 221–224.

[S3] H. J. Sussmann, *Single-input observability of continuous-time systems,* Math. Systems Theory, 12 (1979), pp. 371–393.

[SJ] H. J. Sussmann and V. Jurdjevic, *Controllability of nonlinear systems,* J. Differential Equations, 12 (1972), pp. 95–116.

[SL] H. J. Sussmann and W. Liu, *Limits of highly oscillatory controls and the approximation of general paths by admissible trajectories,* Proceedings of the 30th Conference on Decision and Control, Brighton, United Kingdom, 30 (1991) pp. 437–442.

[WS] Y. Wang and E. D. Sontag, *Order of Input/Output Differential Equations and State Space Dimensions*, preprint, Florida Atlantic University, March 1993.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 387–407, March 1994

004

EQUIVALENCE OF NONLINEAR SYSTEMS TO INPUT-OUTPUT PRIME FORMS*

R. MARINO†, W. RESPONDEK‡, AND A. J. VAN DER SCHAFT§

Abstract. The problem of transforming nonlinear control systems into input-output prime forms is dealt with, using state space, static state feedback, and also output space transformations. Necessary and sufficient geometric conditions for the solvability of this problem are obtained. The results obtained generalize well-known results both on feedback linearization as well as input-output decoupling of nonlinear systems. It turns out that, from a computational point of view, the output space transformation is the crucial step, that is performed by constructing rectifying coordinates for a nested sequence of distributions on the output manifold.

Key words. equivalence, output transformation, input-output prime system, integrable distributions, input-output decoupling

AMS subject classifications. 93C10, 93B17, 58A30

1. Introduction.

We consider smooth (i.e., C^∞) nonlinear systems, depending in an affine way on the inputs $u_1, \dots, u_m$, and having m outputs $y_1, \dots, y_m$

$$(\Sigma) \qquad \begin{aligned} \dot{x} &= f(x) + \sum_{j=1}^{m} g_j(x) u_j, \qquad u = (u_1, \dots, u_m) \in \mathbb{R}^m, \\ y_j &= h_j(x), \qquad j = 1, \dots, m \end{aligned} \tag{1.1}$$

where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_m)$ are local coordinates for the state space manifold M and for the output space manifold Y, respectively. We assume throughout the exisxtence of an equilibrium point $x_0 \in M$ such that $f(x_0) = 0$ and $h(x_0) = 0$. (All results can be adapted to the case $f(x_0) \neq 0$ and/or $h(x_0) \neq 0$; see Remark 2 after the proof of Theorem 6.) Our analysis will be mainly of a local nature (see, however, Theorem 10 and Corollary 11 for global extensions), i.e., we firstly study the system in neighborhoods $V_{x_0} \subset M$ and $W_{y_0} \subset Y$, where $y_0 = h(x_0)$. We also assume throughout that M and Y are connected, and that rank $dh(x)$, with $h = (h_1, \dots, h_m)$, equals m in V_{x_0}, and that the dimension of the distribution $G(x) := \text{span}\,\{g_1(x), \dots, g_m(x)\}$ is m in V_{x_0}. Note that we are restricting ourselves entirely to square systems, i.e., the number of inputs equals the number of outputs.

We address the (local) equivalence of Σ to prime (linear) systems, and to input-output prime (linear) systems. We use the following notion of equivalence.

DEFINITION 1. Consider two systems Σ_1, Σ_2 defined on $(M_1, Y_1), (M_2, Y_2)$ with equilibrium points $x_{01} \in M_1, x_{02} \in M_2$, respectively. We say that Σ_1 is *locally equivalent* to Σ_2, around x_{01} and x_{02}, if there exist:

(i) Neighborhoods $V_{x_{01}} \subset M_1, V_{x_{02}} \subset M_2$ and a diffeomorphism $\varphi : V_{x_{01}} \to V_{x_{02}}$ satisfying $\varphi(x_{01}) = x_{02}$;

* Received by the editors June 24, 1991; accepted for publication (in revised form) June 15, 1992.

† Dipartimento di Ingegneria Elettronica, Seconda Università di Roma, "Tor Vergata," Via O. Raimondo, 00173 Roma, Italy.

‡ Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, 00-950 Warsaw, Poland. This author's work was performed while a visiting professor at the Dipartimento di Ingegneria Elettronica, II Università di Roma "Tor Vergata."

§ Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands.

(ii) a nonsingular state feedback $u = \alpha(x)+\beta(x)v$ defined on $V_{x_{01}}$ such that $\alpha(x_{01}) = 0$ and $\det \beta(x) \neq 0$;

(iii) neighborhoods $W_{y_{01}} \subset Y_1, W_{y_{02}} \subset Y_2$ of $y_{01} = h^1(x_{01})$ and $y_{02} = h^2(x_{02})$, where h^1 and h^2 denote the output maps of Σ_1 and Σ_2 respectively, and a diffeomorphism $\psi : W_{y_{01}} \to W_{y_{02}}$ satisfying $\psi(y_{01}) = y_{02}$,

such that the transformation of Σ_1 under $(\varphi, (\alpha, \beta), \psi)$ equals Σ_2 on the specified neighborhoods.

We recall from [Mo] (see also [He]) the notion of linear prime system.

DEFINITION 2. A system Σ is called a (linear) *prime system* if it is of the form

$$(P)\quad \begin{aligned} y_i &= x_{i1} \\ \dot{x}_{i1} &= x_{i2} \qquad i = 1, \dots, m \\ &\vdots \\ \dot{x}_{i\kappa_i} &= u_i \end{aligned} \tag{1.2}$$

where $x = (x_{11}, \dots, x_{1\kappa_1}, \dots, x_{m1}, \dots, x_{m\kappa_m}) \in M = \mathbb{R}^n, n = \sum_{i=1}^m \kappa_i$, for some integers $\kappa_1, \dots, \kappa_m$, and $y = (y_1, \dots, y_m) \in Y = \mathbb{R}^m$. The integers $\kappa_1, \dots, \kappa_m$ equal the *orders of the zeros at infinity* of the system or the *relative degrees*, as well as the *controllability* or *observability indices.*

More generally we define input-output prime systems.

DEFINITION 3. A system Σ is called an *input-output prime system* if it is of the form

$$(I-O-P)\quad \begin{aligned} y_i &= x_{i1} \\ \dot{x}_{i1} &= x_{i2} \qquad i = 1, \dots, m \\ &\vdots \\ \dot{x}_{i\mu_i} &= u_i \end{aligned} \tag{1.3a}$$

$$\dot{z} = a(z, x) + \sum_{j=1}^{m} b_j(z, x) u_j, \qquad a(z_0, 0) = 0, \tag{1.3b}$$

where $y = (y_1, \dots, y_m) \in Y = \mathbb{R}^m$, and where the state space manifold M has the following special structure. There exists a surjective submersion $\pi : M \to \mathbb{R}^\mu, \mu := \sum_{i=1}^m \mu_i$, with $x = (x_{11}, \dots, x_{m\mu_m}) \in \mathbb{R}^\mu$, and z being complementary local coordinates for M. The integers $\mu_1, \dots, \mu_m$ equal the *orders of the zeros at infinity* or the *relative degrees* of the system, as well as the *observability indices.*

Remark. Observe that the relative degrees are *not* invariant in our problem because we allow for output transformations (see the Example preceeding Algorithm 7); *nor* are the observability indices since they can be changed by feedback. However, the structure at infinity *does* remain unchanged under the considered transformations, and thus this is the right concept to describe the μ_i's as invariants in our problem. Here, the structure of infinity can be defined either geometrically using the V^*-algorithm [NS], [ls2], or by means of dynamic extension $[M]$ since, for input-output prime systems (and their equivalents), both definitions coincide.

We will also be interested in *input-output prime systems of special form*

$$
\begin{aligned}
& & y_i &= x_{i1} & & \\
(1.4a) \qquad & (I-O-P-S) & \dot{x}_{i1} &= x_{i2} & & i=1,\dots,m \\
& & & \vdots & & \\
& & \dot{x}_{i\mu_i} &= u_i & & \\
(1.4b) \qquad & & \dot{z} &= a(z,y) & &
\end{aligned}
$$

with the same specifications as in Definition 3, the difference being that the z-dynamics are only driven by the outputs $y=(y_1,\dots,y_m)$.

The main results of the paper are concerned with identifying, via necessary and sufficient geometric conditions, those nonlinear systems Σ which are locally equivalent to prime systems (Theorem 4), to input-output prime systems (Theorem 6), and to input-output prime systems of special form (Proposition 8). Theorem 10 and Corollary 11 deal with global equivalence issues. The results obtained generalize well-known results both on normal forms for input-output decouplable systems as well as on feedback linearization of systems with no outputs, as we will now briefly indicate.

If outputs are not considered in Σ, and therefore output change of coordinates (iii) is omitted in Definition 1, the problem of local equivalence with prime systems becomes the well-known local *feedback linearization* problem, i.e., local feedback equivalence into linear (Brunovsky) canonical forms

$$
\begin{aligned}
& & \dot{x}_{i1} &= x_{i2} & & \\
(1.5) \qquad & (B) & & \vdots & & i=1,\dots,m, \\
& & \dot{x}_{i\kappa_i} &= u_i & &
\end{aligned}
$$

which was completely solved in [JR] and [HSM]. The solution to this problem is a generalization of a linear result of Brunovsky [Br], stating that any controllable linear system

$$
(1.6) \qquad \dot{x} = Ax + Bu, \quad x \in \mathbb{R}^n, u \in \mathbb{R}^m, \quad \text{rank } B = m,
$$

can be transformed into (B) by the action of the linear feedback group taking the pair (A,B) into $(T(A+BF)T^{-1}, TBG)$ for a linear state space change of coordinates $\tilde{x}=Tx$ and a linear feedback $u=Fx+Gv, \det G \neq 0$. The set of indices $(\kappa_1,\dots,\kappa_m)$, called *controllability indices*, is uniquely associated with (6) and forms a complete set of invariants under the action of the linear feedback group (see also [Wo]). In [Mo] Morse enlarges this group by allowing also for linear output space change of coordinates $\tilde{y}=Hy$, and gives necessary and sufficient conditions for a linear system

$$
\begin{aligned}
& & \dot{x} &= Ax + Bu, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m, \quad y \in \mathbb{R}^m, \\
(1.7) \qquad & (L) & & \\
& & y &= Cx, \qquad \text{rank } B = \text{rank } C = m,
\end{aligned}
$$

to be transformed into a prime system (P) given by (1.2) by the action of the group taking (A,B,C) into $(T(A+BF)T^{-1}, TBG, HCT^{-1})$. We generalize this result of Morse to nonlinear systems Σ in Theorem 4, on the basis of the local feedback linearization theorem [JR], [HSM]. We remark that nonlinear output change of coordinates was introduced in [KR] in the study of asymptotic observers. Furthermore, the problem of local feedback

equivalence (with no output change of coordinates) of Σ to a linear system was studied and solved in [CIRT].

The problem of (local) feedback equivalence, *without* output change of coordinates, of a nonlinear system Σ to an input-output prime system $I - O - P$ has been solved in [IKGM]. Indeed, this problem amounts to the (local) nonlinear *input-output decoupling problem*, as dealt with in [SR], [Fr], and [Si]. The basic tool is the *decoupling matrix*, which generalizes to nonlinear systems the Falb–Wolovich matrix [FW], used in input-output decoupling of linear systems (L). In fact a necessary and sufficient condition for Σ to be input-output decouplable around x_0 is that its decoupling matrix is nonsingular in a neighborhood of x_0. We note that the problem of local equivalence of Σ to $(I - O - P)$ studied in the present paper can be rephrased in this latter terminology as finding a (local) output transformation $\tilde{y} = \psi(y)$ such that Σ, with the resulting *transformed* output functions $\tilde{h}_1 = \psi_1 o h, \ldots, \tilde{h}_m = \psi_m o h$, is locally input-output decouplable. Finally, (local and global) feedback equivalence with no output change of coordinates of Σ into input-output prime systems of special form $(I - O - P - S)$ has been dealt with in [BI], while for linear systems (L) equivalence to $(I - O - P)$ *implies* equivalence to $(I - O - P - S)$, as was implicitly derived in [Mo] (see Remark 2 after Proposition 8).

The results obtained are useful for control applications in the following sense. It is well known (see, e.g., [Is], [NvdS]) that many nonlinear control problems are relatively easily attacked for input-output decouplable systems. Now, in many of these control problems output transformations are naturally allowed, and thus our results enable us to treat in a similar way a class of nonlinear control systems which properly *contains* the input-output decouplable systems. One obvious example of a control problem which naturally *does* allow for output transformations is the (asymptotic) output tracking problem by static state feedback control (see the example after Theorem 6).

2. Main results. Let us first recall the definitions of the following sequences of distributions for a nonlinear system Σ:

$$\begin{aligned} G_1 &:= G := \operatorname{span}\{g_1, \ldots, g_m\} \\ G_{i+1} &:= G_i + [f, G_i], \qquad i = 1, 2, \ldots \end{aligned} \tag{2.1}$$

$$\begin{aligned} S_1 &:= G, \\ S_{i+1} &:= S_i + [f, S_i \cap \ker dh] + \sum_{j=1}^{m} [g_j, S_i \cap \ker dh], \qquad i = 1, 2, \ldots \\ S^* &:= \bigcup_{i \geq 1} S_i. \end{aligned} \tag{2.2}$$

The distributions G_i were introduced in [JR] in the study of the feedback linearization problem, while the algorithm (2.2) and the definition of S^* is taken from [IKGM] (with the difference that, in [IKGM], S_i in the right-hand side of (2.2) is replaced by its involutive closure; see, however, conditions (i), (iii) of Theorem 4). S^*, the smallest conditioned invariant distribution containing G, enjoys the property (see [IKGM])

$$\begin{aligned} [f, S^* \cap \ker dh] &\subset S^*, \\ [g_j, S^* \cap \ker dh] &\subset S^*, \qquad j = 1, \ldots, m, \end{aligned} \tag{2.3}$$

and is a generalization of the notion of the smallest conditioned invariant subspace containing Im B, as introduced in [BM] for a linear system (L). If the distributions $S_i, i = 0, 1, \ldots,$ all have constant dimension, then there exists an integer $i^* \leq n$ such that $S_{i^*} = S^*$.

Following [IKGM] we also recall the construction (if it exists) of V^*, the largest locally controlled invariant distribution contained in kerdh (see also [Hi], and for the linear case [Wo], [BM]). Define the sequence of codistributions

$$\begin{aligned} P_1 &:= dh \\ P_{i+1} &:= P_i + L_f(P_i \cap \text{ann}G) + \sum_{j=1}^{m} L_{g_j}(P_i \cap \text{ann}G), \qquad i = 1, 2, \dots \\ P^* &:= \bigcup_{i \geq 1} P_i \end{aligned} \tag{2.4}$$

where L_f, L_{g_j} denote Lie derivatives, and ann$G$ is the codistribution annihilating G. Then the distribution V^* is the kernel of the codistribution P^*, i.e., $V^* = \ker P^*$ (see, e.g., [Is], [NvdS]).

We finally recall the definition of *characteristic indices* (or *relative degrees*) ρ_i, and of the *decoupling matrix*. For $i = 1, \dots, m$, ρ_i is defined by

$$\begin{aligned} &L_{g_j} L_f^k h_i(x) = 0, \quad k = 0, 1, \dots, \rho_i - 2, j = 1, \dots, m, \quad \text{for all } x \in V_{x_0} \\ &L_{g_j} L_f^{\rho_i - 1} h_i(x) \neq 0, \quad \text{for some } j \in \{1, \dots, m\} \quad \text{and } x \in V_{x_0}. \end{aligned} \tag{2.5}$$

If $\rho_i < \infty, i = 1, \dots, m$, the decoupling matrix $D(x)$ is defined as

$$D(x) = \left(L_{g_j} L_f^{\rho_i - 1} h_i(x) \right)_{i,j=1,\dots,m}. \tag{2.6}$$

We now come to our first main theorem.

THEOREM 4. *Consider a nonlinear system Σ with equilibrium x_0. Σ is locally equivalent to a prime system (P) with equilibrium 0, if and only if the following conditions are satisfied in a neighborhood of x_0:*

(i) *G_i is involutive and of constant dimension, $i = 1, \dots, n-1$;*
(ii) *$G_n = TM$;*
(iii) *$G_i = S_i, i = 1, 2, \dots, n$;*
(iv) *$G_i + \ker dh$ is involutive and of constant dimension, $i = 1, \dots, n-1$.*

Remark 1. Theorem 4 generalizes and clarifies the following result of Morse ([Mo, Thm. 3.1]): The system (L), i.e., the triple (A, B, C), is transformable by $(T, (F, G), H) : (A, B, C) \to (T(A + BF)T^{-1}, TBG, HCT^{-1})$ into a prime system (P) if and only if:

(i)′ $V^* = 0$;
(ii)′ $G_n = \text{Im}(B, AB, \dots, A^{n-1}B) = \mathbb{R}^n$;
(iii)′ $G_i = S_i, i = 1, \dots, n$.

Conditions (i) and (iv) of Theorem 4 are always satisifed for linear systems, while they are crucial integrability conditions in the nonlinear case. Conditions (ii)' and (iii)' of Morse are specializations of conditions (ii) and (iii) of Theorem 4 to the linear case. Condition (i), i.e., $V^* = 0$, is redundant; it is implied by conditions (ii) and (iii). In fact the proof that we will give is entirely different from Morse's and enables us to point out the redundancy of the condition $V^* = 0$ in the original statement of Morse.

Remark 2. Conditions (i) and (ii) are the necessary and sufficient conditions given in [JR] for the system Σ without outputs to be locally feedback equivalent to a linear system in Brunovsky form (B).

Remark 3. While Remarks 1 and 2 clarify the necessity of conditions (i), (ii), and (iii), we may wonder if condition (iv) is not redundant, since already condition (iii) enforces a rather strong compatibility between G_i and kerdh. However, the following example shows

that for $\rho > 2$ condition (iv) is *not* implied by conditions (i), (ii), and (iii). Consider the system

$$\begin{aligned} \dot{x}_1 &= u_1, & y_1 &= x_1 \\ \dot{x}_2 &= x_3, & y_2 &= x_2 + x_1 x_5 \\ \dot{x}_3 &= u_2 & & \\ \dot{x}_4 &= x_5, & y_3 &= x_4 \\ \dot{x}_5 &= x_6 & & \\ \dot{x}_6 &= u_3. & & \end{aligned} \tag{2.7}$$

We easily compute

$$\begin{aligned} G_1 = S_1 &= \operatorname{span}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_6}\right\} \\ \ker dh &= \operatorname{span}\left\{\frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_5} - x_1\frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_6}\right\} \\ G_2 = S_2 &= \operatorname{span}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_5}, \frac{\partial}{\partial x_6}\right\} \\ G_3 = S_3 &= T\mathbb{R}^6 \\ G_1 + \ker dh &= \operatorname{span}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_5}, - x_1\frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_6}\right\} \end{aligned} \tag{2.8}$$

and thus conditions (i), (ii), and (iii) are satisfied, while condition (iv) fails since $G_1 + \ker dh$ is not involutive; in fact,

$$\left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_5} - x_1\frac{\partial}{\partial x_2}\right] = -\frac{\partial}{\partial x_2} \notin G_1 + \ker dh. \tag{2.9}$$

It follows that (2.7) is *not* locally equivalent to a prime system (P).

Remark 4. It is easy to see that if a nonlinear system (Σ) with equilibrium x_0 is locally equivalent to a prime system, then its linearization at x_0, namely, $(\overline{x} = x - x_0, \overline{y} = y - h(x_0))$

$$\begin{aligned} \dot{\overline{x}} &= \frac{\partial f}{\partial x}(x_0)\overline{x} + \sum_{j=1}^{m} g_i(x_0)u_j \\ \overline{y} &= \frac{\partial h}{\partial x}(x_0)\overline{x} \end{aligned}$$

is also equivalent to a prime system. The converse may be false as the following example shows:

$$\begin{aligned} \dot{x}_1 &= u_1, & y_1 &= x_1 \\ \dot{x}_2 &= x_3 + (1 - e^{x_3})u_1, & y_2 &= x_2 \\ \dot{x}_3 &= u_2. & & \end{aligned}$$

In this case the distribution $G_1 = \operatorname{span}\{(\partial/\partial x_1) + (1 - e^{x_3})(\partial/\partial x_2), (\partial/\partial x_3)\}$ is not involutive, so that condition (i) of Theorem 4 is violated. On the other hand the system linearized at the origin is obviously a prime system.

Before giving the proof of Theorem 4, we first recall a lemma which clarifies the meaning of condition (iv). Let $h : M \to Y$ be such that rank $dh(x) = m = \dim Y$ on a neighborhood V_{x_0}. Then $W_{y_0} := h(V_{x_0})$ is a neighborhood of $y_0 = h(x_0)$ in Y.

Furthermore let D be a distribution on M. Then D is said to be *projectable* by h on V_{x_0} if, for all x_1, x_2 in V_{x_0}, we have

$$h(x_1) = h(x_2) \Rightarrow \frac{\partial h}{\partial x}(x_1)(D(x_1)) = \frac{\partial h}{\partial x}(x_2)(D(x_2)). \tag{2.10}$$

If D is projectable by h on V_{x_0}, then we define h_*D as the following distribution on W_{y_0}:

$$(h_*D)(y) = \partial h/\partial x(x)(D(x)), \quad \textit{with } x \in h^{-1}(y) \cap V_{x_0}, \quad y \in W_{y_0}. \tag{2.11}$$

(For the problem of projecting distributions see also [J].)

LEMMA 5. *Let h be such that* rank $dh(x) = \dim Y$ *on V_{x_0}, and let D be involutive and constant dimensional on V_{x_0}. Then D is projectable by h in V_{x_0} to a constant dimensional and involutive distribution h_*D on W_{y_0} if and only if $D + \ker dh$ is involutive and of constant dimension on V_{x_0}.*

Proof. First notice that by the Rank Theorem [Sp] we can take local coordinates $x = (x^1, x^2)$ on V_{x_0} such that $h(x^1, x^2) = x^1$. Then it follows from [vdS] (see also [NvdS, Lem. 14.3]) that if $D + \ker dh$ is involutive and of constant dimension then D is projectable and h_*D is involutive and of constant dimension. Conversely (see [vdS]) it follows trivially that if D is projectable then $D + \ker dh$ is involutive (indeed $D + \ker dh$ is of the form span $\{k(x^1)(\partial/\partial x^1) + \text{span}\{\partial/\partial x^2\}$ for some k). Furthermore, if h_*D is constant dimensional then $D + \ker dh$ is constant dimensional. □

Proof of Theorem 4 (only if). First we note from (2.2) that the definition of S_i is invariant under feedback and output transformations. Suppose that Σ is locally equivalent to (P). Clearly, (P) satisfies conditions (i)–(iv). It follows that also the definition of G_1, and inductively of $G_i, i > 1$, is invariant under feedback, and thus conditions (i)–(iv) are feedback invariant. Thus we can conclude that conditions (i)–(iv) are also satisfied for Σ.

(If.) By virtue of (i) and (iii) it follows that S_{i+1} is alternatively given as

$$S_{i+1} = S_i + [f, S_i \cap \ker dh], \qquad i = 1, 2, \ldots \tag{2.12}$$

since $[g_j, S_i \cap \ker dh] = [g_j, G_i \cap \ker dh] \subset G_i = S_i, j = 1, \ldots, m$.

By conditions (i), (iii), (iv) and Lemma 5,

$$H_i := h_*G_i = h_*S_i, \qquad i = 1, \ldots, n \tag{2.13}$$

are well-defined involutive and constant-dimensional distributions on a neighborhood $W_{y_0} \subset Y$, while by (ii) $H_n = TY$. Obviously $H_1 \subset H_2 \subset \ldots \subset H_n$. It follows that we can define integers

$$\kappa_1 > \kappa_2 > \ldots > \kappa_r > 0 \tag{2.14}$$

such that

$$0 = H_1 = \ldots = H_{\kappa_r - 1} \overset{\subset}{\neq} H_{\kappa_r} = \ldots = H_{\kappa_1 - 1} \overset{\subset}{\neq} H_{\kappa_1} = TY. \tag{2.15}$$

For ease of notation we will assume that

$$\dim H_{\kappa_i} = \dim H_{\kappa_i - 1} + 1, \qquad i = 1, \ldots, m \tag{2.16}$$

implying that $r = m$, and

$$\kappa_1 > \kappa_2 > \ldots > \kappa_m > 0, \qquad \dim H_{\kappa_i} = m - i + 1, i = 1, \ldots, m. \tag{2.17}$$

(Later on we will conclude that $\kappa_1, \ldots, \kappa_m$ are actually the controllability indices of Σ.)

Invoking the generalized Frobenius theorem for the nested sequence of distributions $H_{\kappa_m} \subset H_{\kappa_{m-1}} \subset \ldots \subset H_{\kappa_1}$ (see [JR], [NvdS]), we can choose locally about y_0 in W_{y_0} m independent functions

$$\psi_1, \ldots, \psi_m \tag{2.18}$$

such that

$$\begin{aligned} &\langle d\psi_1, H_{\kappa_1 - 1}\rangle = 0 \\ &\langle d\psi_2, H_{\kappa_2 - 1}\rangle = 0 \\ &\qquad \vdots \\ &\langle d\psi_{m-1}, H_{\kappa_{m-1} - 1}\rangle = 0 \end{aligned} \tag{2.19a}$$

while

$$\begin{aligned} &\langle d\psi_1, H_{\kappa_1}\rangle(x_0) \neq 0 \\ &\langle d\psi_2, H_{\kappa_2}\rangle(x_0) \neq 0 \\ &\qquad \vdots \\ &\langle d\psi_m, H_{\kappa_m}\rangle(x_0) \neq 0. \end{aligned} \tag{2.19b}$$

In the new local coordinates of the output manifold Y, given by

$$\tilde{y}_i := \psi_i(y), \qquad i = 1, \ldots, m, \tag{2.20}$$

we obviously have

$$H_{\kappa_i} = \operatorname{span}\left\{\frac{\partial}{\partial \tilde{y}_m}, \ldots, \frac{\partial}{\partial \tilde{y}_i}\right\}, \qquad i = 1, \ldots, m. \tag{2.21}$$

If we define new output functions

$$\tilde{h}_i := \psi_i o h, \qquad i = 1, \ldots, m \tag{2.22}$$

and recall that $H_i = h_* G_i$ (cf. (2.13)), then (2.19a) yields

$$\begin{aligned} &\langle d\tilde{h}_1, G_{\kappa_1 - 1}\rangle = 0 \\ &\langle d\tilde{h}_2, G_{\kappa_2 - 1}\rangle = 0 \\ &\qquad \vdots \\ &\langle d\tilde{h}_{m-1}, G_{\kappa_{m-1} - 1}\rangle = 0 \end{aligned} \tag{2.23a}$$

while by (2.19b),

$$\begin{aligned} &\langle d\tilde{h}_1, G_{\kappa_1}\rangle(x_0) \neq 0 \\ &\langle d\tilde{h}_2, G_{\kappa_2}\rangle(x_0) \neq 0 \\ &\qquad \vdots \\ &\langle d\tilde{h}_m, G_{\kappa_m}\rangle(x_0) \neq 0. \end{aligned} \tag{2.23b}$$

Let us now compute the decoupling matrix $\tilde{D}(x)$ (cf. (2.6)) of Σ with the newly defined output functions $\tilde{h}_1, \ldots, \tilde{h}_m$ (cf. (2.22)). It readily follows from (2.23), and the Leibniz rule, i.e.,

$$L_X\langle d\varphi, Y\rangle = \langle dL_X\varphi, Y\rangle + \langle d\varphi, ad_X Y\rangle \tag{2.24}$$

for any two vector fields X, Y and function φ, that $\tilde{D}(x)$ is given as (see, e.g., [Is], [NvdS])

$$\tilde{D}(x) = \begin{pmatrix} (-1)^{(\kappa_1 - 1)}\langle d\tilde{h}_1, ad_f^{\kappa_1 - 1} g_1\rangle(x) \ldots (-1)^{(\kappa_1 - 1)}\langle d\tilde{h}_1, ad_f^{\kappa_1 - 1} g_m\rangle(x) \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ (-1)^{(\kappa_m - 1)}\langle d\tilde{h}_m, ad_f^{\kappa_m - 1} g_1\rangle(x) \ldots (-1)^{(\kappa_m - 1)}\langle d\tilde{h}_m, ad_f^{\kappa_m - 1} g_m\rangle(x) \end{pmatrix}. \tag{2.25}$$

We now make the following claim.

CLAIM. *$\tilde{D}(x)$ is nonsingular in a neighborhood U_{x_0} of $x_0 \in M$.*

Once this claim has been proved the rest of the proof of Theorem 4 follows easily. Indeed by the theory of input-output decoupling (see, e.g., [IKGM], [Is], [NvdS]) the functions

$$(\tilde{h}_1, \dots, L_f^{\kappa_1 - 1}\tilde{h}_1, \dots, \tilde{h}_m, \dots, L_f^{\kappa_m - 1}\tilde{h}_m) =: (x_{11}, \dots, x_{1\kappa_1}, \dots, x_{m1}, \dots, x_{m\kappa_m}) \tag{2.26a}$$

are independent on U_{x_0}, and the state feedback

$$u = -\tilde{D}^{-1}(x) \begin{pmatrix} L_f^{\kappa_1}\tilde{h}_1(x) \\ \vdots \\ L_f^{\kappa_m}\tilde{h}_m(x) \end{pmatrix} + \tilde{D}^{-1}(x)v \tag{2.26b}$$

brings the system into the form

$$\begin{aligned} \tilde{y}_i &= x_{i1} \\ \dot{x}_{i1} &= x_{i2} \qquad i = 1, \dots, m, \\ &\vdots \\ \dot{x}_{i\kappa_i} &= v_i \\ \dot{z} &= a(z, x) + b(z, x)v \end{aligned} \tag{2.27}$$

(where $z \in \mathbb{R}^{n-(\kappa_1 + \dots + \kappa_m)}$ are additional coordinates).

Furthermore, it is immediately seen that the S^*-algorithm (cf. (2.2)) applied to (2.27) yields $\dim S^* = \kappa_1 + \dots + \kappa_m$. Then because of feedback invariance of S_i and (ii), i.e., $S_n = G_n = TM$, it follows that $\kappa_1 + \dots + \kappa_m = n$, and thus the z-part in (2.27) is void, implying that Σ with the newly defined output functions $\tilde{h}_1, \dots, \tilde{h}_m$ is feedback equivalent to a prime system (P), with controllability indices $\kappa_1, \dots, \kappa_m$.

Proof of the claim. We use the following induction argument.

Step 1. Consider $\tilde{h}_1$. By (2.23) there exists some $i \in \{1, \dots, m\}$ such that $\langle d\tilde{h}_1, ad_f^{\kappa_1 - 1} g_i \rangle(x_0) \neq 0$. By relabeling $g_1, \dots, g_m$, if necessary, we may thus assume that

$$\langle d\tilde{h}_1, ad_f^{\kappa_1 - 1} g_1 \rangle(x_0) \neq 0. \tag{2.28}$$

Define the functions $\beta_i^1(x) := \langle d\tilde{h}_1, ad_f^{\kappa_1 - 1} g_i \rangle(x), i = 1, \dots, m$, and put locally about x_0

$$\tilde{g}_i := g_i - \frac{\beta_i^1}{\beta_1^1} g_1, \qquad i = 2, \dots, m \tag{2.29}$$

(observe that by (2.28) $\beta_1^1 \neq 0$ locally about x_0). Then, because

$$ad_f^{\kappa_1 - 1}\tilde{g}_i = ad_f^{\kappa_1 - 1} g_i - \frac{\beta_1^1}{\beta_1^1} ad_f^{\kappa_1 - 1} g_i (\mathrm{mod} G_{\kappa_1 - 1}), \tag{2.30}$$

the transformed input vectorfields $\tilde{g}_2, \dots, \tilde{g}_m$ satisfy

$$\langle d\tilde{h}_1, ad_f^{\kappa_1 - 1}\tilde{g}_i \rangle \equiv 0, \quad \text{around } x_0, \quad i = 2, \dots, m. \tag{2.31}$$

For ease of notation we will now omit the tildes above g_i, and thus denote $\tilde{g}_2, \dots \tilde{g}_m$ again by $g_2, \dots, g_m$.

Step $k + 1$. Assume that the functions $\beta_i^j := \langle d\tilde{h}_j, ad_f^{\kappa_j - 1} g_i \rangle$ satisfy

(A1) $\beta_j^j(x_0) \neq 0, \qquad j = 1, \dots, k$

(A2) $\beta_i^j \equiv 0, \quad i = j + 1, \dots, m, \quad j = 1, \dots, k.$

We will show that, after applying feedback, (A1) and (A2) also hold for $j = k+1$. First we note that since $\langle d\tilde{h}_j, G_{\kappa_j - 1}\rangle = 0$ (cf. (2.23a)) repeated use of the Leibniz rule yields

$$\langle dL_f^{\ell}\tilde{h}_j, ad_f^{\kappa_j - \ell - 1}g_i\rangle = (-1)^{\ell}\beta_i^j, \quad \ell = 0, 1, \ldots, \kappa_j - 1, \quad i, j = 1, \ldots, m. \tag{2.32}$$

By using (A1) and (A2) this implies that we have the following "table" for the expressions $\langle dL_f^{\kappa_1 - \kappa_k}\tilde{h}_j, ad_f^{\kappa_k - 1}g_i\rangle, j = 1, \ldots, k, i = 1, \ldots, m$:

$$\begin{array}{lccccccc} & ad_f^{\kappa_k-1}g_1 & ad_f^{\kappa_k-1}g_2 & \ldots & ad_f^{\kappa_k-1}g_k & \ldots & ad_f^{\kappa_k-1}g_m \\ dL_f^{\kappa_1-\kappa_k}\tilde{h}_1 & * & 0 & & & & \\ dL_f^{\kappa_2-\kappa_k}\tilde{h}_2 & & * & & & & \\ \vdots & ? & & & & & 0 \\ d\tilde{h}_k & & & & * & & \end{array} \tag{2.33}$$

where the * elements are all nonzero by (A1). It follows that the map $F = (L_f^{\kappa_1 - \kappa_k}\tilde{h}_1, L_f^{\kappa_2 - \kappa_k}\tilde{h}_2, \ldots, \tilde{h}_k)$ has rank k (the same argument is used in feedback linearization, cf. [HSM], [Is], [NvdS]) and furthermore, since $G_{\kappa_k} + \ker dF = TM$, Lemma 5 implies that G_{κ_k} is projectable by F, while

$$\dim F_* G_{\kappa_k} = k. \tag{2.34}$$

Now consider $\tilde{h}_{k+1}$. Because of (2.23) it follows that there exists some $i \in \{1, \ldots, m\}$ such that

$$\langle d\tilde{h}_{k+1}, ad_f^{\kappa_{k+1} - 1}g_i\rangle(x_0) \neq 0. \tag{2.35}$$

We *claim* that we can take $i \in \{k+1, \ldots, m\}$ having this property. Indeed, otherwise we would have

$$\langle d\tilde{h}_{k+1}, ad_f^{\kappa_{k+1} - 1}g_i\rangle(x_0) = 0, \qquad i = k+1, \ldots, m. \tag{2.36}$$

Now take any $X \in G_{\kappa_{k+1}} \cap \ker dh$, then X is of the form $X = \sum_{i=1}^n \alpha_i ad_f^{\kappa_{k+1} - 1}g_i + Z, Z \in G_{\kappa_{k+1} - 1}$, and with the functions α_i satisfying

$$\begin{aligned} 0 &= \langle d\tilde{h}_{k+1}, \textstyle\sum_{i=1}^n \alpha_i ad_f^{\kappa_{k+1} - 1}g_i\rangle(x_0) \\ &= \textstyle\sum_{i=1}^k \alpha_i(x_0)c_i, \quad \text{with } c_i := \langle d\tilde{h}_{k+1}, ad_f^{\kappa_{k+1} - 1}g_i\rangle(x_0) \end{aligned} \tag{2.37}$$

where at least one of the c_i's is unequal to zero because of (2.35). Now

$$G_{\kappa_{k+1}+1} = S_{\kappa_{k+1}+1} = [f, G_{\kappa_{k+1}} \cap \ker dh] + G_{\kappa_{k+1}}$$

and inductively,

$$G_{\kappa_k} \subset ad_f^{\kappa_k - \kappa_{k+1}}(G_{\kappa_{k+1}} \cap \ker dh) + G_{\kappa_k - 1}. \tag{2.38}$$

Therefore any element of G_{κ_k} is of the form

$$\sum_{i=1}^m \alpha_i ad_f^{\kappa_k - 1}g_i + Z, \qquad Z \in G_{\kappa_k - 1}, \tag{2.39}$$

with $\alpha_i(x_0), i = 1, \ldots k$, satisfying (2.37). Hence, because of table (2.33) and the nontrivial relation (2.37), the space $(F_* G_{\kappa_k})(F(x_0))$ is at most $(k-1)$ dimensional which is in

contradiction with (2.34). Therefore there *does* exist some $i \in \{k+1, \dots, m\}$ such that (2.35) holds. After reordering, if necessary, $g_{k+1}, \dots, g_m$ we may thus assume that

$$\langle d\tilde{h}_{k+1}, ad_f^{\kappa_{k+1}-1} g_{k+1} \rangle (x_0) \neq 0. \tag{2.40}$$

Now define $\beta_i^{k+1} := \langle d\tilde{h}_{k+1}, ad_f^{\kappa_{k+1}-1} g_i \rangle$ and set $\tilde{g}_i := g_i - (\beta_i^{k+1}/\beta_{k+1}^{k+1}) g_{k+1}, i = k+2, \dots, m$. Then as in Step 1, cf. (2.30), we obtain

$$\langle d\tilde{h}_{k+1}, ad_f^{\kappa_{k+1}-1} \tilde{g}_i \rangle \equiv 0, \quad \text{around } x_0, \quad i = k+2, \dots, m. \tag{2.41}$$

Omitting again the tildes above $\tilde{g}_i$ we have thus proved that (A1), (A2) also hold for $j = k+1$. Hence by induction we have proved that (A1), (A2) hold for every $k = 1, \dots, m$, for the feedback transformed system (the feedback arising from successively applying Step 1 up to Step m). It immediately follows from expression (2.25) that the decoupling matrix $\tilde{D}(x_0)$ for this feedback transformed is a lower triangular matrix with nonzero diagonal elements $\beta_1^1(x_0), \dots, \beta_m^m(x_0)$, and thus is nonsingular. Since the rank of the decoupling matrix is invariant under feedback [Is], [NvdS] we have proved the claim. □

As we have already remarked (see Remark 2 after Theorem 4), Theorem 4 and its proof are closely related to the local feedback linearization problem [JR], [HSM]. However, we would like to stress that from a computational point of view the transformation of Σ into a prime system (P) as given by Theorem 4 may be much simpler than the solution to the local feedback linearization problem. In fact for the latter problem we have to find, in some way or another (see [JR], [HSM]), rectifying (Frobenius) coordinates for the whole sequence of distributions $G_1 \subset G_2 \subset \dots \subset G_n = TM$, on the (possibly high-dimensional) state space manifold M. On the other hand, in order to transform Σ into (P) we basically have to find rectifying (Forbenius) coordinates for the projected distributions $H_1 \subset H_2 \subset \dots \subset H_n = TY$ on the output space manifold Y. In general the dimension of Y is much smaller than that of M, and therefore, *assuming* that the projections $H_1, \dots, H_n$ are easily computed, the latter problem is likely to be simpler. We defer a more elaborate computational implementation of Theorem 4 until after the proof of the next theorem, which deals with the more general problem of local equivalence to input-output prime systems. Recall that given two distributions D_1, D_2 on M we call D_1 involutive modulo D_2 if for any two $X, Y \in D_1$ we have $[X, Y] \in D_1 + D_2$. Furthermore, observe that if the codistributions P_i, $i = 1, 2, \dots$ in (2.4) are constant dimensional then V^*, the largest locally controlled distribution contained in kerdh, exists and is constant dimensional (and is given as $V^* = \ker P^*$).

THEOREM 6. *Consider a nonlinear system Σ with equilibrium x_0. Σ is locally equivalent to an input-output prime system $(I - O - P)$ with equilibrium $(0, z_0)$, if and only if the following conditions are satisfied in a neighborhood of x_0*

(i) *P_i is constant dimensional $i = 1, 2, \dots, n$;*
(ii) *G_i is involutive modulo V^*, and $G_i + V^*$ is constant dimensional; $i = 1, \dots, n-1$;*
(iii) *$G_n + V^* = TM$;*
(iv) *$G_i = S_i$ modulo $V^*, i = 1, 2, \dots, n$;*
(v) *$G_i +$ kerdh is involutive and of constant dimension, $i = 1, 2, \dots, n-1$.*

Proof (only if). Suppose that Σ is locally equivalent to $(I-O-P)$. Clearly, $(I-O-P)$ satisfies conditions (i)–(v) (notice that $V^* = \text{span}\{\partial/\partial z\}$). By (2.2) the definition of S_i is invariant under feedback. From the fact that $(I - O - P)$ satisfies condition (ii), it follows that also the definition of G_i is invariant under feedback modulo V^* (i.e., G_i for the feedback transformed system is equal modulo V^* to G_i for the original system). In

particular, since $V^* \subset \ker dh$, it follows that the definition of $G_i + \ker dh$ is invariant under feedback. Thus conditions (i)–(v) are feedback invariant and we can conclude that they are satisfied by Σ.

(If.) By definition of V^* (see, e.g., [IKGM], [Hi]) there exists locally around x_0 a feedback $u = \alpha(x) + \beta(x)v, \det\beta(x) \neq 0$, such that

$$[\tilde{f}, V^*] \subset V^*, \quad [\tilde{g}_j, V^*] \subset V^*, \quad j = 1, \dots, m, \tag{2.42}$$

where $\dot{x} = \tilde{f}(x) + \sum_{j=1}^{m} \tilde{g}_j(x)v_j$ denotes the feedback transformed system. Thus locally around x_0 we can factor out by the distribution V^* to obtain a manifold M' and a factor system (see [IKGM])

$$(\Sigma') \qquad \begin{aligned} \dot{x}' &= f'(x') + \sum_{j=1}^{m} g'_j(x')v_j, \qquad x' \in M', \\ y_j &= h'_j(x'), \qquad j = 1, \dots, m, \end{aligned} \tag{2.43}$$

i.e., around x_0 we have the projection $\pi : M \to M'$, with $V^* = \ker \pi_*$. (Since $V^* \subset \ker dh, h = (h_1, \dots, h_m)$ can be also factored to a map $h' = (h'_1, \dots, h'_m) : M' \to Y$ satisfying $h = h' o \pi$.)

Define the distributions G'_i and S'_i for the factor system Σ'. It is readily checked that G'_i and S'_i satisfy conditions (i)–(iv) of Theorem 4 for the factor system Σ' and around $x'_0 = \pi(x_0)$. Indeed, observe again that under conditions (ii) and (iv) of Theorem 6 the distributions G_i and S_i are feedback invariant modulo V^*. Then it immediately follows that G'_i and S'_i satisfy conditions (i)–(iii) of Theorem 4 applied to Σ'. Finally, since $V^* \subset \ker dh$ it follows that G_i are also feedback invariant modulo $\ker dh$. Thus $\pi_*(G_i + \ker dh) = G'_i + \ker dh'$, and it follows from Lemma 2 (applied to the involutive and constant-dimensional distribution $G_i + \ker dh$ and the mapping $\pi : M \to M'$) that $G'_i + \ker dh'$ is involutive and constant dimensional.

Hence by Theorem 4, Σ' is locally equivalent to a prime system (P) of the form (1.3a) (with $x'_0 = 0$, and $\mu_i, i = 1, \dots, m$, the controllability indices of Σ'). Since the remaining dynamics of Σ are of the general form (1.3b) we conclude that Σ is locally equivalent to $(I - O - P)$ with equilibrium $x_0 = (0, z_0)$. □

Remark 1. Note that the indices $\mu_1, \dots, \mu_m$ are *intrinsically* defined. Indeed if Σ satisfies the conditions of Theorem 6 then $\mu_1, \dots, \mu_m$ are the (intrinsically defined) controllability indices of the factor system Σ', living on M/V^*. In particular it follows that an input-output prime system $(I - O - P)$ *cannot* be equivalent to an input-output prime system with different indices $\mu_1, \dots, \mu_m$.

Remark 2. If Σ satisfies the conditions of Theorem 6 on a neighborhood of a point $\overline{x}$ which is *not* an equilibrium, then Σ will be locally equivalent to an input-output prime system (1.4) with the addition of a constant drift term $f(\overline{x})$. Furthermore, if $f(\overline{x}) \in G(\overline{x})$ then this drift term can be removed by additional feedback. Similarly, if $h(\overline{x}) \neq 0$, then we have to add to the output equation of (1.4a) the constant term $h(\overline{x})$. Of course, this remark already applies to Theorem 4.

Remark 3. It follows from the proof of Theorem 6 that $h_* G_i$ is a well-defined distribution on a neighborhood of $y_0 = h(x_0)$ (i.e., G_i is projectable by h on some neighborhood V_{x_0}), $i = 1, \dots, m$. In fact $h_* G_i = h'_* G'_i$ (with $'$ denoting the factor system Σ'), and the projectability of G'_i by h' to an involutive constant-dimensional distribution on a neighborhood of y_0 follows by an application of Lemma 5 to G'_i and h'. Note, however, that

Lemma 5 as it stands *cannot* be directly applied to G_i and h (satisfying condition (v)), since we do not require G_i to be involutive and constant dimensional (but only modulo V^*).

Note that Theorem 6 generalizes the well-known fact that a nonlinear system Σ whose decoupling matrix $D(x)$ (cf. (2.6)) has rank m around x_0 can be transformed by local state space and feedback transformations into (1.3), see, e.g., [IKGM]. Hence Theorem 6 can also be interpreted as giving the necessary and sufficient conditions for finding a local output transformation $\tilde{y} = \psi(y) = (\psi_1(y), \ldots, \psi_m(y))$ such that the decoupling matrix $\tilde{D}(x)$ for the transformed output functions $\tilde{h}_1 = \psi_1 oh, \ldots, \tilde{h}_m = \psi_m oh$ has rank m around x_0.

Example. Consider the following system on $M = \mathbb{R}^3, Y = \mathbb{R}^2$:

$$\begin{aligned} \dot{x}_1 &= u_2, & y_1 &= x_1 \\ \dot{x}_2 &= x_3, & y_2 &= x_2 + \frac{1}{3}x_1^3 \\ \dot{x}_3 &= u_1. \end{aligned} \tag{2.44}$$

The relative degrees are both 1, while the decoupling matrix $D(x)$ equals

$$D(x) = \begin{pmatrix} 0 & 1 \\ 0 & x_1^2 \end{pmatrix}$$

and thus is singular, implying that the system is *not* input-output decouplable by static state feedback. However it is readily seen that the system satisfies the conditions of Theorem 6 and even of Theorem 4, and in fact we only need the output transformation

$$\psi_1(y) = y_2 - \tfrac{1}{3}y_1^3, \qquad \psi_2(y) = y_1$$

to bring the system into prime form (1.2), with $\kappa_1 = 1, \kappa_2 = 2$ (being the relative degrees of the *transformed* system)! Now suppose we want to asymptotically track a desired smooth trajectory $y^d(t) = (y_1^d(t), y_2^d(t)), t \geq 0$ for (2.44). Using the above output transformation, such a trajectory is transformed into the new coordinates as $\tilde{y}^d(t) = (y_2^d(t) - \frac{1}{3}(y_1^d(t))^3, y_1^d(t)), t \geq 0$, and since (2.44) has been transformed into a prime system the tracking problem is simply solved by a control strategy which is *linear* in the transformed coordinates, namely,

$$\begin{aligned} u_1 = &- K_{11}(x_2 - (y_2^d(t) - \tfrac{1}{3}(y_1^d(t))^3) \\ &- K_{12}(x_3 - (\dot{y}_2^d(t) - (y_1^d(t))^2 \dot{y}_1^d(t))) \\ &+ \ddot{y}_2^d(t) - 2y_1^d(t)(\dot{y}_1^d(t))^2 - (y_1^d(t))^2 \ddot{y}_1^d(t) \\ u_2 = &- K_2(x_1 - y_1^d(t)) + \dot{y}_1^d(t) \end{aligned}$$

where $K_2 < 0$, and K_{11}, K_{12} are designed in such a way that the polynomial $s^2 + K_{12}s + K_{11}$ is Hurwitz.

Notice, furthermore, that the conditions of Theorem 6 imply (see, e.g., [NvdS], [Is]) that Σ is input-output decouplable by *dynamic* state feedback. (In the foregoing example, system (2.44) can be dynamically input-output decoupled for the original output functions by pre-integrating the input u_2 one time). Regarded from this viewpoint, Theorem 6 *avoids* the addition of extra pre-integrators to the system by allowing instead for output transformations.

The proofs of Theorems 4 and 6 immediately yield the following algorithm to transform Σ into a prime or input-output prime system.

ALGORITHM 7. Consider a nonlinear system Σ with equilibrium x_0, and satisfying, in a neighborhood V_{x_0} of x_0, conditions (i)–(v) of Theorem 6. Then Σ can be transformed into $(I-O-P)$ in the following way.

(a) Compute the distributions $H_i := h_* G_i$ on $W_{y_0} = h(V_{x_0}), i = 1, 2, \ldots, n-1$. (By Remark 3 above, H_i are all well defined, involutive, and constant-dimensional distributions.)

(b) Construct rectifying (Frobenius) coordinates $\psi_1, \ldots, \psi_m$ (cf. (2.18), (2.19)) defined on a possibly smaller neighborhood of y_0, for the whole sequence $H_1 \subset H_2 \subset \ldots \subset H_{n-1}$. This defines the output space transformation ψ of Definition 1.

(c) Consider the output functions $\tilde{h}_i := \psi_i o h, i = 1, \ldots, m$, for Σ. Compute the relative degrees $\mu_1, \ldots, \mu_m$ for these output functions and the decoupling matrix $\tilde{D}(x) = (L_{g_j} L_f^{\mu_i - 1} \tilde{h}_i(x))_{i,j=1,\ldots,m}$. Necessarily $\mu_i < \infty, i = 1, \ldots, m$, and rank $D(x) = m$ around x_0. Define the functions

$$x_{ij} := L_f^{(j-1)} \tilde{h}_i, \quad j = 1, \ldots, \mu_i, \quad i = 1, \ldots, m. \tag{2.45}$$

Necessarily these functions are independent around x_0, while

$$V^* = \ker \operatorname{span}\{dx_{ij}, j = 1, \ldots, \mu_i, \quad i = 1, \ldots, m\}. \tag{2.46}$$

Choose complementary coordinates $z = (z_1, \ldots, z_{n'})$ around $x_0 (n' := n - (\mu_1 + \ldots + \mu_m))$. This defines the state space transformation φ of Definition 1.

(d) Compute the regular feedback $u = \alpha(x) + \beta(x)v$ around x_0 as

$$\alpha(x) = -\tilde{D}^{-1}(x) \begin{pmatrix} L_f^{\mu_1} \tilde{h}_1(x) \\ \vdots \\ L_f^{\mu_m} \tilde{h}_m(x) \end{pmatrix}, \qquad \beta(x) = \tilde{D}^{-1}(x). \tag{2.47}$$

This defines the feedback transformation required in Definition 1.

Remark 1. Note that $\mu_1 + \ldots + \mu_m = n$ if and only if $V^* = 0$, in which case Σ is locally equivalent to a prime system.

Remark 2. At some occasions it may be more efficient *not* to check conditions (i)–(v) of Theorem 6 in order to see if Σ is locally equivalent to $(I-O-P)$, but instead to apply *directly* Algorithm 7. If the Algorithm breaks down (e.g., if some distributions H_i are not well defined or not involutive, or if $\tilde{D}(x)$ does not have full rank) then Σ is not locally equivalent to $(I-O-P)$ (while Σ is locally equivalent to $I-O-P$ if Algorithm 7 *does* work).

Example. As an illustration of the above remark we apply Algorithm 7 to the example

$$\begin{aligned} \dot{x}_1 &= u_1, & y_1 &= x_5 + x_5 x_3 - \tfrac{1}{2} x_5 x_2^2 \\ \dot{x}_2 &= u_2, & y_2 &= x_3 - \tfrac{1}{2} x_2^2 \\ \dot{x}_3 &= x_1 + x_2 u_2 \\ \dot{x}_4 &= x_2 + x_1 u_1 \\ \dot{x}_5 &= x_4 - \tfrac{1}{2} x_1^2. \end{aligned}$$

Following the proof of Lemma 5 we first express the system in local coordinates

$$\xi_1 = x_1, \quad \xi_2 = x_2, \quad \xi_3 = x_3 - \tfrac{1}{2} x_2^2, \quad \xi_4 = x_4, \quad \xi_5 = x_5 (1 + x_3 - \tfrac{1}{2} x_2^2)$$

as

$$\begin{aligned}
\dot{\xi}_1 &= u_1, \qquad y_1 = \xi_5 \\
\dot{\xi}_2 &= u_2, \qquad y_2 = \xi_3 \\
\dot{\xi}_3 &= \xi_1 \\
\dot{\xi}_4 &= \xi_2 + \xi_1 u_1 \\
\dot{\xi}_5 &= \frac{\xi_5}{1+\xi_3}(1+\xi_1) + (\xi_4 - \tfrac{1}{2}\xi_1^2)(1+\xi_3).
\end{aligned}$$

(a) Simple computations give

$$\begin{aligned}
G_1 &= \operatorname{span}\left\{\frac{\partial}{\partial \xi_1}, +\xi_1 \frac{\partial}{\partial \xi_4}, \frac{\partial}{\partial \xi_2}\right\} \\
G_2 &= \operatorname{span}\left\{\frac{\partial}{\partial \xi_1}, \frac{\partial}{\partial \xi_2}, \frac{\partial}{\partial \xi_4}, \frac{\partial}{\partial \xi_3} + \frac{\xi_5}{1+\xi_3}\frac{\partial}{\partial \xi_5}\right\} \\
G_3 &= TM \\
H_1 &= 0 \\
H_2 &= \operatorname{span}\left\{\frac{y_1}{1+y_2}\frac{\partial}{\partial y_1} + \frac{\partial}{\partial y_2}\right\} \\
H_3 &= TY.
\end{aligned}$$

According to (2.15) the indices are $\kappa_2 = 2, \kappa_1 = 3$.

(b) From (2.18) the rectifying Frobenius coordinates are

$$\psi_1(y) = \frac{y_1}{1+y_2} \qquad \psi_2(y) = y_2.$$

(c) The transformed output functions $\tilde{h}_1, \tilde{h}_2$ are

$$\tilde{h}_1 = \frac{x_5(1 + x_3 - \frac{1}{2}x_2^2)}{1 + x_3 - \frac{1}{2}x_2^2} = x_5 \qquad \tilde{h}_2 = x_3 - \tfrac{1}{2}x_2^2.$$

The relative degrees are $\mu_1 = 3, \mu_2 = 2$, while the decoupling matrix is $\tilde{D} = I_{2\times 2}$. Hence the functions

$$\begin{aligned}
z_1 &= \tilde{h}_1 = x_5 \\
z_2 &= L_f \tilde{h}_1 = x_4 - \tfrac{1}{2}x_1^2 \\
z_3 &= L_f^2 \tilde{h}_1 = x_2 \\
z_4 &= \tilde{h}_2 = x_3 - \tfrac{1}{2}x_2^2 \\
z_5 &= L_f \tilde{h}_2 = x_1
\end{aligned}$$

give the state space transformation φ of Definition 1.

(d) The regular feedback of definition 1 is

$$\alpha(x) = 0 \qquad \beta(x) = I.$$

In fact, in z-coordinates we have

$$\dot{z}_1 = z_2, \quad \dot{z}_2 = z_3, \quad \dot{z}_3 = u_2 \quad \dot{z}_4 = z_5 \quad \dot{z}_5 = u_1,$$

which is a linear prime system.

The extension of Theorem 6 to local equivalence into input-output prime systems of special form (1.3) reads as follows.

PROPOSITION 8. *Consider a nonlinear system* Σ *with equilibrium* x_0. Σ *is locally equivalent to an input-output prime system of special form* $(I-O-P-S)$, *if and only if, on a neighborhood of* x_0, *conditions (i)–(v) of Theorem 6 are satisfied and, additionally:*

(vi) $V^* \cap S^* = 0$;

(vii) S^* is involutive and constant dimensional.

Proof (only if). Suppose Σ is locally equivalent to $(I-O-P-S)$. From the (only if) part of Theorem 6 it follows that Σ satisfies conditions (i)–(v). Clearly, $(I-O-P-S)$ satisfies conditions (vi) and (vii). Furthermore the definition of S^* is feedback invariant, and thus also Σ satisfies conditions (vi) and (vii).

(If.) By (iii), (iv), and (vi) we have $V^* \oplus S^* = TM$. By Theorem 6, Σ is locally equivalent to $(I-O-P)$, i.e., (1.3). Here z are coordinate functions which are arbitrary except for the fact that they have to be complementary to the coordinate functions $x = (x_{11}, \ldots, x_{m\mu_m})$; see (2.45). In the present case, however, since $V^* \oplus S^* = TM$ and V^* and S^* are involutive and constant dimensional, we can choose z such that span dz = ann S^*. Since, by construction, span dx = ann V^*, cf. (2.46), we thus have

$$V^* = \operatorname{span}\left\{\frac{\partial}{\partial z}\right\}, \qquad S^* = \operatorname{span}\left\{\frac{\partial}{\partial x}\right\}. \tag{2.48}$$

Then, first of all, since $G = S_1 \subset S^* = \operatorname{span}\{\partial/\partial x\}$, we have in (1.3)

$$b_j(z,x) = 0, \qquad j = 1, \ldots, m. \tag{2.49}$$

Second, by definition of S^*, $[f, S^* \cap \ker dh] \subset S^*$, cf. (2.3), and thus, since f is of the form $f = *\partial/\partial x + a(z,x)\partial/\partial z$ and $S^* = \operatorname{span}\{\partial/\partial x\}$

$$\left[*\frac{\partial}{\partial x} + a(z,x)\frac{\partial}{\partial z}, \frac{\partial}{\partial x_{ij}}\right] \subset \operatorname{span}\left\{\frac{\partial}{\partial x}\right\}, \quad j = 2, \ldots, \mu_i, \quad i = 1, \ldots, m. \tag{2.50}$$

(Note that kerdh is everything minus span $\{(\partial/\partial x_{i1}), i = 1, \ldots, m\}$.) This implies that $a(z,x)$ in (1.3) may only depend on z and $x_{i1} = y_i, i = 1, \ldots, m$, and thus (1.4) results.

□

Remark 1. If conditions (ii) and (iv) in Proposition 8 are replaced by the stronger conditions

(ii)′ G_i is involutive and of constant dimension, $i = 1, 2, \ldots, n-1$,

(iv)′ $G_i = S_i, i = 1, \ldots, n$,

then, following [MBE], $a(z,y)$ in (1.4b) will only depend on those (new) output components y_i with i such that $\mu_i = \max\{\mu_1, \ldots \mu_m\}$.

Remark 2. Necessary and sufficient conditions for transforming into (1.4) (without change of output space coordinates) a nonlinear system Σ having invertible decoupling matrix have been identified in [BI]: see also [MBE]. Similar conditions were derived, in a different context, in [vdS]. Notice that in the *linear* case condition (vii) is automatically satisfied. This explains that for a linear system we can always write (even if condition (vi) is not satisfied) the V^* dynamics as being only driven by y, as follows from the Morse canonical form [Mo].

EXAMPLE. Consider the single input system

$$\dot{x}_1 = x_4^3, \qquad y = x_2$$

$$\dot{x}_2 = x_3,$$
$$\dot{x}_3 = x_4$$
$$\dot{x}_4 = u.$$

Easy computations give:

$$V^* = \text{span}\left\{\frac{\partial}{\partial x_1}\right\}, \quad \ker dh = \text{span}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_4}\right\}, \quad S^* = \mathbb{R}^4,$$
$$S^* \cap V^* = \text{span}\left\{\frac{\partial}{\partial x_1}\right\},$$
$$G_1 = \text{span}\left\{\frac{\partial}{\partial x_4}\right\}, \qquad G_2 = \text{span}\left\{\frac{\partial}{\partial x_4}, 3x_4^2\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_3}\right\},$$
$$G_3 = \text{span}\left\{\frac{\partial}{\partial x_4}, 3x_4^2\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_2}\right\}.$$

Theorem 4 does not apply since G_2 is not involutive. Theorem 6 applies, while Proposition 8 does not apply since $V^* \cap S^* \neq 0$.

Now let us proceed to a *global* version of the equivalence results we have obtained so far. Instead of requiring that rank dh(x) and dim $G(x)$ are equal to m in a neighborhood of x_0, we will now have to require this on the whole M. Then by the Rank Theorem (e.g., [Sp]), $h(M)$ is an open part of Y, and without loss of generality, we may restrict to this part of Y and *assume* that $h : M \to Y$ is surjective. The global version of Definition 1 reads now as follows.

DEFINITION 9. Consider two systems Σ_1, Σ_2 defined on $(M_1, Y_1), (M_2, Y_2)$ with equilibrium points $x_{01} \in M_1, x_{02} \in M_2$, respectively. Σ_1 is globally equivalent to Σ_2 if there exist:

(i) A diffeomorphism $\varphi : M_1 \to M_2$, satisfying $\varphi(x_{01}) = x_{02}$;
(ii) a nonsingular feedback $u = \alpha(x) + \beta(x)v$ on M_1 with $\alpha(x_{01}) = 0$ and det $\beta(x) \neq 0$;
(iii) a diffeomorphism $\psi : Y_1 \to Y_2$ with $\psi(h_1(x_{01})) = h_2(x_{02})$ such that the resulting transformation of Σ_1 equals Σ_2.

Since Theorem 6 generalizes Theorem 4 we will only give the global version of Theorem 6, and state as a corollary the global version of Proposition 8.

THEOREM 10. *Consider a nonlinear system* Σ *on* (M, Y) *with equilibrium* x_0, *and assume that* $h : M \to Y$ *is a surjective submersion and that* dim $G(x) = m$, *for all* $x \in M$. *Suppose that conditions* (i)–(v) *of Theorem* 6 *are satisfied on the whole* M, *and that*

(A) There exist globally *defined independent functions* $\psi_1, \ldots, \psi_m$ *on* Y *which are rectifying coordinates for* $H_1, \ldots, H_m$, *i.e.,* (2.19a) *and* (2.19b) *are satisfied for every* $x \in M$ *(the local existence of* $\psi_1, \ldots, \psi_m$ *is already insured by conditions (i)–(v));*

then by Algorithm 7*(c), (d)* V^* *is globally given by* (2.46) *and the feedback* (2.47) *is globally defined. Furthermore, there exists a surjective submersion* $\pi : M \to M'$ *with ker* $\pi_* = V^*$, *while the factor system* Σ', *cf.* (2.43), *is globally defined on* M'.

Assume additionally that

(B) The vectorfields f' *and* $g'_j, j = 1, \ldots, m$, *on* M', *cf. (2.43), are* complete;

then M' *equals* $\mathbb{R}^\mu, \mu = \sum_{i=1}^m \mu_i$, *and thus* Σ *is* globally equivalent *to an input-output system (I − O − P) with equilibrium* $(0, z_0)$. *Conversely, if* Σ *is globally equivalent to*

$(I - O - P)$ then conditions (i)–(v) *of Theorem 6 are satisfied on the whole M, and conditions (A) and (B) hold.*

Remark. Since the feedback (α, β) depends on the choice of $\psi_1, \ldots, \psi_m$, also condition (B), i.e., the completeness of the modified vectorfields $f', g'_j, j = 1, \ldots, m$, may depend on the choice of $\psi_1, \ldots, \psi_m$. This is already illustrated by the following very simple example: Consider the system $\dot{x} = u, y = e^{-x}$ on $M = \mathbb{R}$ and $Y = (0, \infty)$. If we take $\psi(y) = \ell n\, y$ as a global coordinate on Y (which trivially is rectifying, since $H_1 = TY$) then $\tilde{h}(x) = \psi o h(x) = -x$, and $g' = \tilde{g} = -\partial/\partial x$ is complete, implying that the system is globally equivalent to the prime system $\dot{x} = u, y = x$. However if we would take the global rectifying coordinate $\psi(y) = y$, then $g' = e^x \partial/\partial x$ is not complete, and indeed, since $id : (0, \infty) \to \mathbb{R}$ is *not* a diffeomorphism onto $\mathbb{R}$, the system is *not* globally transformed into a prime system.

Proof. Suppose conditions (i)–(v) are satisfied on M, as well as condition (A). Apply Algorithm 7 using the global rectifying coordinates $\psi_1, \ldots, \psi_m$ on Y. Since V^* is constant dimensional on the whole M it follows by a slight adaptation of [HK, Thm. 3.9], see also [IKGM], that V^* can be globally factored out, i.e., there exists a surjective submersion $\pi : M \to M'$ with $\ker \pi_* = V^*$, and the feedback transformed dynamics $\tilde{f}, \tilde{g}_j, j = 1, \ldots, m$ (with (α, β) defined by (2.47)) project to dynamics $f', g'_j, j = 1, \ldots, m$ on M' (note that, in contrast to [HK, Thm. 3.9], we do not require Σ to be accessible; however, condition (iii) of Theorem 6 insures that Σ is "accessible modulo V^*"). Now assume that condition (B) is satisfied. By the local equivalence of Σ with (1.3) it follows that the vectorfields $g'_j, ad_f g'_j, \ldots, ad_f^{\mu_j - 1} g'_j, j = 1, \ldots, m$, are *commuting* and *complete* vectorfields on M' (see [Re], [DBE]). It follows that $M' = \mathbb{R}^{\mu - k} \times S^k$ for some $k \geq 0$. However, since the functions $L_f^j \tilde{h}_i, j = 0, 1, \ldots, \mu_i - 1, i = 1, \ldots, m$, are *global* coordinate functions on M' necessarily $k = 0$ (since S^k is compact). Since $y_j = x_{j1}, j = 1, \ldots, m$, we also have $Y \simeq \mathbb{R}^m$. It follows [Re], [DBE] that Σ' is globally equivalent to a linear system, and thus that Σ is globally equivalent to $I - O - P$. Conversely, if Σ is globally equivalent to $(I - O - P)$ then by the (only if) part of Theorem 6 conditions (i)–(v) are satisfied everywhere. Furthermore, clearly $(I - O - P)$ satisfies Conditions A and B. □

COROLLARY 11. *Σ is globally equivalent to $(I - O - P - S)$ if and only if, in addition to conditions (i)–(v) and conditions A and B of Theorem 10, conditions (vi), (vii) of Proposition 8 are satisfied on the whole M.*

Remark. Analogous reasoning on the global equivalence modulo V^* to a linear system was used in [MRS]. Similar conditions for the global equivalence of a nonlinear system with invertible decoupling matrix into (1.3) or (1.4) were derived in [BI].

3. Conclusions and final remarks. Necessary and sufficient geometric conditions have been given for transforming nonlinear systems into (input-output) prime form (of special form), locally as well as globally. The main novelty (e.g., as compared to normal forms for input-output decouplable systems) is that we allow for output transformations. Actually, as made explicit in Algorithm 7 (see also the example following it), the output transformation is the crucial step in the whole transformation procedure and involves the simultaneous integration of a nested sequence of distributions on the output space manifold (similar to the integration of distributions on the *state* space manifold as in the feedback linearization problem). The results obtained are applicable to control problems where output transformations are naturally allowed, such as output tracking, output regulation, (almost) disturbance decoupling [I], [NvdS], [MRS] and the servomechanism problem. The results enable us to treat the class of nonlinear systems equivalent to input-output prime form very much like the well-studied class of input-output decouplable systems. Finally, as we

have remarked, the use of output transformations may be an alternative to the use of extra pre-integrators for *dynamic* intput-output decoupling. This raises the problem of how output transformations may be used to *minimize* the amount of pre-integrators for dynamic input-output decoupling.

Acknowledgment. Witold Respondek is grateful for the warm hospitality and financial support provided by the Dipartimento di Ingegneria Elettronica, Università di Roma, "Tor Vergata."

REFERENCES

[Br] P. Brunovsky, *A classification of linear controllable systems,* Kybernetika 6 (1970), pp. 173–188.

[BI] C. I. Byrnes and A. Isidori, *Asymptotic stabilization of minimum phase nonlinear systems,* IEEE Trans. Automat. Control, AC-36 (1991), pp. 1122–1137.

[BM] G. Basile and G. Marro, *Controlled and conditioned invariant subspaces in linear system theory,* J. Optim. Theory Appl. 3 (1969), pp. 306–315.

[CIRT] D. Cheng, A. Isidori, W. Respondek, and T. J. Tarn, *Exact linearization of nonlinear systems with outputs,* Math. Systems Theory, 21 (1988), pp. 63–83.

[DBE] W. Dayawansa, W. M. Boothby, and D. L. Elliott, *Global state and feedback equivalence of nonlinear systems,* Systems Control Lett., 6 (1984), pp. 229–234.

[Fr] E. Freund, *The structure of decoupled nonlinear systems,* Internat. J. Control, 21 (1975), pp. 443–450.

[FW] P. L. Falb and W. A. Wolovich, *Decoupling in the design and synthesis of multivariable control systems,* IEEE Trans. Automat. Control, AC-12 (1967), pp. 651–659.

[He] M. Heymann, *The prime structure of linear dynamical systems,* SIAM J. Control Optim. 10 (1972), pp. 460–469.

[Hi] R. M. Hirschorn, *(A, B)-invariant distributions and disturbance decoupling of nonlinear systems,* SIAM J. Control Optim. 19(1981), pp. 1–19.

[HK] R. Hermann and A. J. Krener, *Nonlinear controllability and observability,* IEEE Trans. Automat. Control, AC-22 (1977), pp. 728–740.

[HSM] L. R. Hunt, R. Su and G. Meyer, *Design for multi-input nonlinear systems,* in Différential Geometric Control Theory, R. W. Brockett, R. S. Millman, H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 268–298.

[Is] A. Isidori, *Nonlinear Control Systems,* 2nd ed., Springer-Verlag, New York, 1989.

[Is2] ———, *Nonlinear feedback, structure at infinity and the input-output linearization problem,* in Proc. of MTNS'83, Beer Sheva, P. A. Fuhrmann, ed. Lecture Notes in Control and Information Systems 48, Springer-Verlag, Berlin, New York, 1984, pp. 473–493.

[IKGM] A. Isidori, A. J. Krener, C. Gori-Giorgi and S. Monaco, *Nonlinear decoupling via feedback: A differential geometric approach,* IEEE Trans. Automat. Control, AC-26 (1981), pp. 331–345.

[J] B. Jakubczyk, *Feedback linearization of discrete time systems,* Systems Control Lett., 9(1987), pp. 411–416.

[JR] B. Jakubczyk and W. Respondek, *On linearization of nonlinear control systems,* Bull. Acad. Polon. Sci. Ser. Sci. Math., 28 (1980), pp. 517–522.

[KR] A. J. Krener and W. Respondek, *Nonlinear observers with linearizable error dynamics,* SIAM J. Control Optim., 23 (1985), pp. 197–216.

[M] C. H. Moog, *Nonlinear decoupling and structure at infinity,* Math. Control, Signals Systems, 1 (1988), pp. 257–268.

[Mo] A. S. Morse, *Structural invariants of linear multivariable systems,* SIAM J. Control Optim. 11 (1973), pp. 446–465.

[MBE] R. Marino, W. M. Boothby and D. L. Elliott, *Geometric properties of linearizable control systems,* Math. Syst. Theory, 18 (1985), pp. 97–123.

[MRS] R. Marino, W. Respondek and A. J. van der Schaft, *Almost disturbance decoupling for single-input single-output nonlinear systems,* IEEE Trans. Automat. Control, 34 (1989), pp. 1013–1017.

[NS] H. Nijmeijer and J. M. Schumacher, *Zeros at infinity for affine nonlinear control systems,* IEEE Trans. Automat. Control, 30 (1985), pp. 566–573.

[NvdS] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems,* Springer-Verlag, New York, 1990.

[Re] W. Respondek, *Global aspects of linearization, equivalence to polynomial forms and decomposition of nonlinear systems,* in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess, M. Hazewinkel, eds., D. Reidel, Dordrecht, 1986, pp. 257–284.

[Si] P. K. SINHA, *State feedback decoupling of nonlinear systems,* IEEE Trans. Automat. Control, AC-22 (1977), pp. 487–489.

[Sp] M. A. SPIVAK, *A Comprehensive Introduction to Differential Geometry,* Vol. I, Publish or Perish, Boston, 1970.

[SR] S. N. SINGH AND W. J. RUGH, *Decoupling in a class of nonlinear systems by state variable feedback,* J. Dynamic Systems Meas. Control, 94 (1972), pp. 323–329.

[vdS] A. J. VAN DER SCHAFT, *Observability and controllability for smooth nonlinear systems,* SIAM J. Control Optim., 20 (1982), pp. 338–354.

[Wo] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach,* 3rd ed. Springer-Verlag, New York, 1985.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 408–420, March 1994

005

STABILIZATION AND EXACT BOUNDARY CONTROLLABILITY FOR MAXWELL'S EQUATIONS*

B. V. KAPITONOV†

Abstract. This paper considers Maxwell's equations with dissipative boundary conditions. Under certain geometric conditions imposed on the domain Ω, the results on uniform stabilization of the solutions are established. Exact boundary controllability is then obtained through Russell's "controllability via stabilizability" principle.

Key words. Maxwell's equations, stabilization, exact controllability

AMS subject classifications. 35L, 49E

1. Introduction and problem formulation. Throughout this paper Ω is an open, bounded domain in R^3 with sufficiently smooth boundary $\partial\Omega = S$. In $\Omega \times (0,T)$ we consider the initial boundary value problem for the Maxwell system:

$$\begin{aligned} e_t &= \operatorname{curl}(\mu h), \\ h_t &= -\operatorname{curl}(\lambda e), \\ \operatorname{div} e &= \operatorname{div} h = 0, \end{aligned} \tag{1.1}$$

$$e(x,0) = f_1(x), \qquad h(x,0) = f_2(x), \tag{1.2}$$

$$[\nu, e] - \alpha(h - \nu(h,\nu)) = 0 \qquad (x,t) \in S \times (0,T), \tag{1.3}$$

where e and h are three-dimensional vector-valued functions of $t, x = (x_1, x_2, x_3)$, ν is the unit outer normal, $[\cdot,\cdot]$ and $(\cdot,\cdot)$ are the vector and inner products, $\mu = \mu(x)$ and $\lambda = \lambda(x)$ are scalar functions in Ω (the conditions on μ and λ are presented below), and $\alpha = \alpha(x)$ is a continuously differentiable function on S with $\operatorname{Re}\alpha > 0$.

The equality (1.3) (the Leontovich condition) means that surface S is a conductor and complex-valued function is a surface impedance (cf. [12] for details).

Our first purpose is to prove the uniform stabilization as $t \to \infty$ of solutions of problem (1.1)–(1.3).

Using this result we study the following exact controllability problem:

Given the initial distribution $\{f_1(x), f_2(x)\}$, time $T > 0$, and a desired terminal state $\{g_1(x), g_2(x)\}$ with $\{f_1(x), f_2(x)\}, \{g_1(x), g_2(x)\}$ in appropriate function spaces, find a vector-valued function $p(x,t)$ in a suitable function space such that the solution of (1.1)–(1.2) with boundary condition

$$[\nu, e] - ia(x)(h - \nu(h,\nu))/_S = p(x,t) \qquad (\mathcal{J}ma(x) = 0) \tag{1.4}$$

satisfies

$$e(x,T) = g_1(x), \qquad h(x,T) = g_2(x). \tag{1.5}$$

Here $a(x)$ is scalar continuously differentiable function on S.

* Received by the editors August 26, 1991; accepted for publication (in revised form) September 22, 1992.
† Institute of Mathematics, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia, (boka@math.nsk.su).

Remark 1.1. Due to finite speed of propagation (1.1)–(1.2) the exact controllability problem can have a solution only if T is large enough. The determination of T is part of the problem.

The exact controllability problem for Maxwell's equations ($\lambda(x) \equiv \lambda_0, \mu(x) \equiv \mu_0$ for $x \in \Omega$) with boundary control by means of currents flowing tangentially in the boundary of the region ($a(x) \equiv 0$) has been studied by Russell [17] for a circular cylindrical region, by Kime [8] for a spherical region, and by Lagnese [9] for a general region. In [17] it is assumed that the fields e and h do not depend on the axial coordinate. The control problem can then be transformed into a problem of the exact controllability of two wave equations by means of a single control. The latter control problem is then solved by the moment problem method. A moment problem approach is also used in [8]. In [9] the exact controllability problem for a general region has been studied by means of the Hilbert uniqueness method introduced by Lions [14], [15].

In this paper the controllability problem (1.1)–(1.2), (1.4)–(1.5) is solved using the energy decay of the solution of (1.1)–(1.3). This approach was applied to the wave equation by G. Chen [1]–[3], Lagnese [10], Lasiecka and Triggiani [13], and to the linear elastodynamic systems by Lagnese [11].

2. Well-posedness of (1.1)–(1.3). Let $\lambda(x)$ and $\mu(x)$ be continuously differentiable function in Ω satisfying the conditions $0 < \lambda_0 \le \lambda(x) \le \lambda_1, 0 < \mu_0 \le \mu(x) \le \mu_1$. We denote by $\mathcal{H}$ the Hilbert space of pairs $\{u_1, u_2\}$ of three-component complex-valued functions $u_i \in L_2(\Omega)$ with the inner product

$$\langle \{u_1, u_2\}, \{v_1, v_2\} \rangle_0 = \int_\Omega (\lambda(u_1, \bar{v}_1) + \mu(u_2, \bar{v}_2)) dx.$$

We denote by $\mathcal{H}_1$ the Hilbert space consisting of pairs $u = \{u_1, u_2\}$ such that $\{u_1, u_2\}$, $\{\text{curl } u_1, \text{curl } u_2\} \in \mathcal{H}$. We define the inner product in $\mathcal{H}_1$ by

$$\begin{aligned}\langle \{u_1, u_2\}, \{v_1, v_2\} \rangle_1 = \int_\Omega [&(\text{curl } \lambda u_1, \text{curl } \lambda \bar{v}_1) + (\text{curl } \mu u_2, \text{curl } \mu \bar{v}_2) \\ &+ (c + \lambda)(u_1, \bar{v}_1) + (c + \mu)(u_2, \bar{v}_2)] dx,\end{aligned}$$

where the constant c is chosen so that the norm in $\mathcal{H}_1$ is equivalent to the norm defined by the expression

$$\int_\Omega (|\text{curl } u_1|^2 + |\text{curl } u_2|^2 + |u_1|^2 + |u_2|^2) dx.$$

We further denote by $H_m(\Omega)$ and $H_q(S)$ the usual Sobolev spaces, and by $\| \cdot \|_{m,\Omega}$ and $\| \cdot \|_{q,S}$ the norms in them.

LEMMA 2.1. *Suppose* $\alpha(x) \in C^1(S)$. *The mapping* $\{u_1, u_2\} \to [\nu, u_1] - \alpha(u_2 - \nu(u_2, \nu))$ *from* $C^1(\bar{\Omega})$ *into* $C^1(S)$ *extends by continuity to a continuous linear mapping of* $\mathcal{H}_1 \to H_{-1/2}(S)$, *which we also denote by* $u \to [\nu, u_1] - \alpha(u_2 - \nu(u_2, \nu)) \equiv \omega(\alpha, u)$.

We can now introduce in $\mathcal{H}_1$ the closed subspace $\overset{\circ}{\mathcal{H}}_1(\alpha) = \{u \in \mathcal{H}_1 : \omega(\alpha, u) = 0\}$, which is dense in $\mathcal{H}$.

In $\mathcal{H}$ we define the unbounded operation A:

$$\mathcal{D}(A) = \overset{\circ}{\mathcal{H}}_1(\alpha), \quad Au = \{\text{curl } \mu u_2, -\text{curl } \lambda u_1\}, \quad \{u_1, u_2\} \in \mathcal{D}(A).$$

LEMMA 2.2. *The domain of the adjoint operator* A^* *coincides with* $\overset{\circ}{\mathcal{H}}_1(-\bar{\alpha})$. *For* $v = \{v_1, v_2\} \in \mathcal{D}(A^*)$.

$$A^* v = -\{\text{curl } \mu v_2, -\text{curl } \lambda v_1\}.$$

The proofs of Lemmas 2.1 and 2.2 are carried out in a manner similar to that in [7] for an unbounded domain (see also [6] for $\alpha(x) \equiv 0$). For this reason we do not present these arguments.

We note that the operator A is closed, since it coincides with the operator adjoint to A^*.

It can be shown that for $\mathrm{Re}\alpha(x) \geq 0$ the domain of A contains the set of smooth functions in $C^1(\Omega)$ satisfying $\omega(\alpha, u) = 0$ on the boundary as a dense set. Using this circumstance, it is easy to prove that the operators A and A^* are dissipative, i.e.,

$$\mathrm{Re}\langle Au, u\rangle_0 \leq 0, \quad u \in \mathcal{D}(A); \qquad \mathrm{Re}\langle A^*v, v\rangle_0 \leq 0, \quad v \in \mathcal{D}(A^*).$$

The operator A thus generates a strongly continuous semigroup of contractions $U(t), t > 0$.

Let $M = \{v \in \mathcal{D}(A^*) : A^*v = 0\}$, and let M_1 be the orthogonal complement of M in $\mathcal{H}$. The kernel of A^* is nonempty, since it contains the pairs $\{\lambda^{-1}\nabla\varphi_1, \mu^{-1}\nabla\varphi_2\}$ where $\varphi_i \in \overset{\circ}{H}_1(\Omega) \cap H_2(\Omega)$. It is obvious that $U(t)$ takes $M_1 \cap \mathcal{D}(A)$ into itself. Indeed, if $v \in M$ and $u \in M_1 \cap \mathcal{D}(A)$, then

$$\frac{d}{dt}\langle U(t)u, v\rangle_0 = \langle AU(t)u, v\rangle_0 = \langle U(t)u, A^*v\rangle_0 = 0.$$

We remark that elements $\{u_1, u_2\} \in M_1 \cap \mathcal{D}(A)$ possess the following property: $\mathrm{div} u_i = 0$ in the sense of distributions. It is not hard to show that a sequence of functions $u^n = \{u_1^n, u_2^n\}$ approximating an element $u \in M_1 \cap \mathcal{D}(A)$ can be chosen so that $\mathrm{div} u_i^n = 0$.

As in [7] we can now show that elements $u \in M_1 \cap \mathcal{D}(A)$ are contained in $H_1(\Omega)$ and therefore their traces on the boundary belong to $H_{1/2}(S)$, which makes it possible to treat $\omega(\alpha, u) = 0$ in the usual sense for Sobolev spaces.

The next theorem establishes the solvability of problem (1.1)–(1.3) in the class of functions needed for subsequent investigations.

THEOREM 2.1. *Suppose* $f(x) = \{f_1(x), f_2(x)\} \in M_1 \cap \mathcal{D}(A^n)$, Re $\alpha \geq 0, 0 < \lambda_0 \leq \lambda(x), 0 < \mu_0 \leq \mu(x), |\mathcal{D}_x^\beta \lambda| \leq C, |\mathcal{D}_x^\beta \mu| \leq C(|\beta| \leq n)$, *and* $\alpha(x) \in C^1(S)$. *Then there exists a unique solution* $u = \{e, h\}$ *of* (1.1)–(1.3) *such that* $A^j u \in H_1(\Omega), j = 0, 1, \cdots, n-1$. *Moreover,*

$$\|A^j u\|_0 \leq \|A^j f\|_0, \qquad j = 0, 1, \cdots, n.$$

Proof. It is obvious that the solution of (1.1)–(1.3) is given by $u = U(t)f$. Because of the properties of the semigroup $U(t)$,

$$A^j u = A^j U(t) f = U(t) A^j f \in M_1 \cap \mathcal{D}(A), \qquad j = 0, 1, \cdots, n-1.$$

From this it follows that $A^j u \in H_1(\Omega), j = 0, 1, \cdots, n-1$. Moreover,

$$\|A^j u\|_0 = \|A^j U(t) f\|_0 = \|U(t) A^j f\|_0 \leq \|A^j f\|_0 \qquad (j = 0, 1, \cdots, n),$$

which proves the assertion.

Let $f(x) = \{f_1(x), f_2(x)\} \in \mathcal{H}, f^n = \{f_1^n, f_2^n\} \in \mathcal{D}(A), \|f - f^n\|_0 \to 0$. Then $U(t)f^n$ satisfies the following identity:

$$\int_0^T \left(\left\langle U(t)f^n, \frac{d\Psi}{dt} \right\rangle_0 + \langle U(t)f^n, A^*\Psi\rangle_0 \right) dt = -\langle f, \Psi(0)\rangle_0,$$

where $\Psi \in L_2(0,T;\mathcal{D}(A^*)), \Psi_t \in L_2(0,T;\mathcal{H}), \Psi(T)=0$. From this we easily obtain

$$\int_0^T \left(\left\langle U(t)f, \frac{d\Psi}{dt} \right\rangle_0 + \langle U(t)f, A^*\Psi\rangle_0 \right) dt = -\langle f, \Psi(0)\rangle_0, \tag{2.1}$$

i.e., $U(t)f$ is the weak solution of the problem

$$u_t = Au, \qquad u|_{t=0} = f.$$

We note that $U(t)$ takes M_1 into itself. Indeed, if $g \in M$ and $\Psi(t) = (T-t)g$, then from (2.1) it follows that

$$\int_0^T \langle U(t)f, g\rangle_0 dt = \langle f, g\rangle_0 T.$$

Thus,

$$\langle U(t)f, g\rangle_0 = \langle f, g\rangle_0 \quad \text{for } t \geq 0$$

3. Stabilization. We start from geometrical conditions on Ω.

We consider the problem

$$\Delta\Phi = 1 \quad \text{in } \Omega,$$

$$\left.\frac{\partial\Phi}{\partial\nu}\right|_S = \frac{\text{mes } \Omega}{\text{mes } S},$$

which admits a solution $\Phi(x) \in C^2(\Omega) \cap C^1(\bar{\Omega})$.

For an arbitrary bounded domain Ω with smooth boundary S we define the following quantity:

$$æ(\Omega) = \sup_{\substack{x\in\Omega \\ |\xi|=1}} 2\,\text{Re}\ \Phi_{x_ix_j}\xi^i\bar{\xi}^j,$$

where $\xi = (\xi^1, \xi^2, \xi^3)$ is an arbitrary complex-valued vector. It is obvious that $æ(\Omega) \geq \frac{2}{3}$ and $æ(\{x : |x-x^0| < R\}) = \frac{2}{3}$.

We shall say that Ω is substarlike if

(i) $æ(\Omega) < 1$, or

(ii) $æ(\Omega) \geq 1$; there exists a point $x^0 \in \Omega$ such that for some $0 < \varepsilon \leq 1$

$$(x-x^0, \nu) > -\frac{1}{æ + \varepsilon - 1}\frac{\text{mes } \Omega}{\text{mes } S}.$$

We note that an arbitrary starlike domain ($(x-x^0,\nu) \geq 0$) is substarlike. It will follow from the continuity of $æ(\Omega)$ that there exist substarlike domains which are not starlike.

Henceforth we assume that Ω is substarlike domain.

LEMMA 3.1. *Assume that Ω is substarlike domain. Then there exists a function $\varphi(x) \in C^2(\Omega) \cap C^1(\bar{\Omega})$ such that*

(i) $(\nabla\varphi, \nu) > 0$ *on* S,

(ii) $2\,\text{Re}\varphi_{x_ix_j}\xi^i\bar{\xi}^j - \Delta\varphi|\xi|^2 + |\xi|^2 \leq (1-\omega)|\xi|^2$ *in* Ω, *where* $0 < \omega \leq 1$. *Moreover, for* $æ(\Omega) \geq 1$, $\omega = \varepsilon$.

Proof. For $\theta > 0, x^0 \in \Omega$ we set

$$\varphi(x) = \Phi(x) + \frac{1}{2\theta}|x - x^0|^2$$

so that

$$2\,\mathrm{Re}\ \varphi_{x_i x_j}\xi^i\bar{\xi}^j - \Delta\varphi|\xi|^2 + |\xi|^2 = 2\,\mathrm{Re}\ \Phi_{x_i x_j}\xi^i\bar{\xi}^j - \frac{1}{\theta}|\xi|^2 \leq \left(æ - \frac{1}{\theta}\right)|\xi|^2,$$

$$(\nabla\varphi, \nu) = (\nabla\Phi, \nu) + \frac{1}{\theta}(x - x^0, \nu) = \frac{\text{mes } \Omega}{\text{mes } S} + \frac{1}{\theta}(x - x^0, \nu).$$

Let $æ(\Omega) < 1$. Then for any $\theta > 0$ (ii) holds. If we choose θ so that

$$\theta > 2r(\text{mes } \Omega/\text{mes } S)^{-1} \quad (2r = d = \text{ diameter of } \Omega),$$

we obtain $(\nabla\varphi, \nu) > 0$ on S.

We now assume that $æ(\Omega) \geq 1$. In this case we choose x^0 as in the definition of Ω and set

$$\theta = \frac{1}{æ + \varepsilon - 1}.$$

The proof of Lemma 3.1 is complete.

The proof of stabilization is based on the invariance of the Maxwell system in vacuum ($\lambda(x) \equiv \lambda_0, \mu(x) \equiv \mu_0$) relative to the one-parameter group of dilations in all variables. This property of the Maxwell system leads to the identity

$$\begin{aligned}
&2\mathrm{Re}([\nabla\varphi, \bar{h}] + t\lambda\bar{e}, e_t - \mathrm{curl}(\mu h)) + 2\,\mathrm{Re}([\bar{e}, \nabla\varphi] + t\mu\bar{h}, h_t + \mathrm{curl}(\lambda e)) \\
&\quad = \frac{\partial}{\partial t}\{t(\lambda|e|^2 + \mu|h|^2) + 2\,\mathrm{Re}(\nabla\varphi, [h, \bar{e}])\} \\
&\quad -\mathrm{div}\{2t\lambda\mu\,\mathrm{Re}[h, \bar{e}] + \nabla\varphi(\lambda|e|^2 + \mu|h|^2) - 2\,\mathrm{Re}\ \lambda e(\bar{e}, \nabla\varphi) \\
&\quad -2\,\mathrm{Re}\ \mu h(\bar{h}, \nabla\varphi)\} - \{(\nabla\varphi, \nabla\lambda)|e|^2 + (\nabla\varphi, \nabla\mu)|h|^2\} \\
&\quad -\{2\,\mathrm{Re}\ \varphi_{x_i\bar{x}_j}(\lambda e^i\bar{e}^j + \mu h^i\bar{h}^j) - (\Delta\varphi - 1)(\lambda|e|^2 + \mu|h|^2)\} \\
&\quad -2\,\mathrm{Re}\ \lambda(\nabla\varphi, \bar{e})\mathrm{div}e - 2\,\mathrm{Re}\mu(\nabla\varphi, \bar{h})\mathrm{div}\ h,
\end{aligned} \tag{3.1}$$

where $\varphi(x) \in C^2(\Omega) \cap C^1(\bar{\Omega})$ is constructed in Lemma 3.1.

Let $f = \{f_1, f_2\} \in M_1 \cap \mathcal{D}(A)$ and $\{e, h\} = U(t)f$. From (3.1) after integration over the cylinder $\Omega \times (T_0, T)$ it follows that

$$\begin{aligned}
&\int_\Omega [t(\lambda|e|^2 + \mu|h|^2) + 2\,\mathrm{Re}\ (\nabla\varphi, [h, \bar{e}])]|_{t=T}dx \\
&\qquad - \int_\Omega [t(\lambda|e|^2 + \mu|h|^2) + 2\,\mathrm{Re}(\nabla\varphi, [h, \bar{e}])]|_{t=T_0}dx \\
&\quad = \int_{T_0}^T \int_\Omega [(\nabla\varphi, \nabla\lambda)|e|^2 + (\nabla\varphi, \nabla\mu)|h|^2]dx\,dt \\
&\qquad + \int_{T_0}^T \int_\Omega \{2\,\mathrm{Re}\ \varphi_{x_i x_j}(\lambda e^i\bar{e}^j + \mu h^i\bar{h}^j) \\
&\qquad -(\Delta - 1)(\lambda|e|^2 + \mu|h|^2)\}dx\,dt \\
&\qquad + \int_{T_0}^T \int_S \{2t\lambda\mu\,\mathrm{Re}(\nu, [h, \bar{e}]) + (\nabla\varphi, \nu)(\lambda|e|^2 + \mu|h|^2) \\
&\qquad -2\lambda\,\mathrm{Re}(e, \nu)(\bar{e}, \nabla\varphi) - 2\mu\mathrm{Re}(h, \nu)(\bar{h}, \nabla\varphi)\}dS\,dt \equiv I_1 + I_2 + I_3.
\end{aligned} \tag{3.2}$$

We have

$$2\int_\Omega \mathrm{Re}(\nabla\varphi,[h,\bar{e}])|_{t=T}dx \le \frac{d(\varphi)}{\sqrt{\lambda_0\mu_0}}\int_\Omega(\lambda|e|^2+\mu|h|^2)|_{t=T_0}dx,$$

where $d(\varphi)=\max_{x\in\Omega}|\nabla\varphi|$.

Using the boundary condition, we rewrite the last integral on the right side of (3.2) in the form

$$\begin{aligned}\int_{T_0}^{T}\int_S\{&-2t\lambda\mu|[h,\nu]|^2\,\mathrm{Re}\,\alpha-(\nabla\varphi,\nu)(\lambda|(e,\nu)|^2+\mu|(h,\nu)|^2)\\&+(\nabla\varphi,\nu)(\mu+\lambda|\alpha|^2)|[h,\nu]|^2-2\mathrm{Re}\lambda(\bar{e},\nu)([e,\nu],[\nabla\varphi,\nu])\\&-2\,\mathrm{Re}\,\mu(\bar{h},\nu)([h,\nu],[\nabla\varphi,\nu])\}dSdt\equiv\int_{T_0}^{T}\int_S\mathcal{B}\,dS\,dt.\end{aligned}$$

We now assume that $\mathrm{Re}\,\alpha(x)\ge b>0$. Let α_1,δ be such that

$$|\alpha(x)|\le\alpha_1, x\in S,\qquad(\nabla\varphi,\nu)\ge|\nabla\varphi|\delta,\quad x\in S.$$

We have

$$\begin{aligned}|2\,\mathrm{Re}\,\lambda(\bar{e},\nu)([e,\nu],[\nabla\varphi,\nu])|&\le|\nabla\varphi|\lambda(\delta|(e,\nu)|^2+\delta^{-1}|[e,\nu]|^2),\\|2\,\mathrm{Re}\,\mu(\bar{h},\nu)([h,\nu],[\nabla\varphi,\nu])|&\le|\nabla\varphi|\mu(\delta|(h,\nu)|^2+\delta^{-1}|[h,\nu]|^2).\end{aligned}$$

From these inequalities we obtain

$$\mathcal{B}\le-|[h,\nu]|^2\{2t\lambda_0\mu_0 b-d(\varphi)(\mu_1+\lambda_1\alpha_1^2)(1+\delta^{-1})\}.$$

We choose T_0 so that

$$T_0\ge\frac{d(\varphi)(1+\delta)(\mu_1+\lambda_1\alpha_1^2)}{2\lambda_0\mu_0 b\delta}\equiv T^*.$$

Thus, if $T>T_0\ge T^*$, it follows that

$$\int_{T_0}^{T}\int_S\mathcal{B}\,dS\,dt\le-\int_{T_0}^{T}2\lambda_0\mu_0 b(t-T^*)\int_S|[h,\nu]|^2\,dS\,dt\le0. \tag{3.3}$$

Let us assume for a moment that for some $0<\gamma\le1$

$$I_1+I_2\le(1-\gamma)\int_{T_0}^{T}\int_\Omega(\lambda|e|^2+\mu|h|^2)dxdt. \tag{3.4}$$

From (3.2) we obtain

$$T\|U(T)f\|_0^2\le\left(T^*+\frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}}\right)\|U(T^*)f\|_0^2+(1-\gamma)\int_{T^*}^{T}\|U(t)f\|_0^2dt.$$

Using the Gronwall inequality we find that

$$t\|U(t)f\|_0^2\le\left(T^*+\frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}}\right)\left(\frac{t}{T^*}\right)^{1-\gamma}\|U(T^*)f\|_0^2,\qquad t>T^*.$$

THEOREM 3.1. *Assume that* Ω *is substarlike*, Re $\alpha(x) \geq b > 0$,

$$(\nabla\varphi, \nabla\lambda) \leq \lambda(\omega - \gamma), (\nabla\varphi, \nabla\mu) \leq \mu(\omega - \gamma),$$

where $\varphi(x), \omega$ *are defined in Lemma* 3.1, $0 < \gamma \leq 1$. *Then for all* $f \in M_1, t > T^*$

$$\|U(t)f\|_0^2 \leq \left(T^* + \frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}}\right)(T^*)^{\gamma-1}\frac{1}{t^\gamma}\|f\|_0^2.$$

Proof. We approximate an arbitrary element $f \in M_1$ in the norm of $\mathcal{H}$ by pairs $g^n = \{g_1^n, g_2^n\}$ of smooth vector-functions g_i^n with $\omega(\alpha, g^n) = 0$.

Suppose $f^n = \{g_1^n - \lambda^{-1}\nabla\varphi_1^n, g_2^n - \mu^{-1}\nabla\varphi_2^n\}$, where

$$\operatorname{div}(\lambda^{-1}\nabla\varphi_1^n) = \operatorname{div} g_1^n, \quad \operatorname{div}(\mu^{-1}\nabla\varphi_2^n) = \operatorname{div} g_2^n, \qquad \varphi_i^n|_S = 0.$$

Then $f^n \in M_1 \cap \mathcal{D}(A)$ and

$$\|f^n - f\|_0^2 \leq \|g^n - f\|_0^2 + 2\|g^n\|_0\|g^n - f\|_0 \to 0, \qquad n \to \infty.$$

We have

$$t^\gamma\|U(t)f^n\|_0^2 \leq \left(T^* + \frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}}\right)\|f^n\|_0^2\frac{1}{(T^*)^{1-\gamma}}.$$

Letting $n \to \infty$, we get

$$t^\gamma\|U(t)f\|_0^2 \leq \left(T^* + \frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}}\right)\|f\|_0^2\frac{1}{(T^*)^{1-\gamma}}.$$

COROLLARY 3.1. $U(t)$ *takes the closed subspace* M_1 *into itself and*

$$\|U(t)\|_{M_1 \to M_1} < 1$$

for

$$t > T_1 = \left(T^* + \frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}}\right)^{1/\gamma}(T^*)^{1-1/\gamma}.$$

Using Pazy's theorem [16], we obtain Corollary 3.2.

COROLLARY 3.2. *Suppose* $f(x) = \{f_1, f_2\} \in M_1$. *There exist* $C, \beta > 0$ *such that*

$$\|U(t)f\|_0^2 \leq C\exp(-\beta t)\|f\|_0^2.$$

Remark 3.1. If Ω is a starlike domain, the estimate of Theorem 3.1 holds true under the following assumptions on λ, μ:

$$|\nabla\lambda| < \frac{\lambda}{2r}, \quad |\nabla\mu| < \frac{\mu}{2r} \text{ in } \bar{\Omega}, \qquad 2r = \text{diameter of } \Omega.$$

Indeed, let $\delta > 0$ be such that

$$|\nabla\lambda| \leq \frac{\lambda(1-\delta)}{2r}, \qquad |\nabla\mu| \leq \frac{\mu(1-\delta)}{2r}.$$

We choose $\theta > 0$ so that

$$\theta < \frac{2r\delta}{2r(æ + \varepsilon - 1) + |\nabla\Phi|(1-\delta)}.$$

We then obtain

$$\begin{aligned}
&\frac{1}{\lambda}(\nabla\lambda,\nabla\varphi)\lambda|e|^2+\frac{1}{\mu}(\nabla\mu,\nabla\varphi)|h|^2\\
&\quad=\left\{\frac{1}{\lambda}(\nabla\lambda,\nabla\Phi)+\frac{1}{\lambda\theta}(\nabla\lambda,x-x^0)\right\}\lambda|e|^2\\
&\quad+\left\{\frac{1}{\mu}(\nabla\mu,\nabla\Phi)+\frac{1}{\mu\theta}(\nabla\mu,x-x^0)\right\}\mu|h|^2\\
&\quad<\left(\frac{1}{\theta}-æ+1-\varepsilon\right)(\lambda|e|^2+\mu|h|^2),
\end{aligned}$$

whence we get the required inequality (3.4).

4. Exact controllability. In this section we shall use the estimate of Theorem 3.1 to prove exact controllability to an arbitrary state of solutions of (1.1), (1.2), (1.4).

In $\Omega\times(0,T)$ we consider the following problem:

$$\begin{aligned}
&e_t=\operatorname{curl}(\mu h), && \operatorname{div} e=0,\\
&h_t=-\operatorname{curl}(\lambda e), && \operatorname{div} h=0,\\
&e(x,0)=f_1(x), && h(x,0)=f_2(x),\\
&[\nu,e]-ia(x)(h-\nu(h,\nu))=p(x,t), && (x,t)\in S\times(0,T).
\end{aligned}\tag{4.1}$$

where λ,μ,S satisfy the conditions of Theorem 3.1, $a(x)\in C^i(S), f=\{f_1,f_2\}\in M_1$. Find a vector-function $p(x,t)$ such that the solution of (4.1) satisfies

$$e(x,T)=g_1(x),\qquad h(x,T)=g_2(x)$$

with an arbitrary pair $g=\{g_1,g_2\}\in M_1, T\geq T_1$.

Let $U(t)$ be the semigroup defined above ($\mathcal{J}m\alpha=a(x)$, Re $\alpha(x)=b>0$).

Consider the following equation in M_1

$$w-U^*(T)U(T)w=f-U^*(T)g.$$

The operator $G(T)=U^*(T)U(T)$ takes M_1 into itself and $\|G(T)\|<1$ for $T>T_1$. Thus we can solve this equation for any $f,g\in M_1$ and

$$\|w\|_0\leq \boldsymbol{C}(\|f\|_0+\|g\|_0).$$

Consequently, if we choose $w=(I-G(T))^{-1}(f-U^*(T)g)$, then

$$\{e(x,t),h(x,t)\}=U(t)w-(U^*(T-t)U(T)w-U^*(T-t)g)\equiv\{u,v\}-\{\tilde u,\tilde v\}$$

is a weak solution of (4.1) with

$$p(x,t)=b(v-\nu(v,\nu))+b(\tilde u-\nu(\tilde v,\nu)).$$

We observe that

$$\{e(x,T),h(x,T)\}=g(x)$$

and by the energy identity

$$\|p\|^2_{L_2(S\times(0,T))}\leq C(\|f\|_0^2+\|g\|_0^2).$$

Thus, $p(x,t)$ belongs to $L_2(S \times (0,T))$ and drives the solution of (4.1) to a desired terminal state $g = \{g_1, g_2\}$.

We have the following theorem.

THEOREM 4.1. *Assume that Ω is substarlike, $a(x) \in C^1(S), \lambda(x), \mu(x) \in C^1(\bar{\Omega}), 0 < \lambda_0 \le \lambda(x) \le \lambda_1, 0 < \mu_0 \le \mu(x) \le \mu_1$,*

$$(\nabla\varphi, \nabla\lambda) \le \lambda(w-\gamma), (\nabla\varphi, \nabla\mu) \le \mu(w-\gamma),$$

where $\varphi(x), w$ are defined in Lemma 3.1, $0 < \gamma \le 1$. Then for any $T > 2d(\varphi)(\lambda_0\mu_0)^{-1/2}(1-\gamma)^{1-1/\gamma}\gamma^{-1}$, given any pair of initial data $f = \{f_1, f_2\} \in M_1$ and any pair $g = \{g_1, g_2\} \in M_1$ there exists a boundary control $p(x,t) \in L_2(S \times (0,T))$ such that the corresponding solution of (4.1) satisfies

$$\{e(x,T), h(x,T)\} = \{g_1(x), g_2(x)\}.$$

Moreover,

$$\|p\|^2_{L_2(S\times(0,T))} \le C(\|f\|_0^2 + \|g\|_0^2).$$

We need only to explain that the control time T is an arbitrary quantity greater than

$$T_0 = \frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}} \frac{(1-\gamma)^{1-1/\gamma}}{\gamma}.$$

We note that T_1 is the function of b and

$$\inf\ T_1(b) = \frac{2d(\varphi)}{\sqrt{\lambda_0\mu_0}} \frac{(1-\gamma)^{1-1/\gamma}}{\gamma} = T_0.$$

Let $T > T_0$. We choose $b > 0$ such that

$$T_1(b) < T.$$

Then for $\alpha = b + ia(x)$ $\|U(T)\|_{M_1 \to M_1} < 1$ and we can solve the main relation

$$(I - U^*(T)U(T))w = f - U^*(T)g$$

for any $f, g \in M_1$.

Remark 4.1. If Ω is strictly starlike ($(x - x^0, \nu) > 0$) we can set $\varphi(x) = \frac{1}{2}|x - x^0|^2$. Then

$$d(\varphi) = \max_{\Omega} |\nabla\varphi| \le 2r = d = \text{diameter of } \Omega$$

and we have the best control time $T_0 = 2d/\sqrt{\lambda_0\mu_0}$ for the Maxwell system with $\lambda(x), \mu(x)$ such that

$$\frac{\partial\lambda}{\partial|x|} \le 0, \qquad \frac{\partial\mu}{\partial|x|} \le 0.$$

5. The case of discontinuous $\lambda(x), \mu(x)$. In this section we consider stabilization and exact controllability for Maxwell's equations in multilayered media.

These questions for Euler–Bernoulli beam equation in the one-dimensional case have been studied by G. Chen et al. [4].

Let us assume that $\partial\Omega = S$ is strictly star-shaped with respect to some point $x^0 \in \Omega$, i.e.,

$$(x - x^0, \nu) > 0.$$

With no loss of generality we suppose that $x^0 = 0$.

Assume that $B_k \subset \Omega$ is a bounded domain with sufficiently smooth boundary S_k, $\bar{B}_k \subset B_{k+1}$ for $k = 1, 2, \ldots, n$. Assume that $S_1, S_2, \ldots, S_n$ are star-shaped with respect to the origin, $\lambda(x)$ and $\mu(x)$ lose the continuity on these surfaces. We set

$$\Omega_0 = B_1, \quad \Omega_k = B_{k+1}\backslash\bar{B}_k \quad \text{for } k = 1, 2, \ldots, n-1, \quad \Omega_n = \Omega\backslash\bar{B}_n.$$

We consider here

$$\begin{aligned} &e_t = \operatorname{curl}(\mu h), && \operatorname{div} e = 0, \\ &h_t = -\operatorname{curl}(\lambda e), && \operatorname{div} h = 0, \\ &e(x,0) = f_1(x), && h(x,0) = f_2(x), \\ &[\nu, e] - \alpha(h - \nu(h,\nu)) = 0, && (x,t) \in S \times (0,T), \end{aligned} \tag{5.1}$$

$$\begin{aligned} [\lambda^k \nu, e^k] &= [\lambda^{k-1}\nu, e^{k-1}], \quad [\mu^k \nu, h^k] \\ &= [\mu^{k-1}\nu, h^{k-1}], \quad (x,t) \in S_k \times (0,T), \quad k = 1, \ldots, n, \end{aligned} \tag{5.2}$$

where $\nu = \nu(x)$ (for $x \in S_k$) is the unit normal vector to pointing into the exterior of B_k; $\lambda^k, \mu^k, e^k, h^k$ are the restrictions of corresponding functions on Ω_k.

In $\mathcal{H}$ we can define the unbounded operator $\tilde{A}$ in the same way as in previous sections:

$$\mathcal{D}(\tilde{A}) = \{\{e,h\} \in \mathcal{H} : \operatorname{curl} e^k, \operatorname{curl} h^k \in L_2(\Omega_k) \text{ for } k = 0, 1, \ldots, n,$$

$\omega(\alpha, \{e,h\}) = 0, \{e,h\}$ satisfies (5.2)$\} \equiv \tilde{\mathcal{H}}_1(\alpha)$, and

$$\tilde{A}\{e,h\} = \{\operatorname{curl}(\mu h), -\operatorname{curl}(\lambda e)\} \quad \text{for } \{e,h\} \in \mathcal{D}(\tilde{A}).$$

It can be shown in a similar way that $\tilde{A}$ and $\tilde{A}^*$ are dissipative for Re $\alpha \geq 0$. From it follows that $\tilde{A}$ generates a strongly continuous semigroup of contractions $\tilde{U}(t), t > 0$.

Let $\tilde{M}_1$ be the orthogonal complement of the kernel of $\tilde{A}^*$ in $\mathcal{H}$. It is not hard to show that

$$\operatorname{div} e^k = \operatorname{div} h^k = 0; \qquad e^k, h^k \in H_1(\Omega_k), \quad k = 0, 1, \ldots, n$$

for $\{e,h\} \in \mathcal{D}(\tilde{A}) \cap \tilde{M}_1$.

We remark that element $\Psi = \{\lambda^{-1}\nabla\varphi, 0\}$ belongs to the kernel of $\tilde{A}^*$ for an arbitrary $\varphi \in H_2(\Omega) \cap \overset{\circ}{H}_1(\Omega)$.

Thus, for $\{e,h\} \in \mathcal{D}(\tilde{A}) \cap \tilde{M}_1$ we have

$$\begin{aligned} 0 = \langle\{e,h\}, \Psi\rangle_0 = \int_{\Omega_0} (e^0, \nabla\bar{\varphi})dx + \ldots + \int_{\Omega_n} (e^n, \nabla\bar{\varphi})dx = \int_{S_1} (e^0, \nu)\bar{\varphi}\, dS \\ - \int_{S_1} (e^1, \nu)\bar{\varphi}\, dS + \ldots + \int_{S_n} (e^{n-1}, \nu)\bar{\varphi}\, dS - \int_{S_n} (e^n, \nu)\bar{\varphi}\, dS. \end{aligned}$$

Now we choose φ such that $\varphi = 0$ on $S_1, \dots, S_{j-1}, S_{j+1}, \dots, S_n$. Then

$$\int_{S_j} [(e^{j-1}, \nu) - (e^j, \nu)]\bar{\varphi} dS = 0$$

and we have

$$(e^{j-1}, \nu)|_{S_j} = (e^j, \nu)|_{S_j}, \qquad j = 1, 2, \dots, n. \tag{5.3}$$

It can be shown in the same way that

$$(h^{j-1}, \nu)|_{S_j} = (h^j, \nu)|_{S_j}, \qquad j = 1, 2, \dots, n. \tag{5.4}$$

The interface conditions (5.2)–(5.4) can also be found in the book by Dautray and Lions [5].

We assume that $\lambda^k(x) \equiv \lambda^k, \mu^k(x) = \mu^k$ for $x \in \Omega_k, k = 0, 1, \dots, n$.

Suppose $f \in \mathcal{D}(\tilde{A}) \cap \tilde{M}_1, \{e, h\} = \tilde{U}(t)f$. By an argument similar to the one in §3 we obtain ($\varphi = \frac{1}{2}|x|^2$)

$$\begin{aligned} &\int_\Omega \{t(\lambda|e|^2 + \mu|h|^2) + 2\,\mathrm{Re}(x, [h, \bar{e}])\}|_{t=T} dx \\ &\quad = \int_\Omega \{t(\lambda|e|^2 + \mu|h|^2) + 2\,\mathrm{Re}(x, [h, \bar{e})\}|_{t=T_0} dx \\ &\quad + \int_{T_0}^T \int_S \mathcal{B}\, dS\, dt + \sum_{k=1}^n \int_{T_0}^T \int_{S_k} (\mathcal{B}_{k-1} - \mathcal{B}_k) dS\, dt, \end{aligned} \tag{5.5}$$

where $\mathcal{B}$ is defined as in §3,

$$\begin{aligned} \mathcal{B}_k =& 2t\lambda^k\mu^k\,\mathrm{Re}(\nu, [h^k, \bar{e}^k]) + (x, \nu)(\lambda^k|e^k|^2 + \mu^k|h^k|^2) \\ &-2\,\mathrm{Re}\lambda^k(e^k, \nu)(\bar{e}^k, x) - 2\,\mathrm{Re}\ \mu^k(h^k, \nu)(\bar{h}^k, x). \end{aligned}$$

Using (5.2)–(5.4), we find that

$$\begin{aligned} t\lambda^k\mu^k(\nu, [h^k, \bar{e}^k]) &= t\lambda^{k-1}\mu^{k-1}(\nu, [h^{k-1}, \bar{e}^{k-1}]), \\ \lambda^k|e^k|^2 &= \lambda^k|(e^{k-1}, \nu)|^2 + \frac{(\lambda^{k-1})^2}{\lambda^k}|[e^{k-1}, \nu]|^2, \\ \mu^k|h^k|^2 &= \mu^k|(h^{k-1}, \nu)|^2 + \frac{(\mu^{k-1})^2}{\mu^k}|[h^{k-1}, \nu]|^2, \\ \lambda^k(e^k, \nu)(\bar{e}^k, x) &= \lambda^k(x, \nu)|(e^{k-1}, \nu)|^2 + \lambda^{k-1}(e^{k-1}, \nu)([\bar{e}^{k-1}, \nu], [x, \nu]), \\ \mu^k(h^k, \nu)(\bar{h}^k, x) &= \mu^k(x, \nu)|(h^{k-1}, \nu)|^2 + \mu^{k-1}(h^{k-1}, \nu)([\bar{h}^{k-1}, \nu], [x, \nu]). \end{aligned}$$

Hence

$$\begin{aligned} \mathcal{B}_{k-1} - \mathcal{B}_k = (x, \nu)\{&(\lambda^k - \lambda^{k-1})|(e^{k-1}, \nu)|^2 + (\mu^k - \mu^{k-1})|(h^{k-1}, \nu)|^2 \\ &+ \frac{\lambda^{k-1}}{\lambda^k}(\lambda^k - \lambda^{k-1})|[e^{k-1}, \nu]|^2 + \frac{\mu^{k-1}}{\mu^k}(\mu^k - \mu^{k-1})|[h^{k-1}, \nu]|^2\}. \end{aligned}$$

We now assume that

$$\lambda^k < \lambda^{k-1}, \quad \mu^k < \mu^{k-1}, \quad k = 1, 2, \dots, n; \qquad \mathrm{Re}\ \alpha = b(x) \geq b > 0.$$

Then $\mathcal{B}_{k-1} - \mathcal{B}_k \leq 0$ for $k = 1, 2, \ldots, n$, and from (5.5) it follows

$$T\|\tilde{U}(T)f\|_0^2 \leq \left(T_0 + \frac{2d}{\sqrt{\lambda^n \mu^n}}\right) \|\tilde{U}(T_0)f\|_0^2,$$

where

$$T > T_0 \geq \frac{d(1+\delta)(\mu^0 + \lambda^0 \alpha_1^2)}{2\lambda^n \mu^n b \delta} \equiv \tilde{T}^*, \qquad \frac{(x,\nu)}{|x|} \geq \delta > 0, \quad d = \max_{x \in \Omega} |x|.$$

As above we have

$$\|\tilde{U}(t)\|_{\tilde{M}_1 \to \tilde{M}_1} < 1$$

for

$$t > \tilde{T}^* + 2\frac{d}{\sqrt{\lambda^n \mu^n}}.$$

We can now show exact controllability for control time $T > 2d/\sqrt{\lambda^n \mu^n}$ for Maxwell's equations in multilayered media.

THEOREM 5.1. *Assume that Ω is strictly star-shaped, $\Omega_1, \Omega_2, \ldots, \Omega_n \subset \Omega$ are defined above. Suppose that $\lambda(x), \mu(x)$ are the piecewise constant functions in $\Omega, 0 < \lambda^n < \lambda^{n-1} < \ldots < \lambda^1 < \lambda^0$,*

$$0 < \mu^n < \mu^{n-1} < \ldots < \mu^1 < \mu^0 \qquad (\lambda(x) \equiv \lambda^k, \mu(x) = \mu^k, x \in \Omega_k).$$

Then for any $T > 2d(\lambda^n \mu^n)^{-1/2}$, given any pair of initial data $f = \{f_1, f_2\} \in \tilde{M}_1$ and any pair $g = \{g_1, g_2\} \in \tilde{M}_1$ there exists a boundary control $p(x,t) \in L_2(S \times (0,T))$ such that the corresponding solution of (4.1), (5.2) *satisfies*

$$\{e(x,T), h(x,T)\} = \{g_1(x), g_2(x)\}.$$

Moreover,

$$\|p\|_{L_2(S \times (0,T))}^2 \leq C(\|f\|_0^2 + \|g\|_0^2).$$

REFERENCES

[1] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–274.

[2] ———, *Control and stabilization for the wave equation in a bounded domain*, SIAM J. Control Optim., 17 (1979), pp. 66–81.

[3] ———, *Control and stabilization for the wave equation in a bounded domain*, Part II, SIAM J. Control Optim., 19 (1981), pp. 114–122.

[4] G. CHEN, M. C. DELFOUR, A. M. KRALL, AND G. PAYRE, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[5] R. DAUTRAY AND J. L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vols. 1 and 3, Springer-Verlag, Berlin, New York, 1990.

[6] G. DUVAUT AND J. L. LIONS, *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.

[7] B. V. KAPITONOV, *On exponential decay as $t \to \infty$ of solutions of an exterior boundary value problem for the Maxwell system*, Mat Sb., 180 (1089), pp. 469–490. Math. USSR Sb., 66 (1990), pp. 475–498.

[8] K. A. KIME, *Boundary controllability of Maxwell's equations in a spherical region*, SIAM J. Control Optim., 28 (1990), pp. 294–319.

[9] J. E. LAGNESE, *Exact boundary controllability of Maxwell's equations in a general region*. SIAM J. Control Optim., 27 (1989), pp. 374–388.

[10] ———, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J Differential Equations, 50 (1983), pp. 163–182.

[11] ———, *Boundary stabilization of linear elastodynamic system*, SIAM J. Control Optim., 21 (1983), pp. 968–984.

[12] L. D. Landau and E. M. Lipschitz, *Theoretical Physics*, Vol. 8, Nauka, Moscow, 1982.

[13] I. Lasiecka and R. Triggiani, *Uniform exponential energy decay in a bounded region with $L_2(0,T;L_2(\Omega))$-feedback control in the Dirichlet boundary conditions*, J. Differential Equations, 66 (1987), pp. 340–390.

[14] J. L. Lions, *Contrôlabilitè éxacte des systèmes distribués*, C. R. Acad. Sci. Paris Ser. I. Math., 302 (1986), pp. 471–475.

[15] ———, *Exact controllability, stabilization and perturbations for distributed system*, SIAM Rev., 30 (1988), pp. 1–68.

[16] A. Pazy, *On the applicability of Lyapunov's theorem in Hilbert space*, SIAM J. Math. Anal., 3 (1972), pp. 291–294.

[17] D. L. Russell, *The Dirichlet–Neumann boundary control problem associated with Maxwell's equations in a cylindrical region*, SIAM J. Control Optim., 24 (1986), pp. 199–229.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 421–441, March 1994

006

ON THE BOUNDEDNESS AND STABILITY OF SOLUTIONS TO THE AFFINE VARIATIONAL INEQUALITY PROBLEM*

M. SEETHARAMA GOWDA† AND JONG-SHI PANG‡

Abstract. This paper investigates the boundedness and stability of solutions to the affine variational inequality problem. The concept of a solution ray to a variational inequality defined by an affine mapping and on a closed convex set is introduced and characterized; the connection of such a ray with the boundedness of the solution set of the given problem is explained. In the case of the monotone affine variational inequality, a complete description of the solution set is obtained which leads to a simplified characterization of the boundedness of this set as well as to a new error bound result for approximate solutions to such a variational problem. The boundedness results are then combined with certain degree-theoretic arguments to establish the stability of the solution set of an affine variational inequality problem.

Key words. variational inequality, linear complementarity, solution ray, solution stability, degree theory, error bound

AMS subject classifications. 90C30, 90C33

1. Introduction.

This paper is a continuation of our recent effort in the study of the stability of variational inequalities, complementarity, and related problems. Our previous work [12], [11], [13], [29], which is typical among those of many authors [8], [15], [16], [19], [21], [30], [31], [34], [36], [37], has focused on the analysis of the behavior of a given solution to a problem under perturbation. In the present paper, we shall analyze the behavior of the entire solution set when the problem data are perturbed. Analysis of this kind has previously been performed for linear programs [32], [39], convex quadratic programs [4], [33], smooth generalized equations with bounded, convex solution sets [33], and linear complementarity problems of a certain type [5]. For the sensitivity analysis of a nonsmooth generalized equation, see [22].

The present paper deals with the affine variational inequality problem which is defined as follows. Given a nonempty polyhedron K in R^n, a vector $q \in R^n$, and a matrix $M \in R^{n\times n}$, this problem, denoted AVI (K, q, M), is to find a vector $x \in K$ such that

$$(y - x)^T(q + Mx) \geq 0 \text{ for all } y \in K.$$

Although the main stability results in the paper are obtained for this affine problem, they are derived with the aid of a number of auxiliary results that are valid for the more general case where K is an arbitrary closed convex set in R^n, not necessarily polyhedral. In order to distinguish the latter (semi-affine) case with the (fully) affine case, we shall drop the letter "A" in the prefix "AVI" when the set K is not restricted to be polyhedral. The solution set of the problem (A)VI (K, q, M) is denoted SOL(K, q, M). The primary objective of this paper is to study the behavior of this set as the pair (q, M) is perturbed (with K fixed). In particular, the following concept is central to our study. (Throughout this paper, $\|\cdot\|$ denotes an arbitrary vector norm in R^n.)

* Received by the editors April 6, 1992; accepted for publication (in revised form) November 13, 1992.
† Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21228. GOWDA@UMBC.BITNET.
‡ Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218. MSC_WJP@JHUVMS.HCF.JHU.EDU. The work of this author was based on research supported by National Science Foundation grant DDM-9104078.

DEFINITION 1. The problem VI (K, q, M) is said to be stable if for every $\varepsilon > 0$, there exists $\delta > 0$ such that for every pair $(q', M') \in R^n \times R^{n\times n}$ satisfying $\|q-q'\|+\|M-M'\| < \delta$,

$$\mathrm{SOL}(K, q', M') \cap (\mathrm{SOL}(K, q, M) + \varepsilon\mathcal{B}) \neq \emptyset,$$

where $\mathcal{B}$ denotes the open unit ball associated with the norm.

This stability concept is distinct from that of stability at a solution point [13]. Indeed, in the definition, there is no mention of any particular solution; the concept concerns the solution set of the problem as a whole. Implicit in the above requirement for stability is the nonemptiness of the solution set $\mathrm{SOL}(K, q, M)$ of the given variational inequality problem. In addition, the definition demands that for every ε-neighborhood of $\mathrm{SOL}(K, q, M)$ with $\varepsilon > 0$, there exists a suitable δ-neighborhood of the pair (q, M) such that for all perturbed data (q', M') in the latter neighborhood, the perturbed VI (K, q', M'), in addition to being solvable, must have a solution that lies in the former neighborhood. Hence, stability of the problem VI (K, q, M) implies that its solutions are well behaved in the sense that they will not change too drastically as the data (q, M) are slightly perturbed. In this paper, we shall derive sufficient conditions for the problem AVI (K, q, M) to be stable and investigate some related issues.

Having introduced the above stability concept, we should immediately mention a result of Robinson [33] which states that if M is positive semidefinite (i.e., in the monotone case), the problem AVI (K, q, M) is stable if and only if $\mathrm{SOL}(K, q, M)$ is nonempty and bounded. This result suggests that the boundedness of the solution set of AVI (K, q, M) might have some significance in the stability of this problem in the nonmonotone case. Guided by this insight, we shall undertake, as the first order of business, an in-depth study of the boundedness issue; this will be done for the VI (K, q, M). We shall introduce the notion of a solution ray for a variational inequality which generalizes that for a linear complementarity problem (LCP) introduced originally by Cottle [2]. A characterization of such a ray and conditions for its nonexistence are obtained in both the monotone and general case. These results extend those in [2] and [26] for the LCP, see also [20]. In the study of the monotone affine problem, we obtain a complete description of the solution set of the AVI (K, q, M) analogous to the well-known representation of Adler–Gale [1] for the monotone LCP. This representation of $\mathrm{SOL}(K, q, M)$ can be used to derive an error bound for approximate solutions of the monotone AVI; the derived bound is somewhat different from those obtained recently in [7] by a different approach.

Armed with the boundedness results, we shall then focus on the stability issue. Following the degree-theoretic approach in [16], [11], [13], and [29], we shall introduce a key degree property under which the stability of the AVI (K, q, M) will be established. Sufficient conditions for the validity of the degree property will be derived.

2. Preliminary discussion. In this section, we consider some fundamental facts about the AVI that are motivated by the special case, namely, the LCP. Though elementary, they provide some interesting insights for the AVI and are instrumental to the subsequent development. Included in this discussion is an existence result for the monotone AVI which shows that this problem is solvable if and only if it is "feasible" (in a sense to be made precise later).

To begin, we recall the well-known fact that [3, Prop. 1.5.2] when the set K is a cone in R^n, not necessarily polyhedral, the VI (K, q, M) is equivalent to the generalized complementarity problem:

$$x \in K, q + Mx \in K^*, x^T(q + Mx) = 0,$$

where K^* is the dual cone of K, i.e.,

$$K^* = \{y \in R^n : y^T z \geq 0 \text{ for all } z \in K\}.$$

We denote the latter problem as GCP (K, q, M). The prefix GLCP will be used when K is polyhedral.

In the case of the LCP (q, M) (= GLCP (R^n_+, q, M)), the complementary range $K(M)$ and complementary kernel SOL(0, M) = SOL$(R^n_+, 0, M)$ have played an important role; see [3]. The former consists of all vectors $q \in R^n$ for which the LCP (q, M) has a solution; the latter is a generalization of the null-space concept of a linear transformation. Much is known about these two special sets. With K being an arbitrary closed convex set in R^n, a natural question to ask is what the analogs of these two sets are for the VI (K, q, M).

To deal with the generalization of $K(M)$, we let $\mathcal{R}(K, M)$ denote the set of all vectors q for which SOL(K, q, M) is nonempty. When K is a cone, $\mathcal{R}(K, M)$ is also a cone, but not necessarily convex. In what follows, we give a geometric description of $\mathcal{R}(K, M)$. For this purpose, let $\mathcal{F}_x(K)$ be the cone of feasible directions of K at the point $x \in K$; i.e., $v \in \mathcal{F}_x(K)$ if and only if $x + \tau v \in K$ for all $\tau > 0$ sufficiently small.

PROPOSITION 1. *Let K be a closed convex set in R^n. Then for any vector $q \in R^n$, x is a solution of* VI *(K, q, M) if and only if $x \in K$ and $q + Mx \in \mathcal{F}_x(K)^*$. Consequently,*

$$\mathcal{R}(K, M) = \cup_{x \in K} (\mathcal{F}_x(K)^* - Mx). \tag{1}$$

Proof. Let $q \in \mathcal{R}(K, M)$, and $x \in$ SOL(K, q, M). We claim that $q + Mx \in \mathcal{F}_x(K)^*$. Indeed, let $v \in \mathcal{F}_x(K)$; then $x + \tau v \in K$ for all $\tau > 0$ sufficiently small. Hence, it follows that

$$0 \leq (x + \tau v - x)^T (q + Mx) = \tau v^T (q + Mx),$$

which establishes the claim. Conversely, take $q \in \mathcal{F}_x(K)^* - Mx$, where $x \in K$. For $z \in K$, we have $z - x \in \mathcal{F}_x(K)$ by the convexity of K. Hence,

$$0 \leq (z - x)^T (q + Mx)$$

by the choice of q. This establishes the characterization of solutions to the VI (K, q, M) and the identity (1). □

If K is polyhedral, it is possible to derive an alternate characterization of $\mathcal{R}(K, M)$ which establishes that this set is the union of a finite number of convex polyhedra, hence closed. In order to describe this result, we need an explicit representation of the polyhedron K in terms of a system of linear inequalities:

$$K = \{x \in R^n : Ax \geq b\} \tag{2}$$

for some matrix $A \in R^{m \times n}$ and vector $b \in R^m$. (It is important to point out that although this particular representation of K is used in the proof of several results below, these results are actually valid regardless of the represetation, as long as K is polyhedral.) For each index subset α of $\{1, \ldots, m\}$ with complement $\overline{\alpha}$, the α-face of K is defined by

$$F(\alpha) = \{x \in R^n : A_\alpha x = b_\alpha, A_{\overline{\alpha}} x \geq b_{\overline{\alpha}}\},$$

where A_γ denotes the rows of A indexed by the set $\gamma \subseteq \{1, \ldots, m\}$.

For the alternate characterization of $\mathcal{R}(K, M)$ and the results in the subsequent sections, we shall need some elementary facts from convex analysis [38]. For an arbitrary matrix

C, we shall let pos C denote the polyhedral cone generated by the columns of C. If C is vacuous, we shall let pos C be the singleton $\{0\}$. Also, for an arbitrary subset S of R^n, $\mathcal{H}S$ shall denote the convex hull of S, $\mathcal{C}S$ the conical hull of S, ri S the relative interior of S, and 0^+S the recession cone of S. It is well known that ri S is nonempty and convex for any nonempty convex set $S \subseteq R^n$; moreover, for any two convex sets S_1 and S_2 in R^n, we have

$$\text{ri}(S_1 + S_2) = \text{ri } S_1 + \text{ri } S_2. \tag{3}$$

Another easy fact about the relative interior is as follows. If C is a convex cone in R^n, then

$$C + \text{ri } C = \text{ri } C. \tag{4}$$

The proof of this identity is easy and left to the reader. Finally, a cone $C \subseteq R^n$ is *pointed* if $C \cap (-C) = \{0\}$. It is known that if C is a closed convex pointed cone, then int C^* is nonempty.

The following description of $\mathcal{R}(K, M)$ is reminiscent of the representation of $K(M)$ in terms of the complementary cones in the context of the LCP.

PROPOSITION 2. *In the above setting,*

$$\mathcal{R}(K, M) = \cup_\alpha(\text{pos}(A_\alpha)^T - MF(\alpha)), \tag{5}$$

where the union ranges over all subsets α *of* $\{1, \ldots, m\}$. *Hence,*

$$\mathcal{H}(\mathcal{R}(K, M)) \subseteq (0^+K)^* - MK := \mathcal{F}(K, M). \tag{6}$$

Moreover, if K *is a (polyhedral) cone, then* $\mathcal{R}(K, M)$ *is convex if and only if*

$$\mathcal{R}(K, M) = \mathcal{F}(K, M). \tag{7}$$

Proof. With K represented by (2), it follows that $q \in \mathcal{R}(K, M)$ if and only if there exist vectors $x \in K$ and $u \in R^m_+$ such that

$$\begin{aligned} &0 = q + Mx - A^T u, \\ &v = -b + Ax \geq 0, \quad u \geq 0, \quad u^T v = 0. \end{aligned} \tag{8}$$

From this set of complementarity conditions, the equality (5) follows easily. The inclusion (6) is also obvious by noting that $(0^+K)^* = \text{pos } A^T$. Finally, to prove the last assertion of the proposition, we need to establish only the "only if" part. In turn, it suffices to prove the reverse inclusion in (6). For this purpose, let $q = A^T u - Mx$ for some $u \in R^m_+$ and $x \in K$. Let $F(\alpha)$ be the face of K containing x. Then, we have

$$q = (A_\alpha)^T u_\alpha - Mx + (A_{\overline{\alpha}})^T u_{\overline{\alpha}}.$$

Since K contains the origin, the last expression shows that q belongs to the convex cone generated by $\mathcal{R}(K, M)$. But since $\mathcal{R}(K, M)$ is itself a (generally nonconvex) cone, its conical hull coincides with its convex hull. Consequently, the desired inclusion holds. □

Remark. The identity (5) is actually a special case of (1). This follows because of the fact that a vector x belongs to K if and only if $x \in F(\alpha)$ for some index set α for which $\text{pos}(A_\alpha)^T = \mathcal{F}_x(K)^*$. Consequently, the unions in the two identities are equal.

The set $\mathcal{F}(K, M)$ deserves some further discussion. First of all, we note that its definition is independent of the representation of K. If K is represented by (2), then $\mathcal{F}(K, M)$ consists of all vectors $q \in R^n$ for which there exist vectors $x \in K$ and $u \in R^m_+$ such that $q + Mx = A^T u$; i.e., it contains all vectors q for which the complementarity system (8) is feasible. For this reason, we can think of $\mathcal{F}(K, M)$ as the set of all "feasible" q for the AVI associated with the (fixed) pair (K, M). In this vein, the last assertion of Proposition 2 says that if K is a polyhedral cone, the set of solvable q (for the AVI associated with the fixed pair (K, M)) is convex if and only if for all q, the feasibility of AVI (K, q, M) implies its solvability. This conclusion generalizes the famous observation made by Eaves [6] for the special case of the LCP.

Note that the inclusion (6) is actually valid for an arbitrary closed convex set K. Indeed, since $0^+K \subseteq \mathcal{F}_x(K)$ for all $x \in K$, Proposition 1 and duality imply $\mathcal{R}(K, M) \subseteq \mathcal{F}(K, M)$; (6) follows easily because of the convexity of the latter set.

Next, we consider the generalization of the complementary kernel. For this purpose, we recall [3, Prop. 3.9.23] that for the LCP defined by the (fixed) matrix M, the complementary kernel SOL$(0, M)$ is equal to the singleton $\{0\}$ if and only if for all $q \in K(M)$, SOL(q, M) is compact. The following result generalizes this fact to the VI.

PROPOSITION 3. *Let K be a closed convex set in R^n. Let $\mathcal{S}(K, M)$ denote the solution set of the (homogeneous) $GCP(0^+K, 0, M)$. If $\mathcal{S}(K, M)$ consists of the zero vector alone, then* SOL(K, q, M) *is compact for all $q \in \mathcal{R}(K, M)$. The converse holds if K is in addition a cone.*

Proof. Suppose that for some $q \in \mathcal{R}(K, M)$, SOL(K, q, M) contains a sequence $\{x^k\}$ with $\|x^k\| \to \infty$. Without loss of generality, we may assume that the normalized sequence $\{x^k/\|x^k\|\}$ converges to a limit vector d which is clearly nonzero. We claim that $d \in \mathcal{S}(K, M)$. Indeed, from the inequality $(z - x^k)^T(q + Mx^k) \geq 0$ which holds for all $z \in K$, we deduce, by a standard normalization argument, that $d^T M d \leq 0$. Moreover, for any $y \in 0^+K$ we have $x^k + y \in K$ and

$$y^T(q + Mx^k) \geq 0$$

for all k. The claim follows from a standard normalization argument. If K is in addition a cone, then $K = 0^+K$; the last assertion is now obvious (by the cone property of $\mathcal{S}(K, M)$). □

We note that the equality $\mathcal{S}(K, M) = \{0\}$ provides a sufficient condition for the solution set of the VI (K, q, M) to be bounded for all q. In the next section, we shall study the boundedness of the latter set for a fixed but arbitrary vector q.

In the remainder of this section, we shall relate the three sets $\mathcal{S}(K, M)$, $\mathcal{F}(K, M)$ and $\mathcal{R}(K, M)$ under some additional assumptions on K and M. For this purpose, we recall that a matrix M is said to be copositive on a cone $C \subseteq R^n$ if $x^T Mx \geq 0$ for all $x \in C$; a copositive matrix M on C is *copositive-star* there if $x \in$ SOL$(C, 0, M) \Rightarrow -M^T x \in C^*$. Examples of copositive-star matrices include the positive semidefinite matrices and the copositive-plus matrices; the latter are those copositive matrices M that satisfy $x \in C, x^T Mx = 0 \Rightarrow (M + M^T)x = 0$. See [9], [10], and [14] for more discussion of the copositive matrices and their role in the generalized complementarity problem.

Suppose that M is copositive-star on the recession cone 0^+K. Then

$$\mathcal{S}(K, M) \subseteq \{v \in 0^+K : -M^T v \in (0^+K)^*\}.$$

We claim that the reverse inclusion also holds. Let v be a vector in the right-hand set. To show $v \in \mathcal{S}(K, M)$, it suffices to verify $Mv \in (0^+K)^*$. For this purpose, let $y \in 0^+K$.

Then for all $\tau > 0, y + \tau v \in 0^+K$. The copositivity property of M yields

$$0 \leq (y + \tau v)^T M(y + \tau v) \leq y^T My + \tau y^T Mv$$

which implies $y^T Mv \geq 0$. Consequently, we deduce

$$\mathcal{S}(K, M) = \{v \in 0^+K : -M^T v \in (0^+K)^*\};$$

or equivalently,

$$\mathcal{S}(K, M) = 0^+K \cap (-M(0^+K))^*.$$

If K is polyhedral, then $-M(0^+K)$ is a polyhedral cone; hence, from elementary convex analysis [38], we may deduce

$$\mathcal{S}(K, M)^* = (0^+K)^* - M(0^+K), \tag{9}$$

which implies

$$\mathcal{S}(K, M)^* - MK = (0^+K)^* - MK = \mathcal{F}(K, M). \tag{10}$$

This is the first assertion in the following existence theorem for the AVI.

THEOREM 1. *Let K be a polyhedron in R^n. If M is a copositive-star matrix on 0^+K, then* (10) *holds. If in addition M is positive semidefinite, then* (7) *holds.*

Proof. It suffices to show

$$\mathcal{F}(K, M) \subseteq \mathcal{R}(K, M) \tag{11}$$

under the additional positive semidefiniteness assumption. For this purpose, we represent K by (2). Since M is positive semidefinite, so is the matrix

$$\begin{bmatrix} M & -A^T \\ A & 0 \end{bmatrix},$$

which is the defining matrix for the equivalent (mixed) linear complementarity problem (8). The inclusion (11) now follows from [14, Thm. 3.1]. □

3. Solution ray. Generalizing the definition for the LCP [2], we introduce the following important concept.

DEFINITION 2. A nonzero vector $v \in R^n$ is said to be a (solution) ray of SOL(K, q, M) at the solution $x \in$ SOL(K, q, M) if $x + \tau v \in$ SOL(K, q, M) for all $\tau > 0$.

Clearly, if a solution ray exists, then SOL(K, q, M) is unbounded. Conversely, if either K is polyhedral or SOL(K, q, M) is convex (the latter is true if M is positive semidefinite; see [17], e.g.), and if SOL(K, q, M) is unbounded, then a solution ray exists. To prove this converse, we note that if K is polyhedral, then SOL(K, q, M) is the union of a finite number of convex polyhedra (see, e.g., the proof of Lemma 3.1 in [24]). Hence, if this solution set is unbounded, then one of these latter polyhedra must be unbounded, and thus has an extreme ray which is easily seen to be a solution ray of SOL(K, q, M) as defined above. On the other hand, if SOL(K, q, M) is convex (with K not necessarily polyhedral), then since SOL(K, q, M) is clearly closed, from elementary convex analysis [38] we know that this solution set is unbounded if and only if it has a recession direction which must be a solution ray of SOL(K, q, M).

The next result gives a necessary and sufficient condition for a nonzero vector to be a solution ray of the VI (K, q, M).

PROPOSITION 4. *Let K be a closed convex subset of R^n, and $x \in$ SOL(K, q, M) be given. Then a vector $v \neq 0$ is a solution ray of* SOL(K, q, M) *at x if and only if the following three conditions hold:*

(a) $v \in \mathcal{S}(K, M)$,
(b) $v^T(q + Mx) = 0$,
(c) $(z - x)^T Mv \geq 0$ *for all* $z \in K$.

Proof. The nonzero vector v is a solution ray as described if and only if for all $\tau > 0$ and all $z \in K$, we have $x + \tau v \in K$ and

$$(z - x - \tau v)^T (q + Mx + \tau Mv) \geq 0.$$

Expanding the left-hand side, we obtain the equivalent inequality:

$$0 \leq (z - x)^T (q + Mx) + \tau (z - x)^T Mv - \tau v^T (q + Mx) - \tau^2 v^T Mv. \tag{12}$$

Hence, if properties (a), (b), and (c) hold, then v must be a solution ray of SOL(K, q, M) at x. Conversely, if v is such a ray, then $v \in 0^+ K$ and the inequality (12) implies $v^T Mv \leq 0$. Moreover, letting $z = x$ in the same inequality, dividing by τ and letting $\tau \downarrow 0$, we obtain

$$v^T (q + Mx) \leq 0.$$

We put $z = x + \tau^2 v$ in (12), divide by τ^3, and let $\tau \to \infty$, to obtain $v^T Mv \geq 0$. Hence, $v^T Mv = 0$. From the proof of Proposition 1 and the fact that $\mathcal{F}_x(K)^* \subseteq (0^+ K)^*$, we deduce

$$v^T (q + Mx) \geq 0.$$

Hence, (b) follows. Since (12) holds for all $\tau > 0$, property (c) also holds. Letting $z = x + d$ where $d \in 0^+ K$ is arbitrary, we deduce that $Mv \in (0^+ K)^*$. This establishes (a) and completes the proof of the proposition. Q.E.D.

An equivalent way of stating (c) above is

$$x \in \operatorname{argmin}_{z \in K} z^T Mv. \tag{13}$$

Based on this observation, we may establish a necessary and sufficient condition for SOL(K, q, M) to contain no solution ray.

COROLLARY 1. *Let K be a closed convex subset of R^n and $q \in \mathcal{R}(K, M)$. A necessary and sufficient condition for* SOL(K, q, M) *to have no solution ray is that the implication below holds:*

$$\left.\begin{array}{c} 0 \neq v \in \mathcal{S}(K, M) \\ x \in (\operatorname{argmin}_{z \in K} z^T Mv) \cap \text{SOL}(K, q, M) \end{array}\right\} \Rightarrow v^T (q + Mx) > 0. \tag{14}$$

Proof. The sufficiency is obvious. The necessity is also easy by recalling the fact that if $x \in$ SOL(K, q, M), then $(q + Mx)^T v \geq 0$ for all $v \in 0^+ K$. Hence, if there is no solution ray, the desired implication must hold. □

Motivated by the assumption of the above corollary, we introduce a special property on the pair (K, M).

DEFINITION 3. We say that (K, M) has the sharp property if

$$\left.\begin{array}{c} v \in \mathcal{S}(K, M) \\ x \in \operatorname{argmin}_{z \in K} z^T Mv \end{array}\right\} \Rightarrow v^T Mx \geq 0. \tag{15}$$

Combining the sharp property with Corollary 1, we immediately obtain Corollary 2.

COROLLARY 2. *Let K be a closed convex subset of R^n. If the pair (K, M) has the sharp property, then the* VI (K, q, M) *has no solution ray for all vectors $q \in int(\mathcal{S}(K, M)^*)$ (where* int X *denotes the (topological) interior of a set X).*

Proof. It suffices to note that $q \in \text{int } (\mathcal{S}(K,M)^*)$ if and only if $q^T v > 0$ for all $0 \neq v \in \mathcal{S}(K,M)$. □

In what follows, we shall give several sufficient conditions for the sharp property to hold. Before stating these conditions, we observe that if K is a closed convex set containing the origin, then for every x satisfying the condition (13), we must have

$$x^T M v \leq 0.$$

Moreover, if K is a closed convex cone, then conditions (a) and (c) in Proposition 4 are equivalent to (a) and

$$x^T M v = 0. \tag{16}$$

Indeed, if (a) and (c) hold and $K = 0^+ K$, then by the above observation, we obtain

$$0 \geq -x^T M v \geq 0,$$

where the first inequality holds because $Mv \in (0^+ K)^*$; hence (16) holds. The converse can be proved easily by reversing the argument.

PROPOSITION 5. *Let K be a closed convex subset of R^n. Then the sharp property holds for the pair (K, M) under any one of the following conditions:*

(a) $\mathcal{S}(K,M) = \{0\}$;

(b) *M is positive semidefinite and for every $v \in \mathcal{S}(K,M)$, there exists $y \in K$ such that $v^T M y \geq 0$ (in turn, the latter condition holds if K contains the origin);*

(c) *K contains the origin and*

$$v \in \mathcal{S}(K,M) \Rightarrow (M + M^T)v \in (\mathcal{C}K)^*;$$

(d) *K contains the origin and M is copositive on $\mathcal{C}K$;*

(e) *K is a cone and M is symmetric.*

Proof. Part (a) is trivial. Consider part (b). Let (v, x) be a pair satisfying the left-hand conditions in the sharp implication (15). Since $v^T M v = 0$, the positive semidefiniteness of M implies $Mv + M^T v = 0$. Hence, by the definition of x and the particular vector y associated with v, we obtain

$$v^T M x = -x^T M v \geq -y^T M v = v^T M y \geq 0.$$

Hence the right-hand conclusion in (15) holds.

To prove part (c), let (v, x) be any pair of vectors satisfying the left-hand conditions in (15). As pointed out above, we must have $x^T M v \leq 0$. Also since $x \in K \subseteq \mathcal{C}K$, it follows from the assumption that

$$0 \leq x^T (M + M^T) v \leq v^T (Mx)$$

which is the desired right-hand condition in (15).

If M is copositive on $\mathcal{C}K$, we argue that any vector $v \in \mathcal{C}K$ with $v^T M v = 0$ must satisfy $(M + M^T)v \in (\mathcal{C}K)^*$. Indeed, if v is such a vector, then v is an optimal solution of the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} u^T M u, \\ \text{subject to} & u \in \mathcal{C}K. \end{array}$$

By the variational principle for this problem, it follows that v must satisfy $(M + M^T)v \in (\mathcal{C}K)^*$. Since $0 \in K$, we have $0^+K \subseteq \mathcal{C}K$. From this observation and the previous remark, part (d) can easily be seen to be a special case of (c).

Finally, part (e) is also a special case of (c) because if K is a cone then we must have $K = 0^+K = \mathcal{C}K$. □

Remark. In addition to (d) and (e), part (a) of the above proposition is also a special case of (c). We should also point out that $(\mathcal{C}S)^* = S^*$ for any set $S \subseteq R^n$.

In the degree-theoretic approach to be discussed later, it is essential for the solution sets of a certain family of AVIs to be uniformly bounded. The next result is concerned with this property; it generalizes Corollary 1.

PROPOSITION 6. *Let K be a polyhedral set in R^n. Let $q : [0, 1] \to R^n$ and $M : [0, 1] \to R^{n\times n}$ be continuous mappings. Suppose that for each $t \in [0, 1]$, there exists a $\delta > 0$ such that for all $t' \in [0, 1] \cap [t - \delta, t + \delta]$,*

$$(17)\quad \left.\begin{array}{c} 0 \neq v \in \mathcal{S}(K, M(t)) \\ x \in (\operatorname{argmin}_{z\in K} z^T M(t)v) \cap \operatorname{SOL}(K, q(t'), M(t')) \end{array}\right\} \Rightarrow v^T(q(t') + M(t')x) > 0.$$

Then the union

$$\bigcup_{t\in[0,1]} \operatorname{SOL}(K, q(t), M(t))$$

is bounded.

Proof. Assume the contrary. Then there exist sequences $\{t_k\} \subset [0, 1]$ and $\{x^k\}$ such that $\|x^k\| \to \infty$ and $x^k \in \operatorname{SOL}(K, q(t_k), M(t_k))$ for each k. Without loss of generality, we may assume that (i) the sequence $\{t_k\}$ converges to $\tau \in [0, 1]$, and (ii) the normalized sequence $\{x^k/\|x^k\|\}$ converges to a limit v which must be nonzero. Let K be represented by (2). As in the proof of (5), for each k, there exists an index set $\alpha_k \subseteq \{1, \ldots, m\}$ such that $x^k \in F(\alpha_k)$ and

$$\frac{1}{\|x^k\|}(q(t_k) + M(t_k)x^k) \in \operatorname{pos}(A_{\alpha_k})^T.$$

Since there are only finitely many such index sets, there exist index sets $\tilde{\alpha} \subseteq \{1, \ldots, m\}$ and $\kappa \subseteq \{1, 2, \ldots\}$ such that $\alpha_k = \tilde{\alpha}$ for all $k \in \kappa$. Let $\tilde{\beta}$ be the complement of $\tilde{\alpha}$ in $\{1, \ldots, m\}$. Then $v \in \mathcal{C}(\tilde{\alpha})$ where

$$\mathcal{C}(\tilde{\alpha}) = \{x \in R^n : A_{\tilde{\alpha}}x = 0, A_{\tilde{\beta}}x \geq 0\},$$

and $M(\tau)v \in \operatorname{pos}(A_{\tilde{\alpha}})^T$. Hence, $v \in \mathcal{S}(K, M(\tau))$. Moreover, for all $k \in \kappa$, we can easily prove that

$$x^k \in \operatorname{argmin}_{z\in K} z^T M(\tau)v \quad \text{and} \quad v^T(q(t_k) + M(t_k)x^k) = 0.$$

But this contradicts the assumption applied to $t = \tau$. □

It is useful to point out that the key requirement of the above proposition is easily satisfied if $M(t)$ is a constant for all $t \in [0, 1]$ and each individual $\operatorname{SOL}(K, q(t), M)$ is bounded. Indeed, the proof of this special case also follows from Robinson's well-known result about the locally upper Lipschitzian property of polyhedral multifunctions [35]. Nevertheless, Proposition 6 can not be proved directly from the cited result.

4. The monotone AVI. In this section, we give a complete description of the solution set of the monotone AVI. Based on this description, we derive an error bound result for approximate solutions of such a problem. First, we give a useful property of the solutions of a monotone VI.

LEMMA 1. *Let K be a closed convex set in R^n and M be positive semidefinite. Then there exist a vector $d \in R^n$ and a (nonnegative) scalar σ such that for all $x \in$* SOL(K, q, M),

$$(M + M^T)x = d \quad \text{and} \quad x^T M x = \sigma.$$

If K is a polyhedron represented by (2), *and if (x, u) is any pair satisfying* (8), *then*

$$\sigma + (q^T x - b^T u) = 0.$$

Proof. Let x^1 and x^2 be any two solutions of VI (K, q, M). Then we have

$$(x^1 - x^2)^T(q + Mx^2) \geq 0, \qquad (x^2 - x^1)^T(q + Mx^1) \geq 0.$$

Adding these two inequalities and rearranging terms, we derive

$$-(x^1 - x^2)^T M(x^1 - x^2) \geq 0.$$

The positive semidefiniteness of M easily implies

$$(M + M^T)x^1 = (M + M^T)x^2.$$

Hence, the existence of the vector d follows. From the last equality, we may deduce

$$(x^1)^T M x^1 = (x^2)^T M x^2,$$

which establishes the existence of the scalar σ.

If (x, u) satisfies the complementarity system (8), then it is easy to see that

$$0 = q^T x - b^T u + x^T M x$$

as desired. □

In the rest of this section, we shall focus on the AVI (K, q, M) where K is a nonempty polyhedral set in R^n. We introduce an important extended-valued function $\omega : R^n \to R \cup \{-\infty\}$ associated with this problem:

$$\omega(x) = \min_{z \in K} z^T(q + Mx);$$

note that $\omega(x)$ is the optimal objective value of a linear program parametrized by x. Of particular interest to us is the effective domain of ω; i.e., the set

$$\Omega = \{x \in R^n : \omega(x) > -\infty\}.$$

Since K is polyhedral, it follows that

$$K = \mathcal{H}G + \mathcal{C}H \tag{18}$$

for two finite sets G and H. With this representation, w have $0^+K = \mathcal{C}H$; moreover, it is easy to see that $\omega(x)$ is finite if and only if $q + Mx \in (0^+K)^*$. If K is represented by

(2), then $x \in \Omega$ if and only if there exists a vector $u \geq 0$ such that $0 = q + Mx + A^T u$. Summarizing this discussion, we conclude that Ω is a polyhedral set in R^n; moreover,

$$\begin{aligned} \Omega &= \{x \in R^n : q + Mx \in (0^+K)^*\} \\ &= \{x \in R^n : y^T(q + Mx) \geq 0 \text{ for all } y \in H\}. \end{aligned} \tag{19}$$

The second equality represents Ω in terms of a finite system of linear inequalities.

Next, we consider the set

$$\Omega' := \{x \in R^n : \omega(x) - (\sigma + q^T x) \geq 0\},$$

which must be a subset of Ω. This set, Ω', is also polyhedral and has the representation

$$\Omega' = \{x \in \Omega : z^T(q + Mx) - (\sigma + q^T x) \geq 0 \text{ for all } z \in G\}. \tag{20}$$

Note that if G is empty (or equivalently, if K is a cone), then $x \in \Omega \Leftrightarrow \omega(x) = 0$. In this case, the representation (20) reduces to

$$\Omega' = \{x \in \Omega : -(\sigma + q^T x) \geq 0\}.$$

In what follows, we adopt the convention that if G is empty, any term involving a vector in this vacuous set is interpreted as zero. The convention will enable us to treat this special case as a part of the general framework.

With the above preparation, we may now state the promised representation of the solution set of the monotone AVI (K, q, M).

THEOREM 2. *Let K be a polyhedron in R^n and M be a positive semidefinite matrix. Suppose* SOL$(K, q, M) \neq \emptyset$. *Let d and σ be the two invariants associated with the solutions of the* AVI *(K, q, M) (see Lemma* 1). *Then*

$$\text{SOL}(K, q, M) = \{x \in K \cap \Omega' : (M + M^T)x = d\}. \tag{21}$$

Proof. We first show the inclusion

$$\text{SOL}(K, q, M) \subseteq \{x \in K \cap \Omega' : (M + M^T)x = d\}.$$

Let $x \in$ SOL(K, q, M). It suffices to verify $x \in \Omega'$. As mentioned several times in the previous sections, we have $q + Mx \in (0^+K)^*$, which implies that $x \in \Omega$. Moreover, using the fact that $\sigma = x^T Mx$ and the inequality

$$(z - x)^T(q + Mx) \geq 0,$$

which holds for all $z \in K$, it follows easily that $x \in \Omega'$. This establishes the desired inclusion.

To prove the reverse inclusion, let $x \in K \cap \Omega'$ satisfy $(M + M^T)x = d$. Then by the definition of d, for some solution $\overline{x} \in$ SOL(K, q, M),

$$(M + M^T)x = (M + M^T)\overline{x},$$

which implies $x^T Mx = \overline{x}^T M\overline{x} = \sigma$. Let $z \in K$ be arbitrary. Since $x \in \Omega'$, we have

$$z^T(q + Mx) \geq \sigma + q^T x = x^T(q + Mx),$$

which shows that $x \in$ SOL(K, q, M) as desired. The proof of the theorem is now complete. □

When $K = R^n_+$, the above theorem reduces to the well-known polyhedral representation of the solution set of a monotone LCP [1], [3]. This theorem can also be established by applying the latter result to the equivalent complementarity system (8) and utilizing a standard theorem of the alternatives to remove the multiplier vector u. The proof given here is more direct and brings out the sets Ω and Ω' which play an important role in the error bound analysis.

According to Theorem 1, we know that for a monotone AVI (K, q, M), $\mathrm{SOL}(K, q, M) \neq \emptyset$ if and only if $q \in \mathcal{F}(K, M)$. In principle, we could use the representation (21) to give some simplified characterizations for the boundedness of $\mathrm{SOL}(K, q, M)$ when M is positive semidefinite. Nevertheless, we shall postpone the derivation of these conditions until the next section, where we shall treat a more general situation; see Theorem 7.

An error bound result. As promised, we now show how the representation (21) of the solution set of the AVI (K, q, M) yields an error bound for approximate solutions to this problem that is distinct from the one obtained in [7].

THEOREM 3. *Let K, M, d and σ be as given in Theorem 2. Then there exists a constant $L > 0$ such that for any $x \in \Omega$, there exists a $\overline{x} \in \mathrm{SOL}(K, q, M)$ such that*

$$\|x - \overline{x}\| \leq L[d(x, K) + (\omega(x) - (\sigma + q^T x))_- + \|(M + M^T)x - d\|], \tag{22}$$

where $d(x, K)$ denotes the distance from x to the set K.

Proof. Let K be represented by (2) and (18). The equation (21) defines $\mathrm{SOL}(K, q, M)$ as the solution set of a system of linear inequalities (see also (20)). By the famous Hoffman bound for approximate solutions to such a system [18,25], we deduce the existence of a constant $c > 0$ such that for each $x \in \Omega$, there exists $\overline{x} \in \mathrm{SOL}(K, q, M)$ such that

$$\|x - \overline{x}\| \leq c[\|(Ax - b)_-\| + \max\{(z^T(q + M\dot{x}) - (\sigma + q^T x))_- : z \in G\} + \|(M + M^T)x - d\|].$$

To complete the proof, it remains to show that there exists a constant $c_1 > 0$ such that for all $x \in R^n$,

$$\|(Ax - b)_-\| \leq c_1 d(x, K),$$

and for all $x \in \Omega$

$$\max\{(z^T(q + Mx) - (\sigma + q^T x))_- : z \in G\} \leq (\omega(x) - (\sigma + q^T x))_-. \tag{23}$$

The existence of the constant c_1 is a consequence for the Lipschitz continuity of the function $f(x) = \|(Ax - b)_-\|$. By the definition of the function ω, we have for all $z \in K$,

$$z^T(q + Mx) - (\sigma + q^T x) \geq \omega(x) - (\sigma + q^T x),$$

which easily implies

$$(z^T(q + Mx) - (\sigma + q^T x))_- \leq (\omega(x) - (\sigma + q^T x))_-,$$

from which (23) follows. □

The inequality (22) suggests that for the monotone AVI (K, q, M), the function

$$r(x) = d(x, K) + (\omega(x) - (\sigma + q^T x))_- + \|(M + M^T)x - d\|$$

provides an appropriate residue for all vectors $x \in \Omega$. Note that if $x \notin \Omega$, the $r(x) = \infty$. Hence, the error bound (22) is trivially valid for such a vector x, but it does not offer any

effective information about the distance $d(x, \text{SOL}(K, q, M))$. (Of course, we know for certainty that this x is not a solution of the AVI.) Assumptions similar to this restriction ($x \in \Omega$) are also needed in [7].

Two more remarks can be made about the above error bound result. (1) as can be seen from the proof of Theorem 3, if an explicit inequality representation of K is given, then the distance $d(x, K)$ can be replaced by an appropriate residue function defined by such inequalities. (2) the invariants d and σ appear in the residue function $r(x)$. We believe that, as in the case of the LCP [27] and also in [7], it might be possible to get rid of them under some additional assumptions. Since this topic is not the major concern of the present paper, we choose not to pursue it further.

We close this section by mentioning that error bound results such as Theorem 3 are in general useful for several reasons. (1) (quoting a referee) they provide insights on how approximate solutions approach the solution set; (2) they are instrumental for establishing the rate of convergence of iterative algorithms [24]; and (3) they can be used in the design of inexact iterative algorithms.

5. Main stability results. As in the previous studies [11], [13], [16], [29], our approach to analyze the stability of the AVI is based on degree theory. In addition to the review given in these references, the reader can consult [23], and [28] for the fundamentals of this powerful tool.

Consider the VI (K, q, M) and assume that this problem has a nonempty bounded solution set. Let $\mathcal{O}$ be the family of bounded open sets containing $\text{SOL}(K, q, M)$. In order to employ degree theory, we need to transform the VI into a system of equations. As it is well known (see [17], for example), the VI (K, q, M) is equivalent to

$$F_{(q,M)}(x) := x - \Pi_K(x - (q + Mx)) = 0,$$

where Π_K denotes the projection operator onto the set K. Note that the set K does not appear in the subscript of the mapping $F_{(q,M)}$. The reason for this omission is that, unlike q and M, K is fixed throughout this stability study.

Clearly, for any $D \in \mathcal{O}$, the mapping $F_{(q,M)}$ does not vanish on ∂D, the boundary of the set D. Hence, the degree, denoted deg $(F_{(q,M)}, D)$, of $F_{(q,M)}$ at zero relative to D is well defined. The following theorem is the basis for the degree-theoretic approach to stability analysis.

THEOREM 4. *Let K be a closed convex set in R^n. Suppose* $\text{SOL}(K, q, M)$ *is nonempty and bounded. Assume that for some $D \in \mathcal{O}$, deg $(F_{(q,M)}, D)$ is nonzero. Then* VI (K, q, M) *is stable in the sense of Definition* 1.

Proof. Let $\varepsilon > 0$ be given; consider the open set

$$V = \{x \in R^n : d(x, \text{SOL}(K, q, M)) < \varepsilon\}.$$

Without loss of generality, we can, by restricting ε, assume that $V \subseteq D$. By the excision property of the degree, we have deg $(F_{(q,M)}, D)$ = deg $(F_{(q,M)}, V)$. By the nonexpansiveness of the projection operator, we may choose a suitable $\delta > 0$ such that for all $\|M' - M\| + \|q' - q\| \leq \delta$,

$$\sup_{x \in V} \|F_{(q,M)}(x) - F_{(q',M')}(x)\| < d(0, F_{(q,M)}(\partial V)).$$

By the nearness property of the degree, it follows that deg $(F_{(q',M')}, V) =$ deg$(F_{(q,M)}, V) \neq 0$. Hence, the equation

$$F_{(q',M')}(x) = 0$$

has a solution in V, or equivalently,

$$\mathrm{SOL}(K, q', M') \cap V \neq \emptyset.$$

Consequently, the desired stability conclusion follows. □

Remark. In general, if the VI (K, q, M) is stable, then we must have $q \in \operatorname{int} \mathcal{F}(K, M)$.

Theorem 4 has reduced the stability question related to the VI (K, q, M) to a nonvanishing property of a degree of the (nonsmooth) mapping $F_{(q,M)}$. Our next result, which is based on the homotopy invariance of the degree, provides a useful way to validate the latter degree condition.

THEOREM 5. *Let K be a closed convex set in R^n and $(q, M) \in R^n \times R^{n\times n}$ be arbitrary. Suppose there exists $(q^*, M^*) \in R^n \times R^{n\times n}$ satisfying the following two conditions:*

(A) $\mathrm{SOL}(K, q^*, M^*)$ *is nonempty and bounded, and for some open bounded set D containing* $\mathrm{SOL}(K, q^*, M^*)$, $\deg(F_{(q^*,M^*)}, D)$ *is nonzero;*

(B) *there exists a homotopy $H : [0, 1] \to R^n \times R^{n\times n}$ connecting the pairs (q^*, M^*) and (q, M) such that the set*

$$\bigcup_{t\in[0,1]} \mathrm{SOL}(K, q(t), M(t)) \tag{24}$$

is bounded, where $H(t) = (q(t), M(t))$.

Then, the VI *(K, q, M) is stable; in particular,* $\mathrm{SOL}(K, q, M)$ *is nonempty and bounded, and $q \in \operatorname{int} \mathcal{F}(K, M)$.*

Proof. Without loss of generality, we may assume, using the excision property of the degree, if necessary, that D contains the union in (24). Clearly,

$$\tilde{F}(t) = x - \Pi_K(x - (q(t) + M(t)x))$$

defines a homotopy connecting the two mappings $F_{(q,M)}$ and $F_{(q^*,M^*)}$. By the homotopy invariance of the degree, it follows that $\deg(F_{(q,M)}, D) = \deg(F_{(q^*,M^*)}, D)$; the latter degree is nonzero by assumption. Hence the desired conclusions follow easily from the last theorem. □

Specializing the above theorem to the case where M is positive semidefinite, we obtain the following stability result for the monotone VI (K, q, M).

COROLLARY 3. *Let K be a closed convex set in R^n and M be a positive semidefinite matrix. If* $\mathrm{SOL}(K, q, M)$ *is nonempty and bounded, then the* VI *(K, q, M) is stable.*

Proof. Pick any $x^* \in \mathrm{SOL}(K, q, M)$; define $q^* = -x^*$ and $M^* = I$. Then the map $F_{(q^*,M^*)}$ is equal to the identity map translated by a constant. Hence $\deg(F_{(q^*,M^*)}, D) = 1$ for any open bounded set D containing $\mathrm{SOL}(K, q, M)$. Since M is positive semidefinite, the matrix $M(t) := tM + (1 - t)M^*$ is positive definite for all $t \in [0, 1)$.

Hence with $q(t) := tq + (1 - t)q^*$, it follows that

$$\mathrm{SOL}(K, q(t), M(t)) = \{x^*\}, \quad \text{for all } t \in [0, 1).$$

Consequently, condition (B) of Theorem 5 holds. The desired conclusion follows. □

It is important to point out that the above corollary does not follow from any existing stability theory for the VI; in particular, the results in [33] cannot be used to establish this corollary because they require a certain upper Lipschitzian assumption which has been shown to be valid only in the case of a polyhedral K. Incidentally, the proof of Corollary 3 provides a simple demonstration of one implication in Robinson's result mentioned in the Introduction. This matter will be addressed in full detail in Theorem 7.

Specializing Theorem 5 to the AVI and invoking Proposition 6, we derive the following result.

THEOREM 6. *Let K be a polyhedron in R^n, and $(q, M) \in R^n \times R^{n\times n}$ be arbitrary. Suppose that there exists $(q^*, M^*) \in R^n \times R^{n\times n}$ satisfying the following three conditions:*

(A) $|\mathrm{SOL}(K, q^*, M^*)| = 1$;

(B) *for all $\tilde{q}$ sufficiently close to q^*, $|\mathrm{SOL}(K, \tilde{q}, M^*)| \leq 1$;*

(C) *for each $t \in [0, 1]$, there exists a $\delta > 0$ such that for all $t' \in [0, 1] \cap [t - \delta, t + \delta]$, the implication* (17) *holds for the following homotopies:*

$$M(t) := tM + (1 - t)M^*, \qquad q(t) = tq + (1 - t)q^*.$$

Then the conclusions of Theorem 5 *are valid for the* AVI *(K, q, M).*

Proof. It suffices to show for some bounded open set D containing the unique solution x^* of the AVI (K, q^*, M^*), $\deg(F_{(q^*,M^*)}, D)$ is nonzero. For this purpose, let U be an open neighborhood of q^* such that for all $\tilde{q} \in U$, $|\mathrm{SOL}(K, \tilde{q}, M^*)| \leq 1$. Let ε be a positive scalar with the property that $q^* + (M^* - I)y \in U$ for all y with $\|y\| < \varepsilon$. Then for any such vector y, the equation

$$F_{(q^*,M^*)}(x) = y$$

has at most one solution. Indeed, any solution of this equation is a solution of the AVI $(K, q^* + (M^* - I)y, M^*)$; by the choice of y, the latter AVI has at most one solution. Now, choose an open neighborhood D of x^* such that $\|F_{(q^*,M^*)}(x)\| < \varepsilon$ for all $x \in D$. Hence the restricted map $F_{(q^*,M^*)}\colon D \to R^n$ is one-to-one, and $0 \in F_{(q^*,M^*)}(D)$. Since the degree of an injective map is ± 1 [23, Thm. 3.3.3], it follows that $\deg(F_{(q^*,M^*)}, D)$ is nonzero. □

The above theorem has identified a set of sufficient conditions for the satisfaction of the key assumptions (A) and (B) of Theorem 5. In what follows, we derive various consequences of the latter theorem. Our first corollary pertains to the case where $\mathcal{S}(K, M)$ is a singleton.

COROLLARY 4. *Let K be a polyhedron in R^n. Suppose that M is copositive on 0^+K and $\mathcal{S}(K, M) = \{0\}$. Then the* AVI *$(K, q, M)$ is stable for all $q \in R^n$.*

Proof. Let $M^* = I$ and q^* be arbitrary. It suffices to verify condition (C) of Theorem 6. Since M is copositive on 0^+K, it follows that for all $t \in [0, 1)$, $\mathcal{S}(K, M(t)) = \{0\}$; moreover, $\mathcal{S}(K, M(1)) = \{0\}$ by assumption. □

We shall next derive some stability results for the AVI by relaxing the assumption that $\mathcal{S}(K, M)$ is a singleton. Motivated by the sharp property of the pair (K, M) defined in Section 3, we introduce two important subsets for R^n. Specifically, let $\mathcal{B}(K, M)$ denote the set of vectors $q \in R^n$ for which the implication below holds:

$$\left.\begin{array}{r} 0 \neq v \in \mathcal{S}(K, M) \\ x \in \mathrm{argmin}_{z\in K} z^T M v \end{array}\right\} \Rightarrow v^T(q + Mx) > 0,$$

and let $\mathcal{Q}(K, M)$ denote the set of vectors $q \in R^n$ for which the following (less restrictive) implication holds:

$$\left.\begin{array}{r} v \in \mathcal{S}(K, M) \\ x \in \mathrm{argmin}_{z\in K} z^T M v \end{array}\right\} \Rightarrow v^T(q + Mx) \geq 0.$$

Clearly, the latter set is closed and contains the closure of the former set. Moreover, Corollary 1 implies that if K is polyhedral and $q \in \mathcal{B}(K, M)$, then $\mathrm{SOL}(K, q, M)$ is

bounded if it is nonempty. The next result is concerned with how the two sets just introduced are related to the range $\mathcal{R}(K, M)$ of the pair (K, M) and to the stability of the associated AVIs.

COROLLARY 5. *Let K be a polyhedron in R^n. Suppose that there exists a vector $q^* \in \mathcal{Q}(K, M)$ such that* SOL(K, q^*, M) *is nonempty and bounded, and for some open bounded set D containing* SOL(K, q^*, M), deg $(F_{(q^*, M)}, D)$ *is nonzero. Then, for all $q \in \mathcal{B}(K, M)$, the* AVI *(K, q, M) is stable and* SOL(K, q, M) *is bounded. If in addition int $(\mathcal{S}(K, M)^*)$ is nonempty, then $\mathcal{R}(K, M) \supseteq \mathcal{Q}(K, M)$.*

Proof. Consider the homotopies

$$M(t) := M, \qquad q(t) = tq + (1 - t)q^*.$$

By the choice of q and q^*, it follows that the implication below holds for all $t \in (0, 1]$:

$$\left.\begin{array}{c} 0 \neq v \in \mathcal{S}(K, M) \\ x \in (\text{argmin}_{z \in K} z^T M v) \cap \text{SOL}(K, q(t), M) \end{array}\right\} \Rightarrow v^T(q(t) + Mx) > 0.$$

By assumption, this implication is also valid for $t = 0$ (see Corollary 1). Hence, the two main conditions of Theorem 5 are satisfied; the desired stability conclusion follows readily.

To establish the second assertion, let $u \in$ int $(\mathcal{S}(K, M)^*)$ and $q \in \mathcal{Q}(K, M)$. Then clearly $q + \varepsilon u \in \mathcal{B}(K, M)$ for all $\varepsilon > 0$. Since the range $\mathcal{R}(K, M)$ is a closed set by Proposition 2, it follows that $q \in \mathcal{R}(K, M)$ in view of what has been proved above. □

The above corollary is fairly general in that no particular assumption is imposed on the pair (K, M). It identifies a sufficient condition under which a certain set of vectors q can be associated with the pair (K, M) for which the AVI (K, q, M) is stable. In the sequel, we shall derive various consequences of this corollary by making some special assumptions on (K, M).

If the pair (K, M) has the sharp property, then clearly $\mathcal{S}(K, M)^* \subseteq \mathcal{Q}(K, M)$. In this case, the last corollary implies that if there exists a vector $q^* \in \mathcal{S}(K, M)^*$ satisfying the stated assumption therein, then for all $q \in$ int $(\mathcal{S}(K, M)^*)$, the AVI (K, q, M) is stable. Moreover, if such a vector q indeed exists, then $\mathcal{S}(K, M)^* \subseteq \mathcal{R}(K, M)$.

If K is a pointed polyhedral cone in R^n and M is copositive on K, it is easy to obtain a vector q^* satisfying the degree requirement in Corollary 5. To construct this vector, note that int K^* must be nonempty by the pointedness of K. Pick any vector q^* in the latter interior. Then, by the copositivity of M on K and the interiority of q^*, it is obvious that $|\,\text{SOL}(K, \tilde{q}, M)| = 1$ for all $\tilde{q}$ sufficiently close to q^*; indeed SOL$(K, \tilde{q}, M) = \{0\}$ for any such vector $\tilde{q}$. Hence, as in Theorem 6, we may deduce that deg $(F_{(q^*, M)}, D)$ is nonzero. Summarizing this discussion and recalling part (c) of Proposition 5, we immediately obtain the following consequence of Corollary 5.

COROLLARY 6. *Let K be a pointed polyhedral cone in R^n. If M is copositive on K, then the* GLCP *(K, q, M) is stable for all $q \in$ int $(\mathcal{S}(K, M)^*)$.*

We now come to our main stability theorem for a "copositive AVI." The significance of this result is threefold. Firstly, it recovers Robinson's characterization of the stability of a monotone AVI in terms of the nonemptiness and boundedness of the solution set. Secondly, the theorem adds to this characterization several equivalent conditions each of which is interesting in its own right; collectively, these conditions have appeared in [26] for the special case of the LCP with a copositive-plus matrix. Thirdly and most importantly, the theorem deals with a nonmonotone AVI. It turns out that in order to remove the monotonicity assumption of M, we need to substitute in its place the assumption (A) of Theorem 5. In the proof of the following theorem, the two identities (3) and (4) are needed.

THEOREM 7. *Let $K = \mathcal{H}G + \mathcal{C}H$ be a polyhedron in R^n and M be copositive-plus on 0^+K. Suppose that either*

(A) *there exists a vector $q^* \in R^n$ such that* SOL(K, q^*, M) *is nonempty and bounded, and for some open bounded set D containing* SOL(K, q^*, M), *deg $(F_{(q^*,M)}, D)$ is nonzero, or*

(B) *M is positive semidefinite.*

Consider the five statements below:

(a) *the* AVI *(K, q, M) is stable in the sense of Definition* 1;

(b) SOL(K, q, M) *is nonempty and bounded;*

(c) $q \in$ int $\mathcal{F}(K, M)$;

(d) $q \in$ ri $\mathcal{S}(K, M)^* - MK$;

(e) $q \in$ ri $\mathcal{S}(K, M)^* - M(\mathcal{H}G)$. *Then the following relations hold:*

$$(a) \Leftrightarrow (b) \Leftrightarrow (c) \Rightarrow (d) \Leftrightarrow (e). \tag{25}$$

Moreover, if int $(\mathcal{S}(K, M)^*)$ *is nonempty, then all five statements are equivalent and*

$$\mathcal{R}(K, M) = \mathcal{Q}(K, M) = \mathcal{F}(K, M). \tag{26}$$

Proof. The implication (a) $\Rightarrow$ (c) follows from the remark following Theorem 4. The implication (b) $\Rightarrow$ (a) follows from Corollary 3 when (B) holds. We now show (b) $\Rightarrow$ (a) under the assumption (A). In view of Corollary 5, the proof of this implication consists of verifying that q^* belongs to the set $\mathcal{Q}(K, M)$ and $q \in \mathcal{B}(K, M)$. For the claim about q^*, we shall prove the more general inclusion:

$$\mathcal{F}(K, M) \subseteq \mathcal{Q}(K, M). \tag{27}$$

Let $p \in \mathcal{F}(K, M)$. Then there exists $u \in K$ such that $p + Mu \in \mathcal{S}(K, M)^*$. Let $v \in \mathcal{S}(K, M)$ and $x \in \operatorname{argmin}_{z \in K} z^T Mv$. Then we have

$$x^T Mv \leq u^T Mv,$$

which, by the copositivity-plus property of M on 0^+K, yields

$$v^T(p + Mx) \geq v^T(p + Mu) \geq 0.$$

This establishes (27). Hence q^* belongs to the set $\mathcal{Q}(K, M)$. That q belongs to $\mathcal{B}(K, M)$ can be proved in the following way. Let v be any nonzero vector in $\mathcal{S}(K, M)$ and $x \in \operatorname{argmin}_{z \in K} z^T Mv$. By the argument just given, we know that $q \in \mathcal{Q}(K, M)$. Suppose that $v^T(q + Mx) = 0$. Take any vector $\tilde{z} \in$ SOL(K, q, M). Then, as noted several times before, we have $q + M\tilde{z} \in \mathcal{S}(K, M)^*$. Repeating the proof of the inclusion (27) applied to $p = q$ and $u = \tilde{z}$, we obtain

$$0 = v^T(q + Mx) \geq v^T(q + M\tilde{z}) \geq 0.$$

Hence, $v^T Mx = v^T M\tilde{z}$. Since $Mv = -M^T v$, we deduce

$$x^T Mv = \tilde{z}^T Mv,$$

which implies that $\tilde{z} \in \operatorname{argmin}_{z \in K} z^T Mv$. Summarizing this discussion, we see that the pair $(\tilde{z}, v)$ violates the implication (14) which is necessary (and sufficient) for SOL(K, q, M) to be bounded. This contradiction establishes the implication (b) $\Rightarrow$ (a) under assumption (A).

We now show the implication (c) $\Rightarrow$ (b). Let $q \in \text{int } \mathcal{F}(K, M)$. When (B) holds, Theorem 1 implies that the AVI (K, q, M) is solvable. Suppose its solution set is unbounded. By Proposition 4, there exist a solution $x \in \text{SOL}(K, q, M)$ and a nonzero vector v satisfying the three properties in that proposition. By the positive semi-definiteness of M, it follows that $(M + M^T)v = 0$. The definition of q implies that for some $\varepsilon > 0$ sufficiently small, $q - \varepsilon v \in \mathcal{F}(K, M)$. Hence, there exists a vector $z \in K$ such that $q - \varepsilon v + Mz \in (0^+K)^*$ Since $v \in 0^+K$, we have

$$0 \leq v^T(q - \varepsilon v + Mz) = v^T(q + Mx) - \varepsilon\|v\|^2 - (z - x)^T Mv \leq -\varepsilon\|v\|^2,$$

which is a contradiction. This establishes the desired implication. When (A) holds, by slightly modifying the proof of the inclusion (27), we may establish

$$\text{int } \mathcal{F}(K, M) \subseteq \mathcal{B}(K, M),$$

which, in view of Corollary 5, gives (b).

The implication (c) $\Rightarrow$ (d) is an easy consequence of (10) and (3). Indeed, we have

$$\begin{aligned} \text{int } \mathcal{F}(K, M) &= \text{ri } (\mathcal{S}(K, M)^* - MK) \\ &= \text{ri } \mathcal{S}(K, M)^* - \text{ri } MK \subseteq \text{ri } \mathcal{S}(K, M)^* - MK. \end{aligned}$$

To prove the implication (d) $\Rightarrow$ (e), we first note that by (9), $-M(0^+K) \subseteq \mathcal{S}(K, M)^*$. This implies, by (4),

$$\begin{aligned} \text{ri } \mathcal{S}(K, M)^* - MK &= \text{ri } \mathcal{S}(K, M)^* - M(0^+K) - M(\mathcal{H}G) \\ &\subseteq \text{ri } \mathcal{S}(K, M)^* - M(\mathcal{H}G), \end{aligned}$$

which establishes the implication (d) $\Rightarrow$ (e). The reverse implication (e) $\Rightarrow$ (d) is obvious.

We have now completed the proof of the implications in (25). It remains to establish the two additional conclusions under the nonemptiness assumption of int $\mathcal{S}(K, M)^*$. We first demonstrate the equivalence of all the five statements by proving the implication (d) $\Rightarrow$ (c). But this is obvious from the inclusion,

$$\text{int } \mathcal{S}(K, M)^* - M(K) \subseteq \text{int } \mathcal{S}(K, M)^* - M(K)) = \text{int } \mathcal{F}(K, M),$$

which, incidentally, is easy to see. Finally, the expression (26) holds because by Corollary 5 and the inclusion (27), we have

$$\mathcal{F}(K, M) \supseteq \mathcal{R}(K, M) \supseteq \mathcal{Q}(K, M) \supseteq \mathcal{F}(K, M).$$

Hence, equality holds throughout. This completes the proof of the theorem. □

The significance of the three conditions (c), (d), and (e) can be seen as follows. The condition (c) can be interpreted as saying that the AVI (K, q, M) is "strictly feasible." Conditions (d) and (e) are very much motivated by the LCP (q, M) with a copositive matrix M for which a great deal is known about the relation between the complementary kernel $\mathcal{S}(0, M)$ and various properties of the problem [9], [10], [14]. Indeed, when $K = R^n_+$, we have $G = \{0\}$; moreover, by the argument to follow, the set int $\mathcal{S}(K, M)^*$ is nonempty, hence, statement (e) reduces to $q \in \text{int } \mathcal{S}(K, M)^*$.

The assumption that int $(\mathcal{S}(K, M)^*)$ is nonempty is essential for the validity of the reverse implication (d) $\Rightarrow$ (c) (and hence, for the equivalence of the five statements). In general, we have for any matrix M,

$$(0^+K)^* \subseteq \mathcal{S}(K, M)^*$$

which holds as a consequence of duality and the fact that $\mathcal{S}(K,M) \subseteq 0^+K$. It follows that int $(\mathcal{S}(K,M)^*)$ is nonempty if 0^+K is pointed. In turn, the latter pointedness property holds if 0^+K is contained in the nonnegative orthant R^n_+. So, for instance, if K is given as follows:

$$K = \{x \in R^n : Ax \geq b, x \geq 0\},$$

then $\mathcal{S}(K,M)^*$ must have a nonempty interior regardless of any special property the matrix M might have.

The final result of this paper identifies a sufficient condition under which the assumption (A) in Theorem 7 holds. In order to state this result, we mention an obvious property of an extreme point of a polyhedral set. Namely, if K is polyhedral, then a vector $x \in K$ is an extreme point of K if and only if the associated cone of feasible directions $\mathcal{F}_x(K)$ is pointed. The proof of this fact is easy and left to the reader. Consequently, it follows that if x is an extreme point of K, then int $(\mathcal{F}_x(K)^*)$ is nonempty. We also observe that if $0 \in K \subseteq R^n_+$, then zero must be an extreme point of K. In the following result, we postulate that the polyhedron K has an extreme point c and the matrix M is copositive on the translated set $K - c$.

COROLLARY 7. *Let K be a polyhedron in R^n and M be copositive-plus on 0^+K. Suppose that K has an extreme point c such that*

$$(y-c)^T M(y-c) \geq 0 \qquad \textit{for all } y \in K. \tag{28}$$

Then int $(\mathcal{S}(K,M)^*) \neq \emptyset$ and condition (A) *in Theorem* 7 *holds. Hence all the conclusions of this theorem are valid.*

Proof. Since $\mathcal{S}(K,M) \subseteq 0^+K \subseteq \mathcal{F}_x(K)$ for all $x \in K$, duality and the pointedness of $\mathcal{F}_c(K)$ imply that int $(\mathcal{S}(K,M)^*)$ is nonempty. Hence, it remains to verify the existence of a vector q^* satisfying the properties stipulated in condition (A) of Theorem 7. Let $q^* \in$ int $(\mathcal{F}_c(K)^*) - Mc$. Then, for all $\tilde{q}$ sufficiently close to q^*, we have $\tilde{q} + Mc \in$ int $(\mathcal{F}_c(K)^*)$. By Proposition 1, SOL$(K, \tilde{q}, M) \neq \emptyset$ for such a vector $\tilde{q}$; indeed, we have $c \in$ SOL$(K, \tilde{q}, M)$. We claim that if $\tilde{q} + Mc \in$ int $\mathcal{F}_c(K)^*$, then

$$\text{SOL}(K, \tilde{q}, M) = \{c\}.$$

Assume that $z \neq c$ is another solution of the AVI $(K, \tilde{q}, M)$. Then we have

$$0 \leq (c-z)^T(\tilde{q} + Mz) = -(z-c)^T(\tilde{q} + Mc) - (z-c)^T M(z-c) \leq -(z-c)^T(\tilde{q} + Mc)$$

where the last inequality follows from the condition (28). But since $0 \neq z - c \in \mathcal{F}_c(K)$, we must have $(z-c)^T(\tilde{q} + Mc) > 0$, which is a contradiction.

Consequently, using the same argument as in the proof of Theorem 6, we may deduce that for all vectors y with $\|y\|$ sufficiently small, the equation

$$F_{(q^*,M)}(x) = y$$

has a unique solution, namely, $c + y$. Hence, for some open neighborhood D of c, we must have deg $(F_{(q^*,M)}, D) = 1$. □

From the discussion preceding Corollary 7, we see that if $0 \in K \subseteq R^n_+$ and M is copositive on R^n_+, then $c = 0$ will satisfy the required assumption in this corollary. Alternatively, if K has an extreme point and M is copositive on $K - K$, then c can be taken to be any such extreme point. In these two cases, the conclusions of Theorem 7 are therefore valid.

REFERENCES

[1] I. Adler and D. Gale, *On the Solutions of the Positive Semi-Definite Linear Complementarity Problem,* Tech. Report ORC 75-12, Operations Research Center, University of California, Berkeley, CA, 1975.

[2] R. W. Cottle, *Solution rays for a class of complementarity problems,* Math. Programming Stud., 1 (1974), pp. 59–70.

[3] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem,* Academic Press, Boston, 1992.

[4] J. W. Daniel, *Stability of the solution of the definite quadratic programs*, Math. Programming, 5 (1973), pp. 41–53.

[5] R. D. Doverspike, *Some perturbation results for the linear complementarity problem,* Math. Programming, 23 (1982), pp. 181–192.

[6] B. C. Eaves, *The linear complementarity problem,* Management Sc., 17 (1971), pp. 612–634.

[7] M. C. Ferris and O. L. Mangasarian, *Error Bounds and Strong Upper Semicontinuity for Monotone Affine Variational Inequalities,* Tech. Report #1056, Computer Sciences Department, University of Wisconsin, Madison, WI, November, 1991.

[8] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.

[9] M. S. Gowda, *Pseudomonotone and copositive-star matrices,* Linear Algebra Appl., 113 (1989), pp. 107–110.

[10] ———, *Complementarity problems over locally compact cones,* SIAM J. Control Optimiz., 27 (1989), pp. 836–841.

[11] ———, *Applications of degree theory to linear complementarity problems,* Math. Oper. Res., (1993).

[12] M. S. Gowda and J. S. Pang, *On solution stability of the linear complementarity problem,* Math. Oper. Res., 17 (1992) pp. 77–83.

[13] ———, *Stability analysis of variational inequalities and nonlinear complementarity problems, via the mixed linear complementarity problem and degree theory,* Math. Oper. Res., (1994).

[14] M. S. Gowda and T. I. Seidman, *Generalized linear complementarity problems,* Math. Programming, 46 (1990), pp. 329–340.

[15] C. D. Ha, *Stability of the linear complementarity problem at a solution point,* Math. Programming, 31 (1985), pp. 327–338.

[16] ———, *Application of degree theory in stability of the complementarity problem,* Math. Oper. Res., 12 (1987), pp. 368–376.

[17] P. T. Harker and J. S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms, and applications,* Math. Programming 48 (1990), pp. 161–220.

[18] A. J. Hoffman, *On approximate solutions of systems of linear inequalities,* J. Nat. Bureau Standards, 49 (1952), pp. 263–265.

[19] M. Kojima, *Strongly stable stationary solutions in nonlinear programs,* in Analysis and Computation of Fixed Points, S. M. Robinson, ed. Academic Press, New York, 1980, pp. 93–138.

[20] M. Kojima, S. Mizuno, and A. Yoshise, *Ellipsoids that contain all the solutions of the positive semi-definite linear complementarity problem,* Math. Programming, 48 (1990), pp. 415–435.

[21] J. Kyparisis, *Sensitivity analysis for variational inequalities and nonlinear complementarity problems,* Ann. Oper. Res., 27 (1990), pp. 143–174.

[22] A. J. King and R. T. Rockafellar, *Sensitivity analysis for nonsmooth generalized equations,* Math. Programming, 55 (1992), pp. 193–212.

[23] N. G. Lloyd, *Degree Theory,* Cambridge University Press, Cambridge, U.K., 1978.

[24] Z. Q. Luo and P. Tseng, *Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem,* SIAM J. Optimiz., 2 (1992), pp. 43–54.

[25] O. L. Mangasarian, *A condition number for linear inequalities and linear programs,* in Methods of Operations Research, S. Bamberg and O. Opitz, eds., Verlasgruppe Athenaum / Hain / Scriptor, Hanstein, Honigstein, 1981, pp. 3–15.

[26] ———, *Characterization of bounded solutions of linear complementarity problems,* Math. Programming Stud., 19 (1982), pp. 153–166.

[27] ———, *Error bounds for nondegenerate linear complementarity problems,* Math. Programming, 48 (1990), pp. 437–445.

[28] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables,* Academic Press, New York, 1970.

[29] J. S. Pang, *A degree-theoretic approach to parametric nonsmooth equations with multivalued perturbed solution sets,* Math. Programming, (1993).

[30] Y. Qiu and T. L. Magnanti, *Sensitivity analysis for variational inequalities defined on polyhedral sets,* Math. Oper. Res., 14 (1989), pp. 410–432.

[31] Y. Qiu and T. L. Magnanti, *Sensitivity analysis for variational inequalities,* Math. Oper. Res., 17 (1992), pp. 61–76.

[32] S. M. ROBINSON, *A characterization of stability in linear programming,* Oper. Res., 25 (1977), pp. 435–477.
[33] ———, *Generalized equations and their solutions, part* I: *Basic theory,* Math. Programming Stud., 10 (1979), pp. 128–141.
[34] ———, *Strongly regular generalized equations,* Math. Oper. Res., 5 (1980), pp. 43–62.
[35] ———, *Some continuity properties of polyhedral multifunctions,* Math. Programming Stud., 14 (1981), pp. 206–214.
[36] ———, *Generalized equations and their solutions, part* II: *Applications to nonlinear programming,* Math. Programming Stud., 19 (1982), pp. 200–221.
[37] ———, *An implicit-function theorem for a class of nonsmooth functions,* Math. Oper. Res., 16 (1991), pp. 292–309.
[38] R. T. ROCKAFELLAR, *Convex Analysis,* Princeton University Press, Princeton, NJ, 1970.
[39] A. C. WILLIAMS, *Marginal values in linear programming,* J. Soc. Industrial App. Math., 11 (1963), pp. 82–94.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 442–470, March 1994

007

OPTIMAL CONTROL OF UNBOUNDED DIFFERENTIAL INCLUSIONS*

PHILIP D. LOEWEN† AND R. T. ROCKAFELLAR‡

Abstract. A Mayer problem of optimal control, whose dynamic constraint is given by a convex-valued differential inclusion, is considered. Both state and endpoint constraints are involved. Necessary conditions are proved incorporating the Hamiltonian inclusion, the Euler–Lagrange inclusion, and the Weierstrass–Pontryagin maximum condition. These results weaken the hypotheses and strengthen the conclusions of earlier works. Their main focus is to allow the admissible velocity sets to be unbounded, provided they satisfy a certain continuity hypothesis. They also sharpen the assertion of the Euler–Lagrange inclusion by replacing Clarke's subgradient of the essential Lagrangian with a subset formed by partial convexification of limiting subgradients. In cases where the velocity sets are compact, the traditional Lipschitz condition implies the continuity hypothesis mentioned above, the assumption of "integrable boundedness" is shown to be superfluous, and this refinement of the Euler–Lagrange inclusion remains a strict improvement on previous forms of this condition.

Key words. optimal control, differential inclusion, Hamiltonian inclusion, maximum principle, nonsmooth analysis

AMS subject classification. 49K24

Introduction. This paper describes necessary conditions for optimality in the following Mayer problem of optimal control: Choose an arc (i.e., an absolutely continuous function) $x\colon [a,b] \to \mathbb{R}^n$ to

$$(P)\qquad \begin{array}{ll} \text{minimize} & \ell(x(a),x(b)) \\ \text{subject to} & \dot{x}(t) \in F(t,x(t)) \quad \text{a.e. } t \in [a,b], \\ & (x(a),x(b)) \in S, \\ & x(t) \in X(t) \quad \forall t \in [a,b]. \end{array}$$

Experts will recognize the endpoint constraint $(x(a),x(b)) \in S$ and the state constraint $x(t) \in X(t)$ for all $t \in [a,b]$ as aspects of the model that are indispensable for applications, but which account for considerable complexity in the statement and derivation of necessary conditions. Clarke [2, Chap. III] gives an excellent introduction to this problem and describes several applications. Our main result can be viewed as a generalization of Clarke's necessary conditions in [2, Thm. 3.5.2]; however, the calculus described by Ioffe [6] and Rockafellar [29], and the careful Hamiltonian analysis of Loewen and Rockafellar [13], are important steps along the way from the cited result to the work at hand. The first two sections of [13] describe our reasons for choosing the formulation (P), and the relationship between this version of the problem and others current in the literature.

The results presented here improve upon those in [2] and [13] in three important ways. First, the problem is more general than any considered before, since we do not

*Received by the editors July 31, 1991; accepted for publication (in revised form) September 10, 1992. This work was supported by Natural Science and Engineering Research Council of Canada grant 5-89441, and by National Science Foundation grant DMS-8819586.

†Department of Mathematics, University of British Columbia, Vancouver, Canada, V6T 1Z2.

‡Department of Mathematics, University of Washington, Seattle, Washington 98195.

require the sets $F(t,x)$ of admissible velocities to be bounded. (We insist throughout, however, that these sets be convex.) Second, our necessary conditions are more precise than any previously published, since they involve sharper forms of the transversality condition and the Euler–Lagrange inclusion than those in [2] and [13]. Finally, our method of proof allows a simpler approach to the main result of [13], which is recovered as a corollary. We expect all of these improvements to serve in future developments of the theory.

Several sets of necessary conditions for optimal control problems without boundedness assumptions already exist in the literature. For example, Clarke proves necessary conditions analogous to the Euler–Lagrange equation for such a differential inclusion problem in [1] (see (5.1), below). Although his result does not require the velocity sets to be bounded, it does involve a Lipschitz hypothesis on the state dependence of F—an unacceptably strong condition when the velocity sets are actually unbounded.

Polovinkin and Smirnov [19], [20] prove a form of the Euler–Lagrange inclusion that is sharper than Clarke's using a truncation argument to weaken the Lipschitz hypothesis considerably. Their results also dispense with the convexity condition on the values of the multifunction F. Kaskosz and Łojasiewicz [10] consider a Mayer problem whose dynamic constraint is a controlled differential equation in which both the control sets and the resulting velocity sets are allowed to be unbounded. However, their adjoint inclusions involve Carathéodory selections of the resulting multifunction F, and are not directly comparable to those of our main theorem. (A simple connection in the bounded case is indicated by Loewen and Vinter [14].)

Furthermore, Lipschitz conditions enter [10] at several points, making direct comparison with our main result difficult. The current paper breaks new ground in presenting Hamiltonian necessary conditions for optimality in problem (P) without assuming either that the velocity sets are bounded, or that they display full Lipschitz dependence on the state. Like our previous paper [13], it asserts the Hamiltonian and Eulerian forms of the necessary conditions simultaneously.

Two simple themes underlie our approach: Truncation and strict convexity. Let us explain these ideas before pursuing the details. Suppose $\overline{x}$ solves problem (P). In the case where the optimal solution $\overline{x}$ is Lipschitzian, i.e., $\dot{\overline{x}} \in L^\infty([a,b], \mathbb{R}^n)$, we observe that for any $R > 0$, the arc $\overline{x}$ also solves the version of problem (P) in which the given multifunction F is truncated to produce the bounded multifunction $\widetilde{F}(t,x) := F(t,x) \cap \left(\dot{\overline{x}}(t) + R \operatorname{cl} \mathbb{B}\right)$. Therefore, $\overline{x}$ must fulfill the known necessary conditions for bounded differential inclusions, provided that $\widetilde{F}$ satisfies a suitable Lipschitz condition. Identifying hypotheses on F that ensure this is one of this paper's main contributions. Then, of course, there is the question of relating the necessary conditions derived using $\widetilde{F}$ to those one might expect for F. This is not trivial; §3 contains the detailed arguments. Finally, when $\overline{x}$ is absolutely continuous but not Lipschitzian, we must allow the truncation radius R to vary with time. Our presentation treats this case in parallel with the Lipschitz case.

By coordinating the hypotheses on the multifunction F with the regularity of the solution, we derive the same necessary conditions in both instances. If F is "integrably sub-Lipschitzian in the large" (see Definition 2.3(b)) at every point $(t, \overline{x}(t))$ of $\operatorname{gph} \overline{x}$, the necessary conditions are satisfied without any regularity hypothesis on $\overline{x}$; when $\overline{x}$ is known to be Lipschitzian, we require only that F be "sub-Lipschitzian" (see Definition 2.3(a)) at every point $(t, \overline{x}(t))$ of $\operatorname{gph} \overline{x}$. Strict convexity has a unifying effect on the necessary conditions of nonsmooth optimal control, as noted in our

previous work [13].

Continuing to assume that $\overline{x}$ solves (P), we note that $\overline{x}$ remains optimal for the problem $(\mathcal{P})$ in which the objective function $\ell(x(a), x(b))$ is augmented by an integral term to become

$$\ell(x(a), x(b)) + \int_a^b \left[\sqrt{1 + \left|\dot{x}(t) - \dot{\overline{x}}(t)\right|^2} - 1\right] dt.$$

Hence the Hamiltonian necessary conditions for optimality in $(\mathcal{P})$ must apply to $\overline{x}$. In the bounded case, the analysis of [13] shows that the Hamiltonian inclusion for $\overline{x}$ in problem $(\mathcal{P})$ implies the Hamiltonian inclusion, the sharpened Euler–Lagrange inclusion, and the Weierstrass–Pontryagin maximum condition we ultimately intend to assert for the original problem (P). (This analysis hinges upon the strict convexity of the integrand above as a function of the velocity variable $\dot{x}$.) To make the results of [13] applicable here, we first truncate the problem as described in the previous paragraph, and then introduce strict convexity. The small right-hand side in our transversality inclusion will surprise no one working in the field. Similar transversality conditions appeared first in the work of Mordukhovich [15], who has applied similar ideas to a range of problems in recent years (see [17]).

The new condition is obtained by replacing Clarke's normal cone and subgradient set with their (possibly nonconvex) subsets consisting of limits of proximal normals and proximal subgradients. Clarke actually uses limiting proximal normals to prove his transversality conditions in [2, Thm. 3.5.2], and his proof requires only the slightest modifications to obtain the transversality conditions used here. (This is noted explicitly in [4, Thm. 4.1, footnote].) The sharpened transversality condition also figures in Rowland and Vinter's recent work [31] on necessary conditions for controlled differential equations with free time. We take pains to incorporate it here in order that Theorem 4.3, below, can legitimately claim to have the weakest hypotheses and the strongest conclusions of any set of necessary conditions for the optimal control of differential inclusions on a fixed time interval.

The refinement of the Euler–Lagrange inclusion used here is also obtained by using the cone of limiting proximal normals in place of its convex hull (Clarke's normal cone) on the right-hand side in [2] and [13]. Some convexification is still required, but it now pertains only to the components involving derivatives of the adjoint function instead of to all components at once. A related inclusion has recently been given under considerably stronger hypotheses by Mordukhovich [18]. Our inclusion implies Mordukhovich's, and can be strictly better in certain cases. The key to our refined formulation is the introduction of strict convexity through a suitable integral cost term, as outlined above. A description of Mordukhovich's condition and a detailed comparison with ours appears in §5.

Section 1 describes the starting point for this work—the well known Hamiltonian necessary conditions of Clarke [2] as formulated by Loewen and Rockafellar [13]. It outlines the minor modifications to existing arguments required to sharpen the transversality inclusions as described above. Sections 2 and 3 concern truncation. Section 2 introduces the truncated multifunction $\widetilde{F}$ and describes hypotheses under which it satisfies a suitable Lipschitz condition, while §3 elucidates the relationship between the subgradients of the two Hamiltonians corresponding to the original and truncated multifunctions. Section 4 draws its antecedents together to produce a set of Hamiltonian necessary conditions for unbounded differential inclusion problems in Theorem 4.1. It then brings in strict convexity as outlined above. The methods of

§§2 and 3 (together with Loewen and Rockafellar [13]) then allow the simultaneous derivation of the Hamiltonian inclusion, the refined Euler–Lagrange inclusion, and the Weierstrass–Pontryagin maximum condition. This effort culminates in Theorem 4.3, the main result of this paper. Section 5 concludes the paper with a comparison between Theorem 4.3 and other published work, and gives some examples that clarify the distinctions between the various adjoint inclusions appearing here and elsewhere in the literature.

Readers interested in a quick overview of the work should observe that the notation for generalized derivatives and normals introduced in §1 differs from that in such standard works as Clarke [2]. Clarke subgradients and normals are indicated by the "barred" symbols $\overline{\partial} f(x)$ and $\overline{N}_C(x)$, while proximal subgradients and normals wear a double hat: $\hat{\hat{\partial}} f(x)$ and $\hat{\hat{N}}_C(x)$. The unadorned notation $\partial f(x)$ and $N_C(x)$ is reserved for sets of limiting proximal subgradients and limiting proximal normals.

1. Hypotheses and preliminary results. In this section we establish the technical foundation on which our later results rest. We state the hypotheses under which we later analyse the given problem (P), and review the constraint qualification we must impose when the state constraint is active along the optimal arc. We also review the necessary conditions for bounded differential inclusions due to Clarke [2, Thm. 3.5.2], and observe that they remain valid with a somewhat sharper transversality condition. Since our formulation of the state constraint differs from Clarke's, we use the form of his result appearing in our previous work [13, Thm. 2.8]. (The relationship between these two modes of presentation is clearly spelled out in [13]: While it is almost true to say that a simple change of variable makes them equivalent, the extra analysis appearing in Lemma 2.4 of [13] makes the nontriviality assertion of [13, Thm. 2.8] stronger than Clarke's.) We sharpen the transversality condition in the known result by replacing its right-hand side with a smaller set. Instead of the Clarke subgradient and normal cone, we use the limiting subgradient and the limiting normal cone. These are the fundamental objects in the theory of proximal analysis, which is described in Rockafellar [27], [29], and Clarke [2, § 2.5], for example; see also the book by Mordukhovich [17].

Proximal analysis. Consider a closed set $C \subseteq \mathbb{R}^m$ containing some point c. A vector $\zeta \in \mathbb{R}^m$ is called a *proximal normal to C at c,* written $\zeta \in \hat{\hat{N}}_C(c)$, if there is some $M > 0$ so large that

$$\langle \zeta\,,\, c' - c\rangle \le M|c' - c|^2 \quad \text{for all } c' \in C. \tag{1.1}$$

Theorem 1.2, below, refers to the cone of *limiting normals to C at c,* namely,

$$\begin{aligned} N_C(c) := \big\{\zeta \in \mathbb{R}^m \;:\; &\zeta = \lim_{k\to\infty} \zeta_k \text{ for some sequences} \\ &\zeta_k \in \hat{\hat{N}}_C(c_k) \text{ and } c_k \underset{C}{\to} c\big\}. \end{aligned} \tag{1.2}$$

(Here $c_k \underset{C}{\to} c$ means that $c_k \to c$ and $c_k \in C$ for all k.) The important properties of the limiting normal cone (easily deduced, for example, from [2, §2.5]) are:

(a) If $c \in \operatorname{bdry} C$, then $N_C(c)$ contains nonzero elements;
(b) The multifunction $c' \rightrightarrows N_C(c')$ has closed graph; and
(c) Clarke's normal cone $\overline{N}_C(c)$ is given by

$$\overline{N}_C(c) = \operatorname{cl}\operatorname{co} N_C(c)\,. \tag{1.3}$$

When the object of study is not a set but a locally Lipschitzian function $f: \mathbb{R}^m \to \mathbb{R}$, we apply the previous notions to the set $C := \operatorname{epi} f = \{(x, r) \in \mathbb{R}^m \times \mathbb{R} : r \geq f(x)\}$. This leads to the following definition: Given a point x, a vector ζ is called a *proximal subgradient of f at x*, written $\zeta \in \hat{\partial} f(x)$, if there is some $M > 0$ so large that, on some neighbourhood U of x, we have

$$f(x') \geq f(x) + \langle \zeta , x' - x \rangle - M|x' - x|^2 \quad \forall x' \in U. \tag{1.4}$$

The set of *limiting subgradients of f at x* is defined by

$$\begin{aligned} \partial f(x) = \{\zeta \in \mathbb{R}^m \, : \zeta = \lim_{k \to \infty} \zeta_k \text{ for some sequences} \\ \zeta_k \in \hat{\partial} f(x_k), \ x_k \to x\}. \end{aligned} \tag{1.5}$$

For locally Lipschitzian functions f, the set $\partial f(x)$ is nonempty and compact-valued everywhere, and the multifunction $x' \rightrightarrows \partial f(x')$ has closed graph. Moreover, Clarke's generalized gradient $\overline{\partial} f(x)$ may be obtained from the set of limiting subgradients as follows:

$$\overline{\partial} f(x) = \operatorname{co} \partial f(x). \tag{1.6}$$

(A relationship somewhat more complicated than (1.6) gives $\overline{\partial} f(x)$ in the case where f is assumed only to be lower semicontinuous and extended real valued.) Mordukhovich has used the limiting normal cone in the formulation of necessary conditions since 1976 in [15], [16], [17]. In collaboration with his student A. Y. Kruger, he has extended certain aspects of the theory to infinite-dimensional spaces [11]. More recently, Ioffe [6] has studied the limiting normal cone and limiting subgradient set described here under the names "approximate normal cone" and "approximate subdifferential," and given a more comprehensive extension to the infinite-dimensional case [7], [8], [9].

Hypotheses. Throughout the paper we confine our attention to a relatively open subset Ω of $[a, b] \times \mathbb{R}^n$ having nonempty sections

$$\emptyset \neq \Omega_t = \{x \in \mathbb{R}^n \, : \, (t, x) \in \Omega\} \quad \forall t \in [a, b].$$

In order to treat a local solution $\bar{x}$, we assume that $F(t, x)$ is empty-valued for $(t, x) \notin \Omega$. This makes the requirement that $x(t) \in \Omega_t$ for all t implicit for admissibility in problem (P). (Note that for any continuous function $x: [a, b] \to \mathbb{R}^n$ whose graph lies in Ω, a simple compactness argument implies the existence of some $\varepsilon > 0$ so small that $x(t) + \varepsilon \mathbb{B} \subseteq \Omega_t$ for all $t \in [a, b]$. Here, and throughout the paper, $\mathbb{B}$ denotes the open unit ball in $\mathbb{R}^n$.) Furthermore, we assume the following:

(H1) The endpoint cost functional ℓ is locally Lipschitz on the closed set $S_0 := (\operatorname{cl} \Omega_0) \times (\operatorname{cl} \Omega_1)$, and the localized endpoint constraint set $S \cap S_0$ is closed;

(H2) The sets $F(t, x)$ are nonempty, closed, and convex for each (t, x) in Ω;

(H3) The multifunction F is measurable with respect to the σ-field $\mathcal{L} \times \mathcal{B}$ generated by products of Lebesgue subsets of $[a, b]$ with Borel subsets of $\mathbb{R}^n$;

(H4) The state constraint multifunction X has closed values $X(t)$ and is lower semicontinuous in the sense that, for every point $(t_0, x_0) \in \Omega \cap (\operatorname{gph} X)$ and for every sequence $t_k \to t_0$ in $[a, b]$, there exists a sequence $x_k \to x_0$ satisfying $x_k \in X(t_k)$ for all k.

Jump directions. It is well known that the action of state constraints on an optimal trajectory manifests itself in the necessary conditions by producing discontinuities in

the corresponding adjoint arc. Roughly speaking, the adjoint vector is allowed to jump in an outward normal direction to the constraint set at an instant when the constraint is active. In the general setting proposed here, the possible jump directions lie in the closed convex cone defined as follows. For each (t,x) in $\Omega \cap (\operatorname{gph} X)$:

$$\overline{N}_X(t,x) = \operatorname{cl}\operatorname{co}\{\nu \in \mathbb{R}^n : \nu = \lim_{k\to\infty} \nu_k \text{ for some sequences } \nu_k \in \hat{\overline{N}}_{X(t_k)}(x_k),\ (t_k,x_k) \xrightarrow[\operatorname{gph} X]{} (t,x)\}. \tag{1.7}$$

A discussion of this cone and its relation to other formulations of the state constraint is given in §2 of our previous paper [13]. In that work the same cone was denoted by $N(t,x)$; the change of notation here is meant to emphasize the fact that this cone is related to the multifunction X and that, like Clarke's normal cone, it has closed convex values.

Hypothesis (H4) and definition (1.7) together imply that for any continuous function $x\colon [a,b] \to \mathbb{R}^n$ satisfying $x(t) \in X(t) \cap \Omega_t$ for all t, the convex cone valued multifunction $t \rightrightarrows \overline{N}_X(t,x(t))$ is Borel measurable. In this case, to call an $\mathbb{R}^n$-valued measure μ "$\overline{N}_X(t,x(t))$-valued" means that μ is absolutely continuous with respect to some nonnegative measure μ_0 on $[a,b]$, and that some measurable selection $\nu(t) \in \overline{N}_X(t,x(t))$ satisfies $d\mu(t) \equiv \nu(t)d\mu_0(t)$. (See Rockafellar [22, §5].) Necessary conditions for optimality in which the adjoint function is merely of bounded variation, with jump directions described in terms of cone-valued measures, were first given for convex problems of Bolza by Rockafellar [23], [24], [26].

Constraint qualification. Our necessary conditions require that the cone $\overline{N}_X(t,\overline{x}(t))$ be pointed everywhere on the graph of the optimal arc $\overline{x}$. This constraint qualification is also essential in Clarke's formulation (see [2, Remark 3.2.7(iii)]), as explained by Loewen and Rockafellar [13, §2]. Let us call the state constraint "active" (relative to $\overline{x}$) at any time t when $(t,\overline{x}(t))$ lies on the boundary of $\operatorname{gph} X$, and "inactive" when $(t,\overline{x}(t))$ lies in the interior of $\operatorname{gph} X$. It follows easily from (1.7) that $\overline{N}_X(t,\overline{x}(t))$ collapses to the trivial cone $\{0\}$ if and only if the state constraint is inactive at time t. In particular, if the state constraint is inactive for all $t \in [a,b]$—perhaps because $X(t) \equiv \mathbb{R}^n$—then the constraint qualification mentioned above holds automatically. (Note that there can be times when the state constraint is active even though $\overline{x}(t) \in \operatorname{int} X(t)$ for all t. An example is provided by the arc $\overline{x}(t) = 2t$ and the multifunction $X(t) = \{y : |y| \geq t\}$: The state constraint is active at $t = 0$ even though $\overline{x}(t) \in \operatorname{int} X(t)$ for all t.) Another common case in which the constraint qualification holds automatically arises when the state constraint sets $X(t)$ are convex and have nonempty interior; then the cone $\overline{N}_X(t,x)$ coincides with the usual normal cone $N_{X(t)}(x)$ of convex analysis, and the latter cone is pointed if and only if $\operatorname{int} X(t) \neq \emptyset$.

The state constraint we impose can be given a simple geometric interpretation, based on Rockafellar [25, Thm. 3]. That result states that if a closed subset Ξ of $\mathbb{R}^n$ contains a point ξ at which the Clarke normal cone $\overline{N}_\Xi(\xi)$ is pointed, then there is a neighborhood of ξ in which Ξ is indistinguishable from the isometric linear image of the epigraph of some Lipschitz function on $\mathbb{R}^{n-1}$. (The set Ξ is then called *epi-Lipschitzian at* ξ.) The set $\overline{N}_X(t,x)$ defined by (1.7) is generally larger than Clarke's normal cone $\overline{N}_{X(t)}(x)$ by (1.2)–(1.3), since it contains information not only about the shape of the set $X(t)$, but also about its behavior as t varies.

This leads to the following result, which makes precise the sense in which we can regard our constraint qualification as a requirement of "uniform epi-Lipschitzian behavior" of the multifunction X.

PROPOSITION 1.1. *Let (t,x) be a point in $\Omega \cap \operatorname{gph} X$ at which the cone $\overline{N}_X(t,x)$ is pointed. Then there exist a neighborhood U of (t,x) in $[a,b]\times\mathbb{R}^n$, a linear isometry A on $\mathbb{R}^n$, and a constant L with the following properties. For any $(s,y)\in U\cap\operatorname{gph} X$, there is a Lipschitz function $\phi_{(s,y)}$ of rank L having the property that*

$$A\left(\operatorname{epi}\phi_{(s,y)}\right) = X(s) \quad \text{near } y.$$

(In detail, this conclusion means that there is a neighborhood V of y such that $A\left(\operatorname{epi}\phi_{(s,y)}\right)\cap V = X(s)\cap V$.)

Proof. This result follows from a careful quantitative analysis of the cited theorem of Rockafellar. Details are available in [12]; here we merely indicate the main steps in the proof.

For any $\varepsilon\in(0,1)$ and any unit vector $v\in\mathbb{R}^n$, define the closed, pointed convex cone

$$K_\varepsilon(v) := \{\zeta\in\mathbb{R}^n \,:\, \langle\zeta\,,\,v\rangle \geq \varepsilon|\zeta|\}.$$

Deduce from the hypothesis that there exist some $\varepsilon\in(0,1)$ and some v of unit length, together with a neighborhood U_0 of (t,x) relative to Ω, such that

$$(*) \qquad \overline{N}_{X(s)}(y) \subseteq K_\varepsilon(v) \quad \forall (s,y)\in U_0.$$

Let A be any linear isometry of $\mathbb{R}^n$ into $\mathbb{R}^{n-1}\times\mathbb{R}$ such that $Av = (0,-1)$. (One certainly exists.) Then, taking polars in $(*)$ gives

$$\overline{T}_{A(X(s))}(Ay) \supseteq K_{\varepsilon'}(0,1) \quad \forall (s,y)\in U_0,$$

where $\varepsilon' = \sqrt{1-\varepsilon^2}$. For each (s,y) in U_0, Rockafellar's proof of [25, Thm. 3] provides a Lipschitzian function $\phi_{(s,y)}$ on $\mathbb{R}^{n-1}$ whose epigraph coincides with $A(X(s))$ throughout some neighborhood of Ay. The Lipschitz rank of $\phi_{(s,y)}$ can be estimated using the bound on the size of Clarke's generalized gradient of $\phi_{(s,y)}$ implicit in the identification with $A(X(s))$ and $\operatorname{epi}\phi_{(s,y)}$. (In fact, the estimate gives $L = \varepsilon/\sqrt{1-\varepsilon^2}$.) The conclusion of the proposition now follows, but we have interchanged the names of A and A^{-1} for clarity. □

Note that we have little control over the time-dependence of the functions $\phi_{(s,y)}$ in Proposition 1.1. For example, the multifunction

$$X(t) := \{(x,y) \,:\, y\geq 0\} \quad \text{if } t<\tfrac{1}{2}, \qquad X(t) := \{(x,y) \,:\, y\geq 1\} \quad \text{if } t\geq\tfrac{1}{2}$$

satisfies all our hypotheses but has a discontinuity at $t=\frac{1}{2}$. Since X is convex-valued, $\overline{N}_X(\frac{1}{2},(0,1)) = N_{X(\frac{1}{2})}(0,1) = \{0\}\times(-\infty,0]$. This cone is clearly pointed. When $s<\frac{1}{2}$ and y is near 1, the set $X(s)$ near y looks like the epigraph of $\phi_{(s,y)}\equiv 0$. However, when $s\geq\frac{1}{2}$ and y is near 1, the set $X(s)$ near y looks like the epigraph of the function $\phi_{(s,y)}\equiv 1$.

Hamiltonian necessary conditions. We are now in a position to state the necessary conditions for bounded differential inclusions on which our main results are based. These involve the *Hamiltonian* associated with the multifunction F, defined by $H(t,x,p) := \sup\{\langle p\,,\,v\rangle \,:\, v\in F(t,x)\}$. Theorem 1.2, below, is essentially a transcription of [13, Thm. 2.8], with the exception that the transversality inclusion in part (b) involves limiting subgradients and normals instead of the Clarke subgradients and normals used in [13]. (Clarke subgradients are still required in the Hamiltonian

inclusion.) This distinction is immaterial in the smooth and convex cases for which Clarke's notions are indistinguishable from the corresponding limiting constructions. In general, however, it is possible that the right-hand side of (b) is a proper subset of its counterpart in [13]. A detailed proof of Theorem 1.2 would be both long and repetitive, since many of the steps in the argument are now (or should be) well known. For this reason, we simply outline a derivation of the result based on small adjustments to proofs in the literature. Theorem 4.1, below, will significantly weaken the boundedness and Lipschitz continuity hypotheses (i) and (ii) in the following statement.

THEOREM 1.2. *Assume* (H1)–(H4). *Suppose the arc* $\overline{x}$ *solves problem* (P), *and that the constraint qualification below is satisfied:*

(CQ) *The cone* $\overline{N}_X(t,\overline{x}(t))$ *is pointed for all* t *in* $[a,b]$.

Suppose further that there exist integrable functions ϕ *and* k *such that*

(i) $F(t,x) \subseteq \phi(t)\,\mathrm{cl}\,\mathbb{B}$ *for all* $(t,x) \in \Omega$;

(ii) $F(t,y) \subseteq F(t,x) + k(t)|y-x|\,\mathrm{cl}\,\mathbb{B}$ *for all* $t \in [a,b]$, $x,y \in \Omega_t$.

Then there exist a scalar $\lambda \in \{0,1\}$ *and a function* $p \in BV([a,b];\mathbb{R}^n)$, *not both zero, together with an integrable selection* $\nu(t) \in \overline{N}_X(t,\overline{x}(t))$ *for all* $t \in [a,b]$, *such that*

(a) $(-\dot{p}(t)+\nu(t),\dot{\overline{x}}(t)) \in \overline{\partial}H(t,\overline{x}(t),p(t))$ *for almost all* $t \in [a,b]$;

(b) $(p(a),-p(b)) \in \lambda\partial\ell(\overline{x}(a),\overline{x}(b)) + N_S\,(\overline{x}(a),\overline{x}(b))$;

(c) *The singular part of the measure* dp *is* $\overline{N}_X(t,\overline{x}(t))$*-valued, and in particular is supported on the set*

$$\{t : \overline{N}_X(t,\overline{x}(t)) \neq \{0\}\} = \{t \in [a,b] : (t,\overline{x}(t)) \in \mathrm{bdry\,gph}\,X\}.$$

Outline of proof. Our first step is to reconsider Clarke's necessary conditions for optimality in [2, Thm. 3.5.2], noting that they apply to a slightly different problem than ours. We claim that these remain valid when the transversality condition at the final time [2, p. 143, (2)] is written as $-E \in N_{C_1}(x(b))$, instead of $-E \in \overline{N}_{C_1}(x(b))$. (That is, with the limiting normal cone in place of Clarke's normal cone.) To justify this, we must review the proof of [2, Thm. 3.4.3].

In the notation used there, choosing $x' = x$ in line 4 of Lemma 2 shows immediately that $-v$ is a proximal normal ("perpendicular" in [2]) to the set C_1 at the point c; hence the third displayed conclusion of Lemma 2 can be replaced by

$$\lambda\beta\zeta + p(b) + \int_{[a,b]} \gamma(s)\,\mu(ds) \in -\hat{\hat{N}}_{C_1}(x(b)-u).$$

In the limiting analysis of Step 4, this relation becomes

$$\lambda\beta_0\zeta + \widetilde{p}(b) + \int_{[a,b]} \gamma(s)\,\widetilde{\mu}(ds) = \lambda v_0 \in -N_{C_1}(x(b)).$$

The proof of [2, Thm. 3.4.3] concludes as before, and the method used to make the given solution unique, employed in the proof of [2, Thm. 3.5.2], respects the refined formulation of the transversality condition.

Our second step is to extend the necessary conditions described above to cope with an endpoint cost functional and endpoint constraint set involving both $x(a)$ and

$x(b)$ jointly. To do so, simply note that if an arc $\overline{x}$ solves (P), then the extended arc $(\overline{r}, \overline{x})$ with $\overline{r}(t) \equiv \overline{x}(a)$ solves the following problem:

$$\begin{aligned} \text{minimize} \quad & \ell(r(b), x(b)) \\ \text{subject to} \quad & \dot{r}(t) = 0, \qquad \dot{x}(t) \in F(t, x(t)) \quad \text{a.e. } t \in [a, b], \\ & (r(a), x(a)) \in D, \qquad (r(b), x(b)) \in S, \\ & (r(t), x(t)) \in \mathbb{R}^n \times X(t) \quad \forall t \in [a, b], \end{aligned}$$

where $D = \{(z, z) : z \in \mathbb{R}^n\}$ is the diagonal of $\mathbb{R}^n \times \mathbb{R}^n$. This is a situation to which our refinement of [2, Thm. 3.5.2] can be applied, and the resulting transversality condition is

$$(\dagger) \qquad (p(a), -p(b)) \in \lambda \overline{\partial} \ell(\overline{x}(a), \overline{x}(b)) + N_S(\overline{x}(a), \overline{x}(b)).$$

This differs from the desired conclusion (b) only in its use of the Clarke subgradient of the endpoint cost function ℓ.

Passing from the parametric form of the state constraint $g(t, x(t)) \leq 0$ to the intrinsic form $x(t) \in X(t)$ as described in §2 of [13] does not affect the transversality inclusion. The same methods, therefore, imply that [13, Thm. 2.8] remains valid with the transversality inclusion replaced by (†)—and in particular that conclusion (b) holds if ℓ is smooth.

We now turn to Theorem 1.2. Suppose $\overline{x}$ solves problem (P). Then the extended arc $(\overline{x}, \overline{z})$ in which $\overline{z}(t) \equiv \ell(\overline{x}(a), \overline{x}(b))$ must solve the following problem:

$$\begin{aligned} \text{minimize} \quad & z(b) \\ \text{subject to} \quad & \dot{x}(t) \in F(t, x(t)), \qquad \dot{z}(t) = 0 \quad \text{a.e. } t \in [a, b], \\ & (x(a), x(b), z(b)) \in \operatorname{epi}(\ell + \Psi_S), \qquad z(a) \in \mathbb{R}, \\ & (x(t), z(t)) \in X(t) \times \mathbb{R} \quad \forall t \in [a, b]. \end{aligned}$$

(Here Ψ_S denotes the *indicator function* of the set S, defined by setting $\Psi_S(x) = 0$ if x lies in S, and $\Psi_S(x) = +\infty$ otherwise.) Applying the intermediate form of [13, Thm. 2.8] described above leads to the conclusion that for some $\lambda \in \{0, 1\}$ and $p \in BV([a, b]; \mathbb{R}^n)$, not both zero, and some selection $\nu(t) \in \overline{N}_X(t, \overline{x}(t))$, we have the desired conclusions (a) and (c) of the current theorem, along with the transversality condition

$$(*) \qquad (p(a), -p(b), -\lambda) \in N_{\operatorname{epi}(\ell + \Psi_S)}(\overline{x}(a), \overline{x}(b), \ell(\overline{x}(a), \overline{x}(b))).$$

If $\lambda > 0$, this assertion is equivalent to

$$(p(a), -p(b)) \in \lambda \partial(\ell + \Psi_S)(\overline{x}(a), \overline{x}(b)).$$

Thanks to the calculus rules for limiting subgradients in [6, Thm. 4] of Ioffe, we deduce that

$$(\ddagger) \qquad (p(a), -p(b)) \in \lambda \partial \ell(\overline{x}(a), \overline{x}(b)) + N_S(\overline{x}(a), \overline{x}(b)).$$

In particular, when $\lambda = 1$ we obtain the desired conclusion (b). When $\lambda = 0$, the proximal subgradient formula (see Rockafellar [27, Proof of Thm. 1]) asserts that the vector $(p(a), -p(b), 0)$ appearing on the left-hand side of $(*)$ can be expressed as a limit of some sequence $(p_k(a), -p_k(b), -\lambda_k)$ with $\lambda_k > 0$, along which $(*)$ holds relative to a sequence of base points $(x_k(a), x_k(b), \ell(x_k(a), x_k(b)))$ converging to $(\overline{x}(a), \overline{x}(b), \ell(\overline{x}(a), \overline{x}(b)))$. These sequences therefore satisfy an analogue of $(\ddagger)$ in which a subscript k appears throughout. Taking the limit as $k \to \infty$, we obtain $(\ddagger)$ with $\lambda = 0$. Thus conclusion (b) is also valid in this case. □

Remark. The transversality condition $(*)$ from which (b) is derived in the foregoing proof could conceivably be sharper than (b) in some cases, since it involves the subgradients of the essential endpoint cost functional $\ell + \Psi_S$, instead of the sum of subgradients of its two terms.

2. Localization of unbounded multifunctions. Suppose the arc $\overline{x}$ solves problem (P). Then $\overline{x}$ must also solve any problem with the same objective function as (P) but fewer admissible arcs. Such a problem can be described by replacing the given velocity sets $F(t,x)$ by their bounded subsets

$$\widetilde{F}(t,x) := F(t,x) \cap \big(\dot{\overline{x}}(t) + R(t)\,\mathrm{cl}\,\mathbb{B}\big) \tag{2.1}$$

for some real-valued function $R(t)$. Known necessary conditions for differential inclusion problems with compact right-hand sides, like Theorem 1.2, above, then provide some information about $\overline{x}$. Our goal is to translate this information into necessary conditions that refer only to the data of the original problem (P). This translation is not completely straightforward; neither is it obvious which hypotheses on F and which choice of $R(t)$ will make the application of Theorem 1.2 to the reduced problem both legitimate and informative.

This section deals with the hypotheses. It is rather obvious that for any nonnegative integrable function $R(t)$, the truncated multifunction $\widetilde{F}$ defined above has compact convex values satisfying Theorem 1.2(i). We ensure nonemptiness and Lipschitz continuity by way of the following lemma, whose uncluttered notation is intended to clarify the essential geometry.

LEMMA 2.1. *Let $\overline{v} \in \mathbb{R}^n$, and let F_1 and F_2 be two subsets of $\mathbb{R}^n$ such that for some $\delta > 0$ and $0 < r < R$,*

(i) $F_2 \cap (\overline{v} + R\,\mathrm{cl}\,\mathbb{B}) \subseteq F_1 + \delta\,\mathrm{cl}\,\mathbb{B}$;

(ii) $F_1 \cap (\overline{v} + r\,\mathrm{cl}\,\mathbb{B}) \neq \emptyset$;

(iii) *F_1 is convex.*

Then $F_2 \cap (\overline{v} + R\,\mathrm{cl}\,\mathbb{B}) \subseteq F_1 \cap (\overline{v} + R\,\mathrm{cl}\,\mathbb{B}) + (2R\delta/(R-r))\,\mathrm{cl}\,\mathbb{B}$.

Proof. Without loss of generality, take $\overline{v} = 0$. Choose any $v_2 \in F_2 \cap R\,\mathrm{cl}\,\mathbb{B}$. According to (i), there exists $v_1 \in F_1$ such that $|v_2 - v_1| \leq \delta$. By (ii), we also have some $v_0 \in F_1$ such that $|v_0| \leq r$. Hypothesis (iii) ensures that

$$v_t := (1-t)v_0 + tv_1 \in F_1 \quad \forall t \in [a,b].$$

We estimate

$$\begin{aligned} |v_t| &\leq (1-t)|v_0| + t|v_1| \\ &\leq (1-t)r + t\,(\delta + |v_2|) \\ &\leq r + t(\delta + R - r). \end{aligned}$$

This implies that $v_t \in F_1 \cap R\,\mathrm{cl}\,\mathbb{B}$ when $r + t(\delta + R - r) \leq R$, in particular when

$$0 \leq t \leq \widehat{t} := \frac{R-r}{\delta + R - r}.$$

Let $\widehat{v} = v_{\widehat{t}}$. Then $\widehat{v} \in F_1 \cap (\overline{v} + R\,\mathrm{cl}\,\mathbb{B})$, and

$$\begin{aligned} |v_2 - \widehat{v}| &\leq |v_2 - v_1| + |v_1 - \widehat{v}| \\ &\leq \delta + (1 - \widehat{t})|v_0 - v_1| \\ &\leq \delta + (1 - \widehat{t})\,(|v_0| + |v_1|) \\ &\leq \delta + (1 - \widehat{t})(r + R + \delta) \\ &= 2\delta\left[1 + \frac{r}{\delta + R - r}\right]. \end{aligned}$$

The right-hand side increases if we discard the δ appearing in the denominator. This yields $|v_2 - \widehat{v}| \leq 2\delta R/(R - r)$. Since $v_2 \in F_2$ is arbitrary and $\widehat{v} \in F_1 \cap (\overline{v} + R\,\mathrm{cl}\,\mathbb{B})$, the desired inclusion follows. □

Remark. Clarke's lemma [2, Lem. 3, p. 172] is proven by a very similar argument, but starts with a stronger hypothesis.

Using Lemma 2.1, we now provide a set of sufficient conditions for our localization technique to produce a multifunction satisfying the hypotheses of Theorem 1.2.

PROPOSITION 2.2. *Let Ω and F be given as in the formulation of problem (P); assume* (H2)–(H3). *Let $\overline{x}$ be an F-trajectory. Suppose there exists $\overline{\varepsilon} > 0$ together with nonnegative integrable functions m and R such that $m/R \in L^\infty[a,b]$, and for almost every $t \in [a,b]$ we have*

$$F(t,y) \cap \left(\dot{\overline{x}}(t) + R(t)\,\mathrm{cl}\,\mathbb{B}\right) \subseteq F(t,x) + m(t)|y - x|\,\mathrm{cl}\,\mathbb{B} \quad \forall x, y \in \overline{x}(t) + \overline{\varepsilon}\mathbb{B}. \tag{2.2}$$

Then there is a relatively open subset $\widetilde{\Omega}$ of $[a,b] \times \mathbb{R}^n$ containing the graph of $\overline{x}$ on which the truncated multifunction $\widetilde{F}(t,x)$ of (2.1) *satisfies not only* (H2)–(H3), *but also hypotheses* (i)–(ii) *of Theorem* 1.2.

Proof. Note that the requirement that $\overline{x}(t) + \overline{\varepsilon}\mathbb{B} \subseteq \Omega_t$ for all t is implicit in hypothesis (2.2), since the choice $y = \overline{x}(t)$ forces $F(t,x) \neq \emptyset$ for all $x \in \overline{x}(t) + \overline{\varepsilon}\mathbb{B}$. Therefore any choice of $\varepsilon \in (0, \overline{\varepsilon}]$ will ensure that $\mathrm{gph}\,\overline{x} \subseteq \widetilde{\Omega} \subseteq \Omega$ for the set

$$\widetilde{\Omega} := \{(t,x) \,:\, t \in [a,b], |x - \overline{x}(t)| < \varepsilon\}.$$

We therefore fix $\varepsilon \in (0, \overline{\varepsilon}]$, taking care that

$$\varepsilon m(t)/R(t) \leq 1/2 \quad \text{a.e. } t \in [a,b]. \tag{$*$}$$

This is possible because m/R is essentially bounded by hypothesis.

Let us fix a time $t \in [a,b]$ at which $(*)$ and (2.2) hold, $\dot{\overline{x}}(t)$ exists, and $\dot{\overline{x}}(t) \in F(t, \overline{x}(t))$. (Such t-values form a subset of $[a,b]$ with full measure.) The sets $\widetilde{F}(t,x)$ are evidently compact and convex valued for each $x \in \overline{x}(t) + \varepsilon\mathbb{B}$. To observe that they are nonempty, we choose $y = \overline{x}(t)$ in (2.2). Then

$$\dot{\overline{x}}(t) \in F(t, \overline{x}(t)) \cap \left(\dot{\overline{x}}(t) + R(t)\,\mathrm{cl}\,\mathbb{B}\right) \subseteq F(t,x) + m(t)|\overline{x}(t) - x|\,\mathrm{cl}\,\mathbb{B} \quad \forall x \in \overline{x}(t) + \varepsilon\mathbb{B}.$$

This inclusion implies that

$$F(t,x) \cap \left(\dot{\overline{x}}(t) + \varepsilon m(t)\,\mathrm{cl}\,\mathbb{B}\right) \neq \emptyset \quad \forall x \in \overline{x}(t) + \varepsilon\mathbb{B}, \tag{$\dagger$}$$

and $\widetilde{F}(t,x)$ contains the left-hand side of $(\dagger)$ due to $(*)$. Thus (H2) holds relative to $\widetilde{\Omega}$; the measurability property required by (H3) is evident.

The choice $\phi(t) := |\dot{\bar{x}}(t)| + R(t)$ clearly proves condition (i) of Theorem 1.2. Only condition (ii) remains. With t fixed as above, choose any $x, y \in \bar{x}(t) + \varepsilon\mathbb{B}$ and let $F_1 = F(t,x)$, $F_2 = F(t,y)$. Then the hypotheses of Lemma 2.1 hold with $\bar{v} = \dot{\bar{x}}(t)$, $R = R(t)$, $\delta = m(t)|y-x|$ from (2.2), and $r = \varepsilon m(t) \leq R(t)/2$ from (†). The conclusion is that $\widetilde{F}(t,y) \subseteq \widetilde{F}(t,x) + k(t)|y-x|\,\mathrm{cl}\,\mathbb{B}$, where

$$k(t) = \frac{2R(t)m(t)}{R(t) - \varepsilon m(t)} = \frac{2m(t)}{1 - \varepsilon m(t)/R(t)} \leq 4m(t).$$

Hypothesis (ii) of Theorem 1.2 requires that the function k be integrable; this is ensured by the integrability of m. □

The central hypothesis (2.2) of Proposition 2.2 is a quantitative version of Aubin's *pseudo-Lipschitzian continuity* for the multifunctions $F(t,\cdot)$ at the points $(\bar{x}(t), \dot{\bar{x}}(t))$ along the trajectory $\bar{x}$ (see Rockafellar [28]). Although the conditions of Proposition 2.2 are sufficient for the development of our theory, they require that the arc $\bar{x}$ be known in advance, and offer few suggestions about effective choices of the functions R and m.

Before continuing the development of our argument, we pause to describe hypotheses on the multifunction F that can be used to verify (2.2) along any admissible arc. These involve the following concepts.

DEFINITION 2.3. Let $\Gamma\colon \Omega \rightrightarrows \mathbb{R}^m$ be a multifunction with closed values, and suppose Γ is $\mathcal{L} \times \mathcal{B}$ measurable on Ω. Consider a point $(\bar{t}, \bar{x})$ in Ω.

(a) The multifunction Γ is called *sub-Lipschitzian at* $(\bar{t}, \bar{x})$ if, for every constant $\rho \geq 0$, there exist constants $\varepsilon > 0$ and $\alpha \geq 0$ such that

$$\Gamma(t,y) \cap \rho\,\mathrm{cl}\,\mathbb{B} \subseteq \Gamma(t,x) + \alpha|y-x|\,\mathrm{cl}\,\mathbb{B} \tag{2.3}$$

for all $t \in (\bar{t} - \varepsilon, \bar{t} + \varepsilon) \cap [a,b]$ and all x, y in $\bar{x} + \varepsilon\mathbb{B}$.

(b) The multifunction Γ is called *integrably sub-Lipschitzian in the large at* $(\bar{t}, \bar{x})$ if there exist constants $\varepsilon > 0$ and $\beta \geq 0$, together with a nonnegative function α integrable on $(\bar{t} - \varepsilon, \bar{t} + \varepsilon)$, such that

$$\Gamma(t,y) \cap \rho\,\mathrm{cl}\,\mathbb{B} \subseteq \Gamma(t,x) + (\alpha(t) + \beta\rho)|y-x|\,\mathrm{cl}\,\mathbb{B} \tag{2.4}$$

for all $t \in (\bar{t} - \varepsilon, \bar{t} + \varepsilon) \cap [a,b]$, all x, y in $\bar{x} + \varepsilon\mathbb{B}$, and all $\rho \geq 0$.

Definition 2.3(a) is very similar to the notion of sub-Lipschitzian behaviour introduced by Rockafellar [28], the only difference being that here we consider multifunctions with explicit time-dependence, and require a certain uniformity of the parameters ε and α with respect to t. Rockafellar [28] offers a detailed discussion of (autonomous) sub-Lipschitzian multifunctions and the relationship between this property and the pseudo-Lipschitz continuity introduced by Aubin; he also describes several classes of sub-Lipschitzian multifunctions.

Definition 2.3(b) introduces a new type of sub-Lipschitzian assumption even in the autonomous case. It looks like a stricter hypothesis than that of Definition 2.3(a) because it places certain restrictions on the growth of the right-hand side with ρ. If (b) holds for a constant function $\alpha(t) \equiv \alpha$, then certainly (a) follows; it is not obvious that (b) always implies (a), however, since (b) allows α to depend on t, whereas α must be constant in (a).

Each of these hypotheses has a role as a sufficient condition for the applicability of Proposition 2.2.

PROPOSITION 2.4. *Let Ω and F be given as in the formulation of problem (P); assume* (H2)–(H3). *Let $\overline{x}$ be an F-trajectory. Under either of the two hypotheses below, all the conditions of Proposition* 2.2 *are met. In particular, there is a relatively open subset $\widetilde{\Omega}$ of Ω containing the graph of $\overline{x}$ on which the truncated multifunction $\widetilde{F}$ defined by* (2.1) *satisfies all the hypotheses of Theorem* 1.2.

(a) *The arc $\overline{x}$ is Lipschitzian, and the multifunction F is sub-Lipschitzian at every point $(t, \overline{x}(t))$ in* $\operatorname{gph} \overline{x}$.

(b) *The multifunction F is integrably sub-Lipschitzian in the large at every point $(t, \overline{x}(t))$ in* $\operatorname{gph} \overline{x}$.

Remarks. 1. Note that the two parts of Proposition 2.4 correspond exactly to the two parts of Definition 2.3. Part (b) imposes apparently stricter conditions on F and applies to any arc $\overline{x}$, while part (a) imposes apparently weaker requirements on F but pertains only to Lipschitzian arcs $\overline{x}$.

2. The proof of part (a) below allows for an arbitrarily small positive constant value of R in the localization of (2.1). This may eventually link our results with the necessary conditions for "weak local minima" in the calculus of variations.

3. The conditions of the Proposition make explicit reference to the arc $\overline{x}$, but they would obviously follow from corresponding hypotheses regarding sub-Lipschitzian behavior of F throughout the set Ω.

4. Proposition 2.4 remains valid when Definition 2.3 is weakened by replacing the phrase "$\rho \geq 0$" with "$\rho \geq 0$ sufficiently large" in parts (a) and (b).

Proof of Proposition 2.4. (b) Both hypotheses in the statement of the Proposition must first be extended to the whole interval $[a, b]$ by a compactness argument. We illustrate this just once, taking the more delicate case, situation (b).

Applying Definition 2.3(b) to a point $(s, \overline{x}(s))$ in $\operatorname{gph} \overline{x}$ yields constants $\varepsilon_s > 0$, $\beta_s \geq 0$, and a nonnegative function $\alpha_s(t)$ integrable on $(s - \varepsilon_s, s + \varepsilon_s)$ such that (2.4) holds for any $\rho \geq 0$ and any triple (t, x, y) chosen from the set

$$G_s = \{(t, x, y) \,:\, |t - s| < \varepsilon_s,\ x \in \overline{x}(s) + \varepsilon_s \mathbb{B},\ y \in \overline{x}(s) + \varepsilon_s \mathbb{B}\}.$$

Now each set G_s, $s \in [a, b]$, is open, and the family of these sets covers the compact set $\{(t, \overline{x}(t), \overline{x}(t)) \,:\, t \in [a, b]\}$. Therefore we may extract a finite subcover indexed by $s_1, \ldots, s_N$, and define

$$G := \bigcup_{j=1}^{N} G_j.$$

For simplicity, we have written G_{s_j} as G_j. We define ε_j, β_j, and $\alpha_j(t)$ similarly.

Observe that there exists $\bar{\varepsilon} > 0$ so small that for every $t \in [a, b]$,

$$\{t\} \times (\overline{x}(t) + \bar{\varepsilon}\mathbb{B}) \times (\overline{x}(t) + \bar{\varepsilon}\mathbb{B}) \subseteq G.$$

(If this were not true, then there would be a sequence of points outside G converging to some point $(\bar{t}, \overline{x}(\bar{t}), \overline{x}(\bar{t}))$ in the interior of G, which is a contradiction.) Next, choose $\beta = \max\{\beta_1, \ldots, \beta_N\}$ and define

$$\alpha(t) := \max_{j=1,\ldots,N} \{\alpha_j(t) \,:\, t \in (s_j - \varepsilon_j, s_j + \varepsilon_j)\}.$$

Clearly β is finite and $\alpha(t)$ is integrable. Moreover, for any pair of points (t, x) and (t, y) chosen from the set $\widehat{\Omega} = \{(t, x) \,:\, t \in [a, b],\ x \in \overline{x}(t) + \bar{\varepsilon}\mathbb{B}\}$, we have $(t, x, y) \in G$, so $(t, x, y) \in G_j$ for some $j = 1, \ldots, N$. Thus (2.4) holds for (t, x, y) with parameters

β_j and $\alpha_j(t)$, and therefore it holds with the larger parameters β and $\alpha(t)$. This shows that the set $\widehat{\Omega}$ is a relatively open subset of Ω containing the graph of $\overline{x}$ throughout which (2.4) holds uniformly with parameters $\overline{\varepsilon}$, β, and $\alpha(t)$.

Now consider the function $R(t) = 1 + \alpha(t) + |\dot{\overline{x}}(t)|$. For any t in $[a,b]$, choose $\rho = |\dot{\overline{x}}(t)| + R(t)$ in the extension of (2.4) just proved. This leads to the following estimate, valid for all x, y in $\overline{x}(t) + \overline{\varepsilon}\mathbb{B}$:

$$\begin{aligned} F(t,y) \cap \big(\dot{\overline{x}}(t) + R(t)\,\mathrm{cl}\,\mathbb{B}\big) &\subseteq F(t,y) \cap \big(|\dot{\overline{x}}(t)| + R(t)\big)\,\mathrm{cl}\,\mathbb{B} \\ &\subseteq F(t,x) + \big(\alpha(t) + \beta\,[|\dot{\overline{x}}(t)| + R(t)]\big)\,|y-x|\,\mathrm{cl}\,\mathbb{B}. \end{aligned}$$

This confirms (2.2), where the function

$$m(t) = \alpha(t) + \beta\,[|\dot{\overline{x}}(t)| + R(t)] = (1+\beta)\alpha(t) + 2\beta|\dot{\overline{x}}(t)| + \beta$$

is clearly integrable, as is $R(t)$, while $m(t)/R(t) \le 2 + 2\beta$ almost everywhere. All the hypotheses of Proposition 2.2 are in place; part (b) of the desired result follows.

(a) Under hypothesis (a), we fix any $R > 0$ (perhaps quite small) and let $\rho = R + \|\dot{\overline{x}}\|_\infty$. Then a compactness argument very similar to the one described in detail above leads to a pair of constants $\overline{\varepsilon} > 0$ and $\alpha \ge 0$ for which (2.3) holds for any pair of points (t,x) and (t,y) in $\widehat{\Omega} := \{(t,x) : t \in [a,b], |x - \overline{x}(t)| < \overline{\varepsilon}\}$. In particular, since $\rho \ge R + |\dot{\overline{x}}(t)|$, we have

$$\begin{aligned} F(t,y) \cap \big(\dot{\overline{x}}(t) + R\,\mathrm{cl}\,\mathbb{B}\big) &\subseteq F(t,y) \cap \rho\,\mathrm{cl}\,\mathbb{B} \\ &\subseteq F(t,x) + \alpha|y-x|\,\mathrm{cl}\,\mathbb{B}. \end{aligned}$$

Thus (2.2) holds with the constants $m = \alpha$ and R identified here. □

3. Hamiltonian calculus. We now take up the second question raised at the beginning of §2. Given a multifunction F satisfying our standing hypotheses, and an F-trajectory $\overline{x}$, suppose it is possible to choose a function $R(t)$ for which the localization (2.1) produces a multifunction $\widetilde{F}$ with suitable boundedness and Lipschitz properties. What is the relationship between the Hamiltonian of $\widetilde{F}$ and that of the given multifunction F? More specifically, how are their subgradients linked? We answer these questions using simplified notation that suppresses the time-dependence of F, since we are concerned only with partial subgradients computed at fixed times.

Throughout this section, we consider a multifunction F defined on some neighborhood $\overline{x} + \overline{\varepsilon}\mathbb{B}$ of a given point $\overline{x}$, and taking on nonempty closed convex subsets of $\mathbb{R}^n$ as values. We assume that $F(x)$ depends continuously on x in the set $\overline{x} + \overline{\varepsilon}\mathbb{B}$ in the sense that the inner and outer limits of the sets $F(x')$ share the common value $F(x)$ as $x' \to x$. Given a point $\overline{v}$ in $F(\overline{x})$, we consider the localized multifunction $\widetilde{F}(x) := F(x) \cap (\overline{v} + R\,\mathrm{cl}\,\mathbb{B})$ for some fixed $R > 0$. Like its predecessor F, the multifunction $\widetilde{F}$ has closed convex values on the set $\overline{x} + \overline{\varepsilon}\mathbb{B}$. Concerning $\widetilde{F}$, we assume that

$$\widetilde{F} \text{ is Lipschitz of rank } k \text{ on } \overline{x} + \overline{\varepsilon}\mathbb{B} \tag{3.1}$$

(in particular, $\widetilde{F}$ is nonempty-valued there), and consider a vector $\overline{p}$ with the property that

$$\langle \overline{p}, \overline{v} \rangle \ge \langle \overline{p}, v \rangle \quad \text{for all } v \in \widetilde{F}(\overline{x}). \tag{3.2}$$

Our concern is to relate the subgradients at $(\overline{x},\overline{p})$ of the two Hamiltonians corresponding to F and $\widetilde{F}$, namely

$$\begin{aligned} H(x,p) &:= \sup\{\langle p\,,\,v\rangle \;:\; v \in F(x)\},\\ \widetilde{H}(x,p) &:= \sup\{\langle p\,,\,v\rangle \;:\; v \in \widetilde{F}(x)\}. \end{aligned} \tag{3.3}$$

In particular, we prove that under the hypotheses described above,

$$\overline{\partial}\widetilde{H}(\overline{x},\overline{p}) \subseteq \overline{\partial}H(\overline{x},\overline{p}). \tag{3.4}$$

Recall the Legendre–Fenchel transform that associates to any function $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ its *conjugate*

$$f^*(p) := \sup\{\langle p\,,\,v\rangle - f(v) \;:\; v \in \mathbb{R}^n\}.$$

When f is a proper convex function, its conjugate is as well, and the duality sponsored by this transformation is the cornerstone of many fundamental results in convex analysis. In our current setting, we recognize the Hamiltonians H and $\widetilde{H}$ as the conjugates of certain indicator functions. Using the notation $\Psi_C(v) := 0$ if $v \in C$ and $\Psi_C(v) := +\infty$ if $v \notin C$, we have

$$\begin{aligned} H(x,p) &= \left(\Psi_{F(x)}\right)^*(p),\\ \widetilde{H}(x,p) &= \left(\Psi_{\widetilde{F}(x)}\right)^*(p) = \left(\Psi_{F(x)} + \Psi_{\overline{v}+R\,\mathrm{cl}\,\mathbb{B}}\right)^*(p). \end{aligned}$$

One important consequence of this observation is that the possibly extended-valued function H is lower semicontinuous on $(\overline{x}+\overline{\varepsilon}\mathbb{B})\times\mathbb{R}^n$. Indeed, the continuity of F assumed above implies that epi $\Psi_{F(x)}$ varies continuously with x on $\overline{x}+\overline{\varepsilon}\mathbb{B}$. According to Wijsman's theorem [32, Thm. 6.2], epi-continuity is preserved under the Legendre–Fenchel transform. In particular, epi $H(x,\cdot)$ also varies continuously with x on $\overline{x}+\overline{\varepsilon}\mathbb{B}$. The very definition of epi-continuity now implies that H is lower semicontinuous near $(\overline{x},\overline{p})$.

Infimal convolution. Addition of proper convex functions corresponds to infimal convolution of their conjugates under the Legendre–Fenchel transform: According to *Convex Analysis* [21, Thm. 16.4], we have the following identity for all x near $\overline{x}$ and all $p \in \mathbb{R}^n$:

$$\begin{aligned} \widetilde{H}(x,p) &= \left(\Psi^*_{F(x)} \,\square\, \Psi^*_{\overline{v}+R\,\mathrm{cl}\,\mathbb{B}}\right)(p)\\ &= \inf\left\{\Psi^*_{F(x)}(p-z) + \Psi^*_{\overline{v}+R\,\mathrm{cl}\,\mathbb{B}}(z) \;:\; z \in \mathbb{R}^n\right\}\\ &= \inf\{H(x,p-z) + \langle \overline{v}\,,\,z\rangle + R|z| \;:\; z \in \mathbb{R}^n\}. \end{aligned} \tag{3.5}$$

(The hypotheses of [21, Thm. 16.4] require that for each x near $\overline{x}$, the convex sets $\mathrm{ri}(\mathrm{dom}\,H(x,\cdot))$ and $\mathrm{ri}(\mathrm{dom}\,\widetilde{H}(x,\cdot))$ have a point in common. However, because $\widetilde{F}$ is bounded, the latter set is the whole space $\mathbb{R}^n$; the former set is nonempty, so this hypothesis holds.)

Subgradient analysis. Equation (3.5) expresses $\widetilde{H}$ as the value function associated with a minimization problem depending upon the parameters x and p. Proximal analysis is a powerful technique for estimating the subgradients of such functions: the

situation we now face is covered by Rockafellar [29, Thm. 8.3]. If we write $f(z,x,p) := H(x,p-z) + \langle \overline{v}\,,\, z\rangle + R|z|$, that result affirms that

$$\begin{aligned}\overline{\partial}\widetilde{H}(\overline{x},\overline{p}) \subseteq \operatorname{cl}\operatorname{co}\Bigg[&\bigcup_{z\in\Sigma(\overline{x},\overline{p})} \left\{(\pi,v) \,:\, (0,\pi,v)\in \overline{\partial} f(z,\overline{x},\overline{p})\right\} \\ &+ \bigcup_{z\in\Sigma(\overline{x},\overline{p})} \left\{(\pi,v) \,:\, (0,\pi,v)\in \overline{\partial}^{\infty} f(z,\overline{x},\overline{p})\right\}\Bigg],\end{aligned} \tag{3.6}$$

where $\Sigma(\overline{x},\overline{p})$ denotes the set of all points $z\in\mathbb{R}^n$ at which the infimum in (3.5) is attained. Our assumption (3.2) and the convexity of the set $F(\overline{x})$ together ensure that one such point is $z=0$, where $\widetilde{H}(\overline{x},\overline{p}) = \langle \overline{p}\,,\, \overline{v}\rangle = H(\overline{x},\overline{p})$. However, because $\overline{v}\in F(\overline{x})$, we have $H(\overline{x},\overline{p}-z) \geq \langle \overline{p}-z\,,\, \overline{v}\rangle$ for any $z\in\mathbb{R}^n$; hence

$$H(\overline{x},\overline{p}-z) + \langle \overline{v}\,,\, z\rangle + R|z| \geq H(\overline{x},\overline{p}) + R|z|.$$

The right-hand side above strictly exceeds the minimum value $H(\overline{x},\overline{p})$ for all z except $z=0$. Therefore $\Sigma(\overline{x},\overline{p}) = \{0\}$, and inclusion (3.6) simplifies to

$$\begin{aligned}\overline{\partial}\widetilde{H}(\overline{x},\overline{p}) \subseteq \operatorname{cl}\operatorname{co}\Big[&\left\{(\pi,v) \,:\, (0,\pi,v)\in \overline{\partial} f(0,\overline{x},\overline{p})\right\} \\ &+ \left\{(\pi,v) \,:\, (0,\pi,v)\in \overline{\partial}^{\infty} f(0,\overline{x},\overline{p})\right\}\Big].\end{aligned} \tag{3.7}$$

Before completing our analysis of (3.7), we must confirm that the derivation of (3.6) from [29, Thm. 8.3] is justified. This requires that we verify three hypotheses. First, the function $\widetilde{H}$ must be finite at $(\overline{x},\overline{p})$. This requires only that $\widetilde{F}(\overline{x})$ be nonempty, which we have assumed from the start. Second, a certain constraint qualification must hold at $(z,\overline{x},\overline{p})$ for every $z\in\Sigma(\overline{x},\overline{p})$. This turns out to be trivial because the constraint structure of our problem is so much simpler than that involved in the general situation of the cited theorem. Third, there must exist constants $\varepsilon>0$ and $\overline{\alpha} > \widetilde{H}(\overline{x},\overline{p})$ such that the following set is bounded:

$$S := \{(z,x,p) \,:\, H(x,p-z) + \langle \overline{v}\,,\, z\rangle + R|z| \leq \overline{\alpha},\ |(x,p)-(\overline{x},\overline{p})| \leq \varepsilon\}.$$

To prove this, we apply the Lipschitz hypothesis (3.1), which implies (since $\overline{v}\in\widetilde{F}(\overline{x})$) that for any x in $\overline{x}+\overline{\varepsilon}\mathbb{B}$,

$$(\overline{v} + k|x-\overline{x}|\operatorname{cl}\mathbb{B}) \cap \widetilde{F}(x) \neq \emptyset.$$

It follows that for any such x, and for any $p,z\in\mathbb{R}^n$,

$$H(x,p-z) \geq \widetilde{H}(x,p-z) \geq \langle p-z\,,\, \overline{v}\rangle - k|x-\overline{x}||p-z|.$$

Therefore, if we choose $\varepsilon>0$ small enough that $R-\varepsilon k > \varepsilon/2$, any triple (z,x,p) satisfying the defining inequalities in S will obey

$$\begin{aligned}\overline{\alpha} &\geq H(x,p-z) + \langle \overline{v}\,,\, z\rangle + R|z| \\ &\geq R|z| + \langle p\,,\, \overline{v}\rangle - k|x-\overline{x}||p-z| \\ &\geq (R-k|x-\overline{x}|)\,|z| - (|\overline{v}| + k|x-\overline{x}|)\,|p| \\ &\geq (\varepsilon/2)|z| - (|\overline{v}| + k\varepsilon)(|\overline{p}|+\varepsilon).\end{aligned}$$

This clearly imposes an upper bound on $|z|$, and it follows that the set S is bounded. Our verification of the hypotheses of [29, Thm. 8.3] is complete, and we can apply its conclusions (3.6) and (3.7) with confidence.

To complete our derivation of inclusion (3.4), it remains only to compute the subgradient sets appearing in (3.7). Recall that $f(z,x,p) = H(x,p-z)+\langle \overline{v}\,,\, z\rangle + R|z|$; thus f is the sum of a lower semicontinuous function and a continuous convex function. According to Thm. 8.1 (the sum rule) and Cor. 7.1.2 (the chain rule) of Rockafellar [29], we have

$$\begin{aligned}\overline{\partial} f(0,\overline{x},\overline{p}) &\subseteq \{(-v,\pi,v) \,:\, (\pi,v)\in \overline{\partial} H(\overline{x},\overline{p})\} + (\overline{v}+R\,\mathrm{cl}\,\mathbb{B})\times\{(0,0)\}\\ \overline{\partial}^{\infty} f(0,\overline{x},\overline{p}) &\subseteq \left\{(-v,\pi,v) \,:\, (\pi,v)\in \overline{\partial}^{\infty} H(\overline{x},\overline{p})\right\}.\end{aligned}$$

Thus (3.7) yields

$$\begin{aligned}\overline{\partial}\widetilde{H}(\overline{x},\overline{p}) &\subseteq \mathrm{cl}\,\mathrm{co}\Big(\Big[\overline{\partial} H(\overline{x},\overline{p})\cap(\mathbb{R}^n\times(\overline{v}+R\,\mathrm{cl}\,\mathbb{B}))\Big]\\ &\qquad + \Big[\overline{\partial}^{\infty} H(\overline{x},\overline{p})\cap(\mathbb{R}^n\times\{0\})\Big]\Big)\\ &\subseteq \mathrm{cl}\,\mathrm{co}\left(\overline{\partial} H(\overline{x},\overline{p})+\overline{\partial}^{\infty} H(\overline{x},\overline{p})\right).\end{aligned}\tag{3.8}$$

Since the set on the left side is nonempty, the set on the right must also be nonvoid. This forces $\overline{\partial} H(\overline{x},\overline{p})\neq\emptyset$, a situation in which $\overline{\partial}^{\infty} H(\overline{x},\overline{p})$ is known to equal the recession cone of $\overline{\partial} H(\overline{x},\overline{p})$. In particular, $\overline{\partial} H(\overline{x},\overline{p})+\overline{\partial}^{\infty} H(\overline{x},\overline{p})$ is a subset of the closed convex set $\overline{\partial} H(\overline{x},\overline{p})$. Therefore (3.8) implies (3.4), and the objective of this section is accomplished.

4. General necessary conditions. We now combine the efforts of the first three sections to prove necessary conditions for optimality in (P) without the boundedness and Lipschitz continuity assumptions used previously.

Our first result, Theorem 4.1, extends the Hamiltonian necessary conditions of Theorem 1.2 to the unbounded case. Although this is a significant advance in itself, it is superseded by Theorem 4.3 below, in which the same hypotheses are used to produce an adjoint function satisfying the Hamiltonian inclusion, a refined Euler–Lagrange inclusion, and the Weierstrass–Pontryagin maximum condition, simultaneously. Our purpose in proving Theorem 4.1 first is to clarify the roles of §§2 and 3 in eliminating boundedness assumptions. This provides a convenient point at which to reflect on what has been achieved, and to prepare for the next step.

Hypotheses (H1)–(H4) mentioned in the statement below are listed in §1; the notions required in assumptions (i) and (ii) are described in Definition 2.3. As observed in §2, assumptions (i) and (ii) can be replaced by stronger hypotheses requiring appropriate sub-Lipschitzian behaviour at every point of Ω if the arc $\overline{x}$ is not known in advance.

THEOREM 4.1 (Hamiltonian necessary conditions). *Assume* (H1)–(H4). *Suppose that the arc $\overline{x}$ solves problem (P), and that the constraint qualification below is satisfied:*

(CQ) *The cone $\overline{N}_X(t,\overline{x}(t))$ is pointed for all t in $[a,b]$.*

Suppose further that one of the following two conditions holds:

(i) *The arc $\overline{x}$ is Lipschitzian, and the multifunction F is sub-Lipschitzian at every point $(t,\overline{x}(t))$ of* $\mathrm{gph}\,\overline{x}$; *or*

(ii) *The multifunction F is integrably sub-Lipschitzian in the large at every point $(t, \overline{x}(t))$ of* $\operatorname{gph} \overline{x}$.

Then there exist a scalar $\lambda \in \{0, 1\}$ and a function $p \in BV([a,b]; \mathbb{R}^n)$, not both zero, such that we have the following:

(a) *The Hamiltonian inclusion*

$$(-\dot{p}(t), \dot{\overline{x}}(t)) \in \overline{\partial} H(t, \overline{x}(t), p(t)) - \overline{N}_X(t, \overline{x}(t)) \times \{0\} \quad \text{a.e. } t \in [a,b];$$

(b) *The transversality inclusion*

$$(p(a), -p(b)) \in \lambda \partial \ell(\overline{x}(a), \overline{x}(b)) + N_S(\overline{x}(a), \overline{x}(b)); \text{ and}$$

(c) *The singular part of the measure dp is $\overline{N}_X(t, \overline{x}(t))$-valued and, in particular, is supported on the set*

$$\{t : \overline{N}_X(t, \overline{x}(t)) \neq \{0\}\} = \{t \in [a,b] : (t, \overline{x}(t)) \in \operatorname{bdry} \operatorname{gph} X\}.$$

Remark. The interpretation of inclusion (a) in Theorem 4.1 is the same as that given in Theorem 1.2. That is, (a) asserts that for some integrable selection $\nu(t) \in \overline{N}_X(t, \overline{x}(t))$ for all $t \in [a,b]$, we have

$$(-\dot{p}(t) + \nu(t), \dot{\overline{x}}(t)) \in \overline{\partial} H(t, \overline{x}(t), p(t)) \quad \text{a.e. } t \in [a,b].$$

Proof. Under either hypothesis (i) or (ii), Proposition 2.4 describes a choice of $R(t)$ for which the truncated multifunction $\widetilde{F}(t,x) := F(t,x) \cap (\dot{\overline{x}}(t) + R(t) \operatorname{cl} \mathbb{B})$ satisfies both assumptions (i) and (ii) of Theorem 1.2. Of course, the arc $\overline{x}$ is a trajectory for $\widetilde{F}$, and consequently solves the problem $(\widetilde{P})$ defined by replacing F with $\widetilde{F}$ in (P). Apply Theorem 1.2 to $\overline{x}$ in $(\widetilde{P})$; this produces a constant λ and an adjoint function p of bounded variation, not both zero, together with a selection $\nu(t)$ of $\overline{N}_X(t, \overline{x}(t))$, satisfying all the conclusions of Theorem 1.2. Let us denote these by $(\widetilde{\text{a}})$–$(\widetilde{\text{c}})$, since they involve the multifunction $\widetilde{F}$ and its associated Hamiltonian $\widetilde{H}$.

We show that these three conditions imply the desired conclusions (a)–(c) for the same λ, p, and ν. Indeed, conditions $(\widetilde{\text{b}})$ and $(\widetilde{\text{c}})$ are the same as the desired assertions (b) and (c), while $(\widetilde{\text{a}})$ implies (a). To justify the latter assertion, fix $t \in (a,b)$ and consider the multifunctions $F(t, \cdot)$ and $\widetilde{F}(t, \cdot)$. Our assumption of either (i) or (ii) implies that the given multifunction $F(t, \cdot)$ is continuous in the weak sense required in §3, and that the truncated multifunction $\widetilde{F}(t, \cdot)$ satisfies the Lipschitz condition (3.1) (see Proposition 2.4). Hypothesis (3.2) for $\overline{p} = p(t)$ is a well-known consequence of $(\widetilde{\text{a}})$ (see Clarke [2, Prop. 3.2.4(d)]). The conclusion is that for almost every time $t \in [a,b]$, $\overline{\partial} \widetilde{H}(t, \overline{x}(t), p(t)) \subseteq \overline{\partial} H(t, \overline{x}(t), p(t))$. Hence (a) follows from $(\widetilde{\text{a}})$, as required. □

Strict convexity. The crucial observation that allowed us to unify the adjoint inclusions of Hamilton, Euler–Lagrange, and Weierstrass–Pontryagin in [13] was that the Hamiltonian inclusion actually implies the other two inclusions when $\dot{\overline{x}}(t)$ is almost always an extreme point of the (convex) velocity set $F(t, \overline{x}(t))$. We now use the same observation to extend Theorem 4.1. Let us continue under the hypotheses of that result.

Consider the function

$$L(t,v) := \sqrt{1 + |v - \dot{\overline{x}}(t)|^2} - 1.$$

Notice that L is nonnegative, smooth, and strictly convex, with $L(t,\dot{\overline{x}}(t)) \equiv 0$ and $L_v(t,\dot{\overline{x}}(t)) \equiv 0$. Observe also that for each fixed t, the function $L(t,\cdot)$ is globally Lipschitzian of rank 1 on $\mathbb{R}^n$. These properties are important in our analysis of the following auxiliary problem, whose state (x,y) evolves in $\mathbb{R}^n \times \mathbb{R}$:

$$(\mathcal{P}) \qquad \begin{array}{ll} \text{minimize} & \ell(x(a),x(b)) + y(b) \\ \text{subject to} & (\dot{x}(t),\dot{y}(t)) \in [F(t,x(t)) \times \mathbb{R}] \cap \operatorname{epi} L(t,\cdot) \quad \text{a.e. } t \in [a,b], \\ & (x(a),x(b)) \in S, \qquad y(a) = 0, \\ & (x(t),y(t)) \in X(t) \times \mathbb{R} \quad \forall t \in [a,b]. \end{array}$$

It is clear that any absolutely continuous function $(x(t),y(t))$ admissible for the auxiliary problem $(\mathcal{P})$ has a first component admissible for the original problem (P), while the second component obeys $y(b) \geq 0$. Therefore, the objective value in $(\mathcal{P})$ is always at least as large as the objective value in (P). However, the arc $(\overline{x},\overline{y})$ for which $\overline{y}(t) \equiv 0$ is admissible for $(\mathcal{P})$, and has an objective value equal to the minimum value in (P). Therefore, it must be optimal in $(\mathcal{P})$.

The dynamic constraint in $(\mathcal{P})$ involves the unbounded multifunction $\mathcal{F}\colon \Omega \times \mathbb{R} \rightrightarrows \mathbb{R}^n \times \mathbb{R}$ defined by

$$\mathcal{F}(t,x,y) := [F(t,x) \times \mathbb{R}] \cap \operatorname{epi} L(t,\cdot).$$

We now show that $\mathcal{F}$ inherits the sub-Lipschitzian property of F along $\overline{x}$, and consequently admits a truncation displaying the boundedness and Lipschitz continuity properties required for the application of Theorem 1.2. Since the y dependence of $\mathcal{F}$ is trivial, we suppress it in the notation below.

LEMMA 4.2. *Suppose that hypothesis* (i) *or* (ii) *of Theorem* 4.1 *holds for the multifunction F relative to the arc $\overline{x}$. Then the same hypothesis holds for $\mathcal{F}$ relative to $\overline{x}$. Moreover, there exists a nonnegative function R such that both truncated multifunctions below satisfy hypotheses* (i) *and* (ii) *of Theorem* 1.2 *on some relatively open subset of $[a,b] \times \mathbb{R}^n$ containing the following graph of $\overline{x}$:*

$$\widetilde{F}(t,x) := F(t,x) \cap \left[\dot{\overline{x}}(t) + R(t)\,\mathrm{cl}\,\mathbb{B}\right],$$
$$\widetilde{\mathcal{F}}(t,x) := \mathcal{F}(t,x) \cap \left[(\dot{\overline{x}}(t),0) + R(t)(\mathrm{cl}\,\mathbb{B} \times [-1,1])\right].$$

Proof. (ii) Suppose F satisfies hypothesis 4.1(ii) relative to $\overline{x}$. Let any $\overline{t} \in [a,b]$ be given. Then by hypothesis, there must be constants $\varepsilon > 0$ and $\beta \geq 0$, together with a nonnegative function α integrable on $(\overline{t}-\varepsilon,\overline{t}+\varepsilon) \cap [a,b]$, such that

$$(*) \qquad F(t,x') \cap \rho\,\mathrm{cl}\,\mathbb{B} \subseteq F(t,x) + (\alpha(t) + \beta\rho)\,|x'-x|\,\mathrm{cl}\,\mathbb{B}$$

for all $t \in (\overline{t}-\varepsilon,\overline{t}+\varepsilon) \cap [a,b]$, all x,x' in $\overline{x}(\overline{t}) + \varepsilon\mathbb{B}$, and all $\rho \geq 0$.

To prove a similar statement involving $\mathcal{F}$, let any $\rho \geq 0$ and $t \in (\overline{t}-\varepsilon,\overline{t}+\varepsilon) \cap [a,b]$ be given, together with any two points $x,x' \in \overline{x}(\overline{t}) + \rho\,\mathrm{cl}\,\mathbb{B}$. Then for any point (v',r') in $\mathcal{F}(t,x') \cap \rho(\mathrm{cl}\,\mathbb{B} \times [-1,1])$, we have $v' \in F(t,x') \cap \rho\,\mathrm{cl}\,\mathbb{B}$. Thus $(*)$ provides a point $v \in F(t,x)$ such that $|v'-v| \leq (\alpha(t)+\beta\rho)|x'-x|$. Now $L(t,\cdot)$ is Lipschitz of rank 1, and $r' \geq L(t,v')$. Hence there must be a point $r \geq L(t,v)$ for which $|r'-r| \leq |v'-v|$. Thus (v,r) is a point in $\mathcal{F}(t,x)$ for which

$$\begin{aligned} |(v',r') - (v,r)| &\leq |v'-v| + |r'-r| \\ &\leq 2|v'-v| \\ &\leq 2(\alpha(t)+\beta\rho)|x'-x|. \end{aligned}$$

Since (v', r') is arbitrary, this argument proves that

$$(**) \quad \mathcal{F}(t,x') \cap \rho\,(\mathrm{cl}\,\mathbb{B} \times [-1,1]) \subseteq \mathcal{F}(t,x) + 2\,(\alpha(t) + \beta\rho)\,|x' - x|\,(\mathrm{cl}\,\mathbb{B} \times [-1,1])\,.$$

Hypothesis 4.1(ii) for $\mathcal{F}$ follows.

Now if we multiply both α and β in $(*)$ by 2, we find that both multifunctions F and $\mathcal{F}$ are integrably sub-Lipschitzian in the large at $(\bar{t}, \overline{x}(\bar{t}))$ with the same choices of ε, $2\alpha(t)$, and 2β in the definition. Reviewing the proof of Proposition 2.4, we deduce that the function $R(t) := 1 + 2\alpha(t) + \left|\dot{\overline{x}}(t)\right|$ provides a truncation radius for which each multifunction $\widetilde{F}$, $\widetilde{\mathcal{F}}$ satisfies hypotheses (i) and (ii) of Theorem 1.2 on some neighborhood of $\mathrm{gph}\,\overline{x}$. We restrict attention to the intersection of these two neighborhoods to obtain the desired conclusion.

(i) If F satisfies hypothesis 4.1(i) relative to $\overline{x}$, then an argument similar to that just given shows that for every point $\bar{t} \in [a,b]$ and every $\rho \geq 0$, there exist constants $\varepsilon > 0$ and $\alpha \geq 0$ such that

$$(\dagger) \qquad F(t,y) \cap \rho\,\mathrm{cl}\,\mathbb{B} \subseteq F(t,x) + 2\alpha|y - x|\,\mathrm{cl}\,\mathbb{B}$$

and

$$(\ddagger) \qquad \mathcal{F}(t,y) \cap \rho\,(\mathrm{cl}\,\mathbb{B} \times [-1,1]) \subseteq \mathcal{F}(t,x) + 2\alpha|y - x|\,(\mathrm{cl}\,\mathbb{B} \times [-1,1])$$

for all $t \in (\bar{t} - \varepsilon, \bar{t} + \varepsilon) \cap [a,b]$ and all x, y in $\overline{x}(\bar{t}) + \varepsilon\mathbb{B}$. Just as above, the proof of Proposition 2.4 shows that any constant value of $R > 0$ provides a truncation radius suitable for both multifunctions F and $\mathcal{F}$ at once. □

Just as in the proof of Theorem 4.1, the arc $(\overline{x}, 0)$ that solves $(\mathcal{P})$ remains optimal for the problem $(\widetilde{\mathcal{P}})$ obtained from $(\mathcal{P})$ by changing $\mathcal{F}$ to $\widetilde{\mathcal{F}}$. We apply Theorem 1.2 to deduce that there exist a scalar $\lambda \geq 0$ and a function $(p,q)\colon [a,b] \to \mathbb{R}^n \times \mathbb{R}$ of bounded variation, not both zero, together with a selection $\nu(t) \in \overline{N}_X(t, \overline{x}(t))$ for all $t \in [a,b]$ such that

(a) $(-\dot{p}(t) + \nu(t), \dot{\overline{x}}(t), 0) \in \overline{\partial}\widetilde{\mathcal{H}}(t, \overline{x}(t), p(t), q(t))$ almost everywhere in $[a,b]$, $\dot{q}(t) = 0$ almost everywhere in $[a,b]$;

(b) $(p(a), -p(b)) \in \lambda\partial\ell(\overline{x}(a), \overline{x}(b)) + N_S\,(\overline{x}(a), \overline{x}(b))$, $q(b) = -\lambda$;

(c) The singular part of the measure (dp, dq) is $\overline{N}_X(t, \overline{x}(t)) \times \{0\}$-valued and, in particular, is supported on the set

$$\left\{t \,:\, \overline{N}_X(t, \overline{x}(t)) \neq \{0\}\right\} = \{t \in [a,b] \,:\, (t, \overline{x}(t)) \in \mathrm{bdry}\,\mathrm{gph}\,X\}\,.$$

Here we use the fact that $\widetilde{\mathcal{F}}$ is independent of y to simplify the Hamiltonian inclusion; we need only deal with the reduced Hamiltonian given by

$$\widetilde{\mathcal{H}}(t,x,p,q) := \sup\left\{\langle p\,,\, v\rangle + qr \,:\, (v,r) \in \widetilde{\mathcal{F}}(t,x)\right\}.$$

Conditions (a)–(c) together imply that the adjoint function's $q(t)$ component is actually constant, with the value $-\lambda$. Thus conclusions (b) and (c) reduce to the expected transversality and support conditions associated with the adjoint function p, while conclusion (a) may be written as follows:

$$(4.1) \qquad (-\dot{p}(t) + \nu(t), \dot{\overline{x}}(t), 0) \in \overline{\partial}\widetilde{\mathcal{H}}(t, \overline{x}(t), p(t), -\lambda) \quad \text{a.e. } t \in [a,b].$$

In the remainder of this section we use inclusion (4.1) to show that the function p satisfies the Hamiltonian inclusion, a refined Euler–Lagrange inclusion, and the Weierstrass–Pontryagin maximum condition for the original problem (P).

The maximum condition. We have seen in Lemma 4.2 that $\widetilde{\mathcal{F}}$ is a multifunction satisfying hypotheses (i) and (ii) of Theorem 1.2. Under these assumptions, Clarke [2, Prop. 3.2.4(d)] shows that inclusion (4.1) implies

$$(p(t), -\lambda) \in N_{\widetilde{\mathcal{F}}(t,\overline{x}(t))}\left(\dot{\overline{x}}(t), 0\right) \quad \text{a.e. } t \in [a,b]. \tag{4.2}$$

However, for each $t \in [a,b]$, the compact set $\widetilde{\mathcal{F}}(t,\overline{x}(t))$ coincides with the unbounded set $\mathcal{F}(t,\overline{x}(t))$ on a neighborhood of $(\dot{\overline{x}}(t), 0)$. Hence these two sets have the same normal cone at this point. Using the calculus of convex normal cones (Rockafellar [21], or [29, Cor. 8.1.1]), we deduce that for almost every $t \in [a,b]$,

$$\begin{aligned}(p(t), -\lambda) &\in N_{\mathcal{F}(t,\overline{x}(t))}\left(\dot{\overline{x}}(t), 0\right) \\ &= N_{F(t,\overline{x}(t))\times\mathbb{R}\cap \operatorname{epi} L(t,\cdot)}\left(\dot{\overline{x}}(t), 0\right) \\ &\subseteq N_{F(t,\overline{x}(t))\times\mathbb{R}}\left(\dot{\overline{x}}(t), 0\right) + N_{\operatorname{epi} L(t,\cdot)}\left(\dot{\overline{x}}(t), 0\right) \\ &= N_{F(t,\overline{x}(t))}\left(\dot{\overline{x}}(t)\right) \times \{0\} + \{0\} \times (-\infty, 0].\end{aligned}$$

(The last step uses the fact that $N_{\operatorname{epi} L(t,\cdot)}\left(\dot{\overline{x}}(t), 0\right)$ is the convex cone generated by $\partial L(t,\dot{\overline{x}}(t)) \times \{-1\} = \{(0,-1)\}$.) The first component of this inclusion gives the desired maximum condition for p, namely,

$$p(t) \in N_{F(t,\overline{x}(t))}\left(\dot{\overline{x}}(t)\right) \quad \text{a.e. } t \in [a,b]. \tag{4.3}$$

The Hamiltonian inclusion. Observe that the function $\widetilde{\mathcal{H}}$ can be written as follows:

$$\widetilde{\mathcal{H}}(t,x,p,q) = \begin{cases} \sup\{\langle p\,,\, v\rangle + qL(t,v) \,:\, v \in \widetilde{F}(t,x)\}, & \text{if } q < 0, \\ \sup\{\langle p\,,\, v\rangle \,:\, v \in \widetilde{F}(t,x)\} + qR(t), & \text{if } q \geq 0. \end{cases}$$

This is precisely the sort of function studied in §4 of our previous work [13], where we examined its relationship to the function below:

$$\widetilde{\mathcal{H}}_\lambda(t,x,p) := \sup\left\{\langle p\,,\, v\rangle - \lambda L(t,v) \,:\, v \in \widetilde{F}(t,x)\right\}.$$

Lemma 4.2 ensures that for each fixed $t \in [a,b]$, the multifunction $\widetilde{F}(t,\cdot)$ in this expression obeys the standing assumptions (A1)–(A3) of [13, §4].

This observation allows us to apply [13, Thm. 4.4] to inclusion (4.1), and thereby derive

$$(-\dot{p}(t) + \nu(t), \dot{\overline{x}}(t)) \in \overline{\partial}\widetilde{\mathcal{H}}_\lambda(t,\overline{x}(t),p(t)) \quad \text{a.e. } t \in [a,b]. \tag{4.4}$$

In the case where $\lambda = 0$, $\widetilde{\mathcal{H}}_\lambda$ coincides with $\widetilde{H}$. Thus inclusion (4.4) is equivalent to

$$(-\dot{p}(t) + \nu(t), \dot{\overline{x}}(t)) \in \overline{\partial}\widetilde{H}(t,\overline{x}(t),p(t)) \quad \text{a.e. } t \in [a,b]. \tag{4.5}$$

In the case where $\lambda = 1$, inclusion (4.4) implies (4.5) via [13, Cor. 4.3(b)]. To justify this, fix any time t where (4.4) holds and apply Clarke [2, Prop. 2.5.3] to deduce that $\dot{\overline{x}}(t) \in \partial_p\widetilde{\mathcal{H}}_1(t,\overline{x}(t),p(t))$. Conversely, elementary convex analysis shows that any vector v lying in $\partial_p\widetilde{\mathcal{H}}_1(t,\overline{x}(t),p(t))$ must maximize the function $v' \mapsto \langle p\,,\, v'\rangle - L(t,v')$ over the set $\widetilde{F}(t,\overline{x}(t))$. Because this function is strictly concave by construction, only one maximizer can exist, namely $\dot{\overline{x}}(t)$. Consequently, $\overline{\partial}_p\widetilde{\mathcal{H}}_1(t,\overline{x}(t),p(t)) = \{\dot{\overline{x}}(t)\}$.

The union appearing in [13, Cor. 4.3(b)] therefore involves the choices $v = \dot{\overline{x}}(t)$ and $z \in \overline{\partial}_v L(t, \overline{x}(t)) = \{0\}$. This implies that the right-hand side of (4.4) is a subset of the right-hand side of (4.5).

With (4.5) in hand, we note that the Hamiltonian analysis of §3 allows us to replace $\widetilde{H}$ with H in (4.5) just as we did in the proof of Theorem 4.1. The result is the desired Hamiltonian inclusion for p:

$$(-\dot{p}(t) + \nu(t), \dot{\overline{x}}(t)) \in \overline{\partial} H(t, \overline{x}(t), p(t)) \quad \text{a.e. } t \in [a, b]. \tag{4.6}$$

The Euler–Lagrange inclusion. Once again we rely upon the technical results of Loewen and Rockafellar [13], as extended by Rockafellar [30]. To streamline the discussion, we fix a time $t \in [a, b]$ at which (4.1) holds, and suppress t in the notation below. Thus our starting point is the inclusion

$$(-\dot{p} + \nu, \dot{\overline{x}}, 0) \in \overline{\partial} \widetilde{\mathcal{H}}(\overline{x}, p, -\lambda). \tag{4.7}$$

Rockafellar [30, Thm. 3.1] provides a far-reaching analogue of [13, Lemma 4.5] in which limiting normals and subgradients replace Clarke normals and subgradients. Applying this result to our problem (with $f = \Psi_{\widetilde{\mathcal{F}}}$), we find that

$$\begin{aligned} -\partial(-\widetilde{\mathcal{H}})(\overline{x}, p, -\lambda) \subseteq \{(-u, v, \ell) \,:\, & (u, p, -\lambda) \in N_{\operatorname{gph} \widetilde{\mathcal{F}}}(\overline{x}, v, \ell), \\ & (p, -\lambda) \in N_{\widetilde{\mathcal{F}}(x)}(v, \ell)\}. \end{aligned} \tag{4.8}$$

Since the function $\widetilde{\mathcal{H}}$ is Lipschitzian, we have $\overline{\partial} \widetilde{\mathcal{H}} = -\overline{\partial}(-\widetilde{\mathcal{H}}) = \operatorname{co} -\partial(-\widetilde{\mathcal{H}})$.

Thus it must be possible to express the point $(-\dot{p}+\nu, \dot{\overline{x}}, 0)$ as a convex combination of elements from the right side of (4.8). That is, there must be some $N \in \mathbf{N}$ and some constants $\alpha_i \geq 0$ with $\sum \alpha_i = 1$ such that

$$(-\dot{p} + \nu, \dot{\overline{x}}, 0) = \sum_{i=1}^{N} \alpha_i (-u_i, v_i, \ell_i), \tag{4.9a}$$

where, for each i,

$$(u_i, p, -\lambda) \in N_{\operatorname{gph} \widetilde{\mathcal{F}}}(\overline{x}, v_i, \ell_i), \qquad (p, -\lambda) \in N_{\widetilde{\mathcal{F}}(\overline{x})}(v_i, \ell_i). \tag{4.9b}$$

We have already shown that $(p, -\lambda) \in N_{\widetilde{\mathcal{F}}(\overline{x})}(\dot{\overline{x}}, 0)$ (see (4.2)). We now add the observation that $(\dot{\overline{x}}, 0)$ is an extreme point of the set $\widetilde{\mathcal{F}}(\overline{x})$. This is obvious, since $\widetilde{\mathcal{F}}(\overline{x})$ is the intersection of the compact convex sets $\operatorname{epi} L$ and $\widetilde{F}(\overline{x}) \times \mathbb{R}$, and $(\dot{\overline{x}}, 0)$ is an extreme point of the first of these by the strict convexity of L. It follows that $(\dot{\overline{x}}, 0)$ is the only point (v, ℓ) for which $(p, -\lambda) \in N_{\widetilde{\mathcal{F}}(\overline{x})}(v, \ell)$. This forces $(v_i, \ell_i) = (\dot{\overline{x}}, 0)$ in (4.9), and thus implies

$$\dot{p} - \nu \in \operatorname{co} \left\{ u \,:\, (u, p, -\lambda) \in N_{\operatorname{gph} \widetilde{\mathcal{F}}}(\overline{x}, \dot{\overline{x}}, 0) \right\}. \tag{4.10}$$

To simplify this assertion, temporarily think of L as a function of both x and v ($L(x, v) \equiv L(v)$) in order to write $\operatorname{gph} \widetilde{\mathcal{F}} = \operatorname{epi}\left(L + \Psi_{\operatorname{gph} \widetilde{F}}\right)$. This allows us to transcribe the inclusion characterizing the right-hand side of (4.10) as

$$(u, p, -\lambda) \in N_{\operatorname{epi}\left(L + \Psi_{\operatorname{gph} \widetilde{F}}\right)}(\overline{x}, \dot{\overline{x}}, 0).$$

The same arguments used in the last paragraph of the proof of Theorem 1.2, together with the observation that $\partial L(\overline{x},\dot{\overline{x}}) = \{(0,0)\}$, show that this inclusion implies

$$(u,p) \in N_{\operatorname{gph}\widetilde{F}}\left(\overline{x},\dot{\overline{x}}\right) = N_{\operatorname{gph}F}\left(\overline{x},\dot{\overline{x}}\right).$$

The equality here holds because the sets $\operatorname{gph}\widetilde{F}$ and $\operatorname{gph}F$ coincide on a neighborhood of the point $(\overline{x},\dot{\overline{x}})$, so their limiting normal cones at this point are identical. Using this statement in (4.10) leads to the following inclusion, in which we revert to fully explicit notation:

$$\dot{p}(t) - \nu(t) \in \operatorname{co}\left\{u \,:\, (u,p(t)) \in N_{\operatorname{gph}F(t,\cdot)}\left(\overline{x}(t),\dot{\overline{x}}(t)\right)\right\}. \tag{4.11}$$

This inclusion holds for all t outside a null subset of $[a,b]$. It is the form of the Euler–Lagrange inclusion we wish to record.

Main result. We now summarize the results of the derivation above. Conclusions (a)–(c) in the following formal statement have already been established as lines (4.3), (4.6), and (4.11) above. We remind the reader that our notation differs slightly from that of Clarke [2], as indicated in the last paragraph of the introduction.

THEOREM 4.3 (General necessary conditions). *Assume* (H1)–(H4). *Suppose that the arc $\overline{x}$ solves problem (P), and that the constraint qualification below is satisfied:*

(CQ) *The cone $N_X(t,\overline{x}(t))$ is pointed for all t in $[a,b]$.*

Suppose further that one of the following two conditions holds:

(i) *The arc $\overline{x}$ is Lipschitzian, and the multifunction F is sub-Lipschitzian at every point $(t,\overline{x}(t))$ of $\operatorname{gph}\overline{x}$; or*

(ii) *The multifunction F is integrably sub-Lipschitzian in the large at every point $(t,\overline{x}(t))$ of $\operatorname{gph}\overline{x}$.*

Then there exist a scalar $\lambda \in \{0,1\}$ and a function $p \in BV([a,b];\mathbb{R}^n)$, not both zero, such that for almost all $t \in [a,b]$, we have

(a) *The Hamiltonian inclusion*

$$(-\dot{p}(t),\dot{\overline{x}}(t)) \in \overline{\partial}H(t,\overline{x}(t),p(t)) - N_X(t,\overline{x}(t)) \times \{0\};$$

(b) *The Euler–Lagrange inclusion*

$$\dot{p}(t) \in \operatorname{co}\left\{u \,:\, (u,p(t)) \in N_{\operatorname{gph}F(t,\cdot)}\left(\overline{x}(t),\dot{\overline{x}}(t)\right)\right\} + \overline{N}_X(t,\overline{x}(t)); \text{ and}$$

(c) *The Weierstrass–Pontryagin maximum condition*

$$\left\langle p(t)\,,\,\dot{\overline{x}}(t)\right\rangle = max\{\langle p(t)\,,\,v\rangle \,:\, v \in F(t,\overline{x}(t))\}.$$

The adjoint function p also satisfies the following:

(d) *The transversality inclusion*

$$(p(a),-p(b)) \in \lambda\partial\ell(\overline{x}(a),\overline{x}(b)) + N_S\left(\overline{x}(a),\overline{x}(b)\right); \text{ and}$$

(e) *The singular part of the measure dp is $N_X(t,\overline{x}(t))$-valued, and in particular is supported on the set*

$$\{t \,:\, N_X(t,\overline{x}(t)) \neq \{0\}\} = \{t \in [a,b] \,:\, (t,\overline{x}(t)) \in \operatorname{bdry}\operatorname{gph}X\}.$$

Remarks. 1. If the state constraint is inactive along the optimal arc $\overline{x}$ and the endpoint constraint set S has the form $C \times \mathbb{R}^n$ or $\mathbb{R}^n \times D$ for some closed sets C, D, then we may take $\lambda = 1$ in Theorem 4.3. This is not completely obvious from the theorem's statement, but it does follow from the proof given above. To see this, note that the scalar λ and the function p described in the conclusions of the theorem actually arise as the dual variables in the auxiliary problem $(\widetilde{\mathcal{P}})$.

Problem $(\widetilde{\mathcal{P}})$ has a Hamiltonian $\widetilde{\mathcal{H}}$ for which the mapping $x' \mapsto \mathcal{H}(t, x', p)$ is Lipschitz of rank $k(t)|p|$ for some integrable function k. Under the extra assumptions above, we have $\nu(t) \equiv 0$, so the adjoint function p must be absolutely continuous and satisfy the differential inequality $|\dot{p}(t)| \leq k(t)|p(t)|$ almost everywhere. If the endpoint conditions described above are satisfied, assuming $\lambda = 0$ leads to either $p(a) = 0$ or $p(b) = 0$. In either case, Gronwall's lemma implies $p(t) \equiv 0$, which is a contradiction.

2. A separated form of the Hamiltonian inclusion can be asserted concurrently with (a)–(c) above. To derive it, note that since $\widetilde{H}$ is locally Lipschitz, inclusion (4.5) implies

$$(-\dot{p}(t) + \nu(t), \dot{\overline{x}}(t)) \in \operatorname{co}\left[\partial_x \widetilde{H}(t, \overline{x}(t), p(t)) \times \partial_p \widetilde{H}(t, \overline{x}(t), p(t))\right] \quad \text{a.e. } t \in [a, b].$$

The second component simply reiterates (4.3), while the first asserts that

$$-\dot{p}(t) + \nu(t) \in \operatorname{co} \partial_x \widetilde{H}(t, \overline{x}(t), p(t)) \quad \text{a.e. } t \in [a, b].$$

Arguments similar to those in §3 allow us to replace $\widetilde{H}$ with H in this statement: The result is the separated Hamiltonian inclusion

$$\begin{aligned} -\dot{p}(t) &\in \operatorname{co} \partial_x H(t, \overline{x}(t), p(t)) - \overline{N}_X(t, \overline{x}(t)), \\ \dot{\overline{x}}(t) &\in \partial_p H(t, \overline{x}(t), p(t)) \quad \text{a.e. } t \in [a, b]. \end{aligned} \tag{4.12}$$

5. Conclusion. Theorem 4.3 is the most general set of necessary conditions available for differential inclusion control problems. To substantiate this claim, we review the literature and discuss several pertinent examples in this section.

The bounded case. Note first that Theorem 4.3 is a strict extension of our best previous result for bounded differential inclusions, the case $L \equiv 0$ of [13, Thm. 1.1]. To prove this, it suffices to show that the boundedness and Lipschitz continuity hypotheses of [13], which coincide with conditions (i)–(ii) of the current Theorem 1.2, imply the hypotheses of Theorem 4.3. Indeed, suppose that the multifunction F satisfies condition (ii) of Theorem 1.2. Then the choices $\beta = 0$ and $\alpha = k$ in Definition 2.3(b) show that F is integrably sub-Lipschitzian in the large at every point (t, x) in Ω. Hence hypothesis (ii) of Theorem 4.3 is satisfied; the conclusions either reproduce those of [13, Thm. 1.1] or are strictly stronger. Note in particular that hypothesis (i) is completely superfluous both in [13] and in Theorem 1.2.

The two conclusions of Theorem 4.3 that differ from their counterparts in [13, Thm. 1.1] are the Euler–Lagrange inclusion (b) and the transversality inclusion (d). The right-hand side of (d) is always a subset of its cognate phrased in terms of Clarke subgradients and normals, although the two right-hand sides coincide whenever the function ℓ and the set S are Clarke regular at the point $(\overline{x}(a), \overline{x}(b))$. Likewise, the Euler–Lagrange inclusion (b) readily implies (but may not be equivalent to) the more familiar form involving Clarke's normal cone,

$$(\dot{p}(t), p(t)) \in \operatorname{co} N_{\operatorname{gph} F}\left(\overline{x}(t), \dot{\overline{x}}(t)\right) - \overline{N}_X(t, \overline{x}(t)) \times \{0\} \text{ a.e.} \tag{5.1}$$

The formulation in (b) has the advantage of applying the convex hull only to variables associated with the derivatives of the adjoint function p. The routine use of weak convergence of the derivatives both in existence theory and in the derivation of necessary conditions makes it difficult to imagine making do with less convexity than this.

The possibility of refining the Euler–Lagrange inclusion in our main theorem was suggested by a recent preprint of Boris Mordukhovich, a pioneer in the systematic reduction of convexity hypotheses in nonsmooth analysis. His manuscript [18] introduces a version of the Euler–Lagrange inclusion whose counterpart in our problem would read as follows:

$$(\dot{p}(t) - \nu(t), \dot{\overline{x}}(t)) \in \operatorname{co}\{(u,v) : (u,p(t)) \in N_{\operatorname{gph} F(t,\cdot)}(\overline{x}(t),v), \quad p(t) \in N_{F(t,\overline{x}(t))}(v)\}. \tag{5.2}$$

(Here, as in §4, $\nu(t)$ is a selection of $\overline{N}_X(t,\overline{x}(t))$. Mordukhovich's work does not allow for state constraints, so his version of (5.2) involves an absolutely continuous function p and $\nu \equiv 0$.) This is clearly a consequence of inclusion (4.11). The two are equivalent if, for almost every t, the maximum value of $\langle p(t), v\rangle$ over $v \in F(t,\overline{x}(t))$ is attained at the unique point $v = \dot{\overline{x}}(t)$. Without this hypothesis, however, the right-hand side of (5.2) may be a proper superset of the right-hand side of (4.11). This is demonstrated by Example 5.2, below. Thus the necessary conditions of Mordukhovich [18] are strictly superseded by those given here. Indeed, Rockafellar's dualization result [30, Thm. 3.1], used to prove the Euler–Lagrange inclusion (4.11), implies that under Mordukhovich's hypotheses in [18], inclusion (5.2) actually follows from the Hamiltonian inclusion in Theorem 1.2.

Although several technical results from our previous work [13] were used to prove Theorem 4.3, this paper's development starts from [13, Thm. 2.8]. Since we recover [13, Thm. 1.1] as a corollary (at least in the case $L \equiv 0$), this paper provides a much simpler alternative to the formidable sequential arguments of [13, §3]. This makes sense, because the sequences of adjoint functions required there arose directly out of a less sophisticated truncation procedure than the one introduced in the current work.

The simultaneous assertion of the adjoint inclusion in both Hamiltonian and Eulerian forms is a significant feature that Theorem 4.3 shares with the main result of [13]. The relationship between these two inclusions in their various forms is still not completely understood. For example, we now show that the Euler–Lagrange inclusion in Clarke's form (5.1) bears no simple relationship to the Hamiltonian inclusion.

EXAMPLE 5.1. *There exist a compact convex valued, Lipschitzian multifunction* $F\colon \mathbb{R} \rightrightarrows \mathbb{R}$ *and a pair of arcs* x, p *on* $[a,b]$ *such that for all* $t \in [a,b]$, *the Clarke form of the Euler–Lagrange inclusion holds, i.e.,*

$$(\dot{p}(t), p(t)) \in \overline{N}_{\operatorname{gph} F}(x(t), \dot{x}(t)). \tag{5.3a}$$

However, the following two inclusions fail:

$$(\dot{p}(t), \dot{x}(t)) \in \operatorname{co}\{(u,v) : (u,p(t)) \in N_{\operatorname{gph} F}(x(t),v),\ p(t) \in N_{F(x(t))}(v)\}, \tag{5.3b}$$

$$(-\dot{p}(t), \dot{x}(t)) \in \overline{\partial} H(x(t),p(t)). \tag{5.3c}$$

Proof. Let $F(x) := [-|x|, |x|]$. This multifunction is compact convex valued and Lipschitz continuous; its graph is the plane set obtained by filling in the vertical space between the lines $y = x$ and $y = -x$. The limiting normal cone to $\operatorname{gph} F$ at the point $(0,0)$ consists of the two lines $y = \pm x$ in the plane; the corresponding Clarke normal

cone is therefore the whole space $\mathbb{R}^2$. Thus for the arc $x(t) \equiv 0$, the right-hand side in Clarke's form of the Euler–Lagrange inclusion (5.3a) is simply $\mathbb{R}^2$, so any arc p will serve. On the other hand, Mordukhovich's form of the Euler–Lagrange inclusion (5.3b) makes a nontrivial restriction on the choice of p. The inclusion $p(t) \in N_{F(0)}(v)$ forces $v = 0$, so that (5.3b) becomes

$$(\dot{p}(t), 0) \in \operatorname{co}\{(u, 0) : |u| = |p(t)|\} = [-|p(t)|, |p(t)|] \times \{0\}.$$

Any arc p that obeys $|\dot{p}(t)| > |p(t)|$ will confirm (5.3a) but violate (5.3b). For example, $p(t) = e^{2t}$ will serve. By the result of Rockafellar cited above, inclusion (5.3c) implies (5.3b). Hence the same choice of p must also violate (5.3c). Of course, this can be confirmed directly by noting that the Hamiltonian corresponding to F is

$$H(x, p) = \sup\{pv : |v| \le |x|\} = |px|,$$

and that for $p \neq 0$, we have $\overline{\partial} H(0, p) = [-|p|, |p|] \times \{0\}$. □

Note that in Example 5.1, the Mordukhovich form of the Euler–Lagrange inclusion is equivalent to the refined form used in Theorem 4.3 because $F(0)$ is a one-point set. Thus Example 5.1 shows that the Euler–Lagrange inclusion in Clarke's form (5.1) does not imply the Hamiltonian inclusion, and that (5.1) can be strictly weaker than our refined Euler–Lagrange inclusion (4.11). However, it does not rule out the possibility that (4.11) implies the Hamiltonian inclusion.

Our next example shows that the Hamiltonian inclusion implies neither Euler–Lagrange inclusion (4.11), nor (5.1) "pointwise"; recall, however, that the Hamiltonian inclusion does imply the Euler–Lagrange inclusion (5.2) in Mordukhovich's form.

EXAMPLE 5.2. *There exist a compact convex valued, Lipschitzian multifunction* $F: \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ *and a collection of points* x, v, p, u *in* $\mathbb{R}^2$ *such that*

$$(-u, v) \in \overline{\partial} H(x, p) \quad \textit{but} \quad (u, p) \notin \overline{N}_{\operatorname{gph} F}(x, v).$$

In particular,

$$u \notin \operatorname{co}\{u' : (u', p) \in N_{\operatorname{gph} F}(x, v)\}.$$

Proof. Define $F: \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ as follows:

$$F(x_1, x_2) := \{(t, t|x_1| + r) : t \in [-1, 1], r \in [a, b]\}.$$

For each $x = (x_1, x_2)$ in $\mathbb{R}^2$, the set $F(x)$ is a solid parallelogram in the plane. The corresponding Hamiltonian is

$$H(x_1, x_2, p_1, p_2) = \big|p_1 + p_2|x_1|\big| + \max\{p_2, 0\}.$$

We consider the points $x = (0, 0)$, $v = (0, 0)$, and $p = (0, -1)$. With these choices, $F(x)$ is the plane rectangle $[-1, 1] \times [a, b]$, and p is an outward normal vector to this set at the boundary point v. The crucial feature of this example is that the hyperplane $x_2 = 0$ that supports the set $F(x)$ at v intersects the set $F(x)$ in more than one point. (In other words, the maximum of $\langle p, v \rangle$ over v in $F(x)$ is attained at infinitely many points.) Clarke's subgradient of H at the point $(x, p) = (0, 0, 0, -1)$ can be calculated using [2, Thm. 2.5.1]; it is the two-dimensional square $[-1, 1] \times \{0\} \times [-1, 1] \times \{0\}$ in $\mathbb{R}^4$. One point in this square is $(1, 0, 0, 0)$, which suggests the choice $u = (-1, 0)$. We claim that $(u, p) = (-1, 0, 0, -1)$ lies outside $\overline{N}_{\operatorname{gph} F}(x, v)$.

To prove this, note that up to a permutation of the coordinates, $\operatorname{gph} F = E \times \mathbb{R}$ for the set $E := \{(x_1, t, |x_1|t + r) : x_1 \in \mathbb{R},\ t \in [-1, 1],\ r \in [a, b]\}$. Near the point $(0, 0, 0)$, E coincides with the epigraph of the function $g\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $g(y, t) := t|y|$. This function g is Lipschitzian, and it is easy to show that $\overline{\partial} g(0, 0) = \{(0, 0)\}$. Therefore

$$\overline{N}_E(0,0,0) = \overline{N}_{\operatorname{epi} g}(0,0,g(0,0)) = \bigcup_{\lambda \geq 0} \lambda \left[\overline{\partial} g(0,0) \times \{-1\}\right] = \{(0,0)\} \times (-\infty, 0].$$

We deduce that

$$\overline{N}_{\operatorname{gph} F}(0,0,0,0) = \{(0,0,0)\} \times (-\infty, 0].$$

In particular, $\overline{N}_{\operatorname{gph} F}(x, v)$ does not contain the point $(u, p) = (-1, 0, 0, -1)$, even though $(-u, v) \in \overline{\partial} H(x, p)$. □

It follows from Example 5.2 that the Mordukhovich form of the Euler–Lagrange inclusion (5.2) may fail to imply either the Clarke form (5.1) or the sharper form (4.11). To see this, recall that the Hamiltonian inclusion in Example 5.2 implies the Mordukhovich inclusion (5.2) by the result of Rockafellar [30] cited above. Hence this is an example in which (5.2) holds, but both (5.1) and (4.11) fail.

The unbounded case. Since Theorem 4.3 incorporates a form of the Euler–Lagrange inclusion at least as sharp as (5.1), it subsumes the main result of Clarke [1]. That result requires the multifunction F to display integrably Lipschitz dependence on the state, a hypothesis strictly stronger than assumption (ii) of Theorem 4.3.

Conditions (i) and (ii) of Theorem 4.3 are not directly comparable to the basic hypothesis of Polovinkin and Smirnov [19]. Their truncation scheme involves a constant truncation radius in place of the positive-valued function $R(t)$ in (2.1), and their work involves explicit assumptions about the behaviour of the truncated multifunction $\widetilde{F}$ along the nominal arc $\overline{x}$. Section 2 in this paper has the advantage of introducing hypotheses only on the pointwise behaviour of the given multifunction F near the nominal arc. Indeed, the whole of §2 can be viewed as a set of verifiable sufficient conditions for a weakened form of Polovinkin and Smirnov's "Condition 1" [19, p. 662] to hold.

Polovinkin and Smirnov's conclusions [19], [20] pertain to differential inclusions whose right-hand side may take on nonconvex values, whereas the convexity of the sets $F(t, x)$ is crucial to our approach. However, their work offers only a version of the Euler–Lagrange inclusion, whereas ours incorporates a Hamiltonian inclusion as well. Even in the case of bounded differential inclusions, no one knows whether the Hamiltonian inclusion is a correct necessary condition in the absence of this convexity hypothesis.

A detailed comparison of our Euler–Lagrange inclusion with that in [19, (17)] is beyond the scope of this discussion. However, two comments are in order. First, the approach in [19], [20] is completely different from ours. It is based on "linearizing" the given differential inclusion about the nominal arc, and examining the manner in which solutions of the linearized system provide approximations for the resulting reachable set. (A similar approach is taken by Frankowska [5], and has recently been extended to second-order approximations by Zheng [33].) Second, we note that in Example 5.1, the inclusion [19, (17)] is equivalent to (5.3a). (In general, [19, (17)] is a sharper condition than (5.3a).) As such, it may generate adjoint arcs that satisfy neither the Hamiltonian inclusion (4.6), nor the refined Euler–Lagrange inclusion (4.11). Thus we have at least one example in which our results outperform those of Polovinkin and Smirnov.

Let us note that Theorem 4.3 cannot be obtained simply by reformulating problem (P) as an instance of the generalized problem of Bolza. For simplicity, we discuss only the case without state constraints by setting $X(t) \equiv \mathbb{R}^n$. Then the definition $L(t,x,v) := \Psi_{F(t,x)}(v)$ puts problem (P) into the following form:

$$(P_B) \qquad \begin{aligned} &\text{minimize } \ell(x(a),x(b)) + \int_0^1 L(t,x(t),\dot{x}(t))\,dt \\ &\text{subject to } (x(a),x(b)) \in S. \end{aligned}$$

The Hamiltonian for this problem is the same as the one we have already associated with F. In particular, since the sets $F(t,\overline{x}(t))$ are not necessarily bounded, the convex functions $p \mapsto H(t,\overline{x}(t),p)$ are not necessarily finite-valued everywhere. This places the current instance of (P_B) beyond the scope of the necessary conditions in Clarke [2, Chap. 4], since the strong Lipschitz condition used there tacitly requires the finiteness of H. (See [2, Remark 4.2.1].) Likewise, the possibility that H could take the value $+\infty$ makes it impossible to verify the basic growth condition assumed in Clarke [3].

Thus Theorem 4.3 not only generalizes the necessary conditions formulated explicitly in terms of differential inclusions, but also lies beyond the reach of the best results on the generalized problem of Bolza. Indeed, there is good reason to expect that Theorem 4.3 may lead to strict improvements of the necessary conditions for the Bolza problem. The authors are now pursuing this prospect.

REFERENCES

[1] F. H. CLARKE, *Optimal solutions to differential inclusions,* J. Optim. Theory Appl., 19 (1976), pp. 469–478.

[2] ———, *Optimization and Nonsmooth Analysis,* John Wiley, New York, 1983.

[3] ———, *Hamiltonian analysis of the generalized problem of Bolza,* Trans. Amer. Math. Soc., 301 (1987), pp. 385–400.

[4] ———, *Methods of Dynamic and Nonsmooth Optimization,* CBMS-NSF Regional Conference Series in Applied Math., vol. 57, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.

[5] H. FRANKOWSKA, *Contingent cones to reachable sets of control systems,* SIAM J. Control Optim., 27 (1989), pp. 170–198.

[6] A. D. IOFFE, *Approximate subdifferentials and applications* 1: *The finite dimensional theory,* Trans. Amer. Math. Soc., 281 (1984), pp. 389–416.

[7] ———, *Approximate subdifferentials and applications* 2, Mathematika, 33 (1986), pp. 111–128.

[8] ———, *Approximate subdifferentials and applications* 3: *The metric theory,* Mathematika, 36 (1989), pp. 1–38.

[9] ———, *Proximal analysis and approximate subdifferentials,* J. London Math. Soc., 41 (1990), pp. 175–192.

[10] B. KASKOSZ, AND S. ŁOJASIEWICZ, JR., *Boundary trajectories of systems with unbounded controls,* J. Optim. Theory Appl., 70 (1991), pp. 539–559.

[11] A. Y. KRUGER AND B. S. MORDUKHOVICH, *Extremal points and the Euler equation in a nonsmooth optimization problem,* Dokl. Acad. Nauk BSSR, 24 (1980), pp. 684–687. (In Russian.)

[12] P. D. LOEWEN, *Quantitative analysis of epi-Lipschitzian sets and functions,* University of British Columbia, 1991 preprint.

[13] P. D. LOEWEN AND R. T. ROCKAFELLAR, *The adjoint arc in nonsmooth optimization,* Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.

[14] P. D. LOEWEN AND R. B. VINTER, *Pontryagin-type necessary conditions for differential inclusion problems,* Systems Control Lett., 9 (1987), pp. 263–265.

[15] B. S. MORDUKHOVICH, *Maximum principle in the problem of time optimal response with nonsmooth constraints,* J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[16] B. S. MORDUKHOVICH, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems,* Soviet Math. Dokl., 22 (1980), pp. 526–530.

[17] ———, *Approximation Methods in Problems of Optimization and Control,* Nauka, Moscow 1988. (In Russian; English translation to appear.)

[18] ———, *On variational analysis of differential inclusions,* Wayne State University, 1991, preprint.

[19] E. S. POLOVINKIN AND G. V. SMIRNOV, *An approach to the differentiation of many-valued mappings, and necessary conditions for optimization of solutions of differential inclusions,* Differential Equations, 22 (1986), pp. 660–668; Differen. Uravn., 22 (1986), pp. 944–954. (In Russian.)

[20] ———, *Time-optimum problem for differential inclusions,* Differential Equations, 22 (1986), pp. 940–952; Differen. Uravn., 22 (1986), pp. 1351–1365. (In Russian.)

[21] R. T. ROCKAFELLAR, *Convex Analysis,* Princeton University Press, Princeton, New Jersey, 1970.

[22] ———, *Integrals which are convex functionals* II, Pacific J. Math., 39 (1971), pp. 429–469.

[23] ———, *State constraints in convex problems of Bolza,* SIAM J. Control, 10 (1972), pp. 691–715.

[24] ———, *Dual problems of Lagrange for arcs of bounded variation,* in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 155–192.

[25] ———, *Clarke's tangent cones and the boundaries of closed sets in* $\mathbb{R}^n$, Nonlinear Anal., Theory, Meth. & Appl., 3 (1979), pp. 145–154.

[26] ———, *Optimality conditions for convex control problems with nonnegative states and the possibility of jumps,* in Game Theory and Mathematical Economics, O. Moeschlin, ed., North-Holland, Amsterdam, 1981, pp. 339–349.

[27] ———, *Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization,* Math. Oper. Res., 6 (1981), pp. 424–436.

[28] ———, *Lipschitzian properties of multifunctions,* Nonlinear Anal., Theory, Meth. & Appl., 9 (1985), pp. 867–885.

[29] ———, *Extensions of subgradient calculus with applications to optimization,* Nonlinear Analysis, Theory, Methods & Applications, 9 (1985), pp. 665–698.

[30] ———, *Dualization of subgradient conditions for optimality,* Nonlinear Anal. Theory, Methods, Appl., 20 (1993), pp. 627–646.

[31] J. D. L. ROWLAND AND R. B. VINTER, *Dynamic optimization problems with free time and active state constraints,* Imperial College, 1990, preprint.

[32] R. A. WIJSMAN, *Convergence of sequences of convex sets, cones and functions* II, Trans. Amer. Math. Soc., 123 (1966), pp. 32–45.

[33] H. ZHENG, *Second-order necessary conditions for differential inclusion problems,* University of British Columbia, 1991, preprint.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 471–479, March 1994

008

AN EVASION GAME WITH AN INFINITE NUMBER OF STATES*

V. J. BASTON† AND F. A. BOSTOCK†

Abstract. The paper considers a two-person zero-sum discrete gunner-evader game Γ, which takes place on a semi-infinite line. The game is modeled as a recursive game with an infinite number of states. The theory of such games is far from being complete, and it is not even known whether they always have a solution. Even when they do have a solution, the optimal or ϵ-optimal strategies for the players may require a knowledge of past actions. It is shown that Γ has a solution and that the players have stationary optimal or ϵ-optimal strategies.

Key words. two-person game, zero-sum game, infinite recursive matrix game, time lag system

AMS subject classifications. 90D20, 90D05

1. Introduction.

Firing games in which there is a time lag have attracted attention over a period of years. This class of games occurs in different guises in a variety of situations such as a bomber-battleship problem [7], [9], [11], and [12] or a tank manoeuvring to avoid gunfire [13] (and [15]). More recently there has been considerable interest in problems in which an evader, moving on a discrete set of points, tries to avoid being hit by a gunner. Baston and Bostock [1], [2] have treated the case where the number of points is finite. Lee [16], [17] and Sakaguchi [19] have investigated the case where there is a safe point for the evader when he moves on a line, while Garnaev [10] deals with the situation where the evader moves on the infinite two-dimensional integer grid with a general set of safe points. In these papers the gunner has a given number of bullets to start with, but Bernhard, Colomb, and Papavassilopoulos [4] considered the situation in which the gunner has an unlimited supply.

The game we investigate in this paper is broadly similar to the game in [1], but it differs in the crucial aspect of having an infinite number of states. The theory of such games (i.e., infinite recursive games) is far from being complete. Indeed it is not yet known whether every bounded infinite recursive game always has a solution. In 1972 Orkin [18] presented this as an open question for games in which the components are matrices. Our search of the subsequent literature shows this important question to be as yet unanswered.

An introductory description of our game is as follows. Let A_r denote the point r of the x-axis where r is a nonnegative integer. An evader starts at some given point A_s and at discrete intervals of time $t = 1, 2, \ldots$ chooses to move to one of the points adjacent to him or stay where he is. A gunner with a single bullet may at each of the same discrete intervals of time either fire the bullet at one of the points A_r or hold his fire. It is assumed that the gunner always hits the point at which he aims and that the bullet takes one unit of time to reach its target. The payoff to the gunner is 1 if he hits the evader, μ (where $|\mu| < 1$) if he fires and misses, and 0 if he never fires. Although the above description superficially defines our game, the choices of strategy spaces for the players will play an important role when we come to model it. This is dealt with more fully in the next section.

*Received by the editors October 24, 1990; accepted for publication (in revised form) October 8, 1992.

†Faculty of Mathematical Studies, University of Southampton, Southampton SO9 5NH, United Kingdom.

2. Preliminary notions and the value for $\mu = 0$. We obtain a solution for our game by modeling it as a recursive matrix game Γ^μ with a countably infinite number of states $\Gamma_1^\mu, \Gamma_2^\mu, \ldots$. The fundamental paper on the theory of recursive games with a finite number n of states was written by Everett [8]. In that paper a strategy X for a player was defined as a sequence $X = X^1, X^2, \ldots$, where each term X^t was a sequence $X_1^t, \ldots, X_n^t$ with the significance that X_r^t is a mixed strategy to be used at time t when playing in state r. These strategies employed history only insofar as a player was aware of the stage t he had reached. We shall denote the sets of such strategies for player 1 and player 2 by $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively; the same symbols will be used for the corresponding sets of strategies in an infinite recursive matrix game. Strategies independent of t are said to be stationary.

In a later paper [18] Orkin introduced a wider class of strategies for the players (in a finite recursive matrix game) by taking a strategy for player 1 (player 2) as a probability distribution on the rows (columns) of each component game matrix, where the probability distribution could depend on the preceding sequence of moves (i.e., the history). The sets of such strategies for P_1 and P_2 will be denoted by $\mathcal{H}_1$ and $\mathcal{H}_2$; the same symbols will be used for the corresponding sets of strategies in an infinite recursive matrix game. We shall refer to members of $\mathcal{H}_1$ and $\mathcal{H}_2$ as history-remembering strategies. We shall also need the concept of a semi-Markov strategy; this is a history-remembering strategy that ignores history apart from acknowledging the starting component and the stage play has reached. The expectation corresponding to strategies X and Y is denoted by $E(X, Y)$. A bar over a symbol for a recursive game will be used if and only if history-remembering strategies are available. It is well known that for any finite recursive matrix game Λ there always exist stationary optimal or ϵ-optimal strategies. Orkin points out that such strategies are equally effective against history-remembering strategies, but it is now known that the restriction of being stationary is unnecessary, even in the infinite case (see Lemma 2).

We now set up the matrices for the component games of Γ^μ. Note that the point A_0 is of special significance in the game; if the evader moves to A_0 in the course of play, he can then either stay at A_0 or return to A_1. In this sense we can think of A_0 as a barrier. If the game starts at A_0, it is easy to see as follows that the value is $(1 + \mu)/2$. The gunner can achieve $(1 + \mu)/2$ by firing straightaway with equal probability at A_0 and A_1, whereas the evader can clearly hold the gunner down to this by, at each stage, staying where he is or moving to the right with equal probability. The interpretation of the component game Γ_r^μ is that the evader starts at A_r. Thus the component matrices are given by

$$\Gamma_1^\mu = \begin{pmatrix} 1 & \mu & \mu \\ \mu & 1 & \mu \\ \mu & \mu & 1 \\ (1+\mu)/2 & \Gamma_1^\mu & \Gamma_2^\mu \end{pmatrix}$$

and for $r = 2, 3, \ldots$ by

$$\Gamma_r^\mu = \begin{pmatrix} 1 & \mu & \mu \\ \mu & 1 & \mu \\ \mu & \mu & 1 \\ \Gamma_{r-1}^\mu & \Gamma_r^\mu & \Gamma_{r+1}^\mu \end{pmatrix}.$$

In Γ_r^μ the columns 1, 2, 3, respectively, represent the pure strategies for the evader in which he moves to the points A_{r-1}, A_r, A_{r+1}; the first three rows, respectively, represent the pure strategies for the gunner in which he fires at the points A_{r-1}, A_r, A_{r+1}, and row 4 represents the strategy of not firing.

At this point we find it convenient to restrict our attention to the case $\mu = 0$, and for brevity write Γ^0 as Γ. For each infinite real vector $W = (w_1, w_2, \ldots)$ let $M_r(W)$ denote the real matrix obtained by substituting w_i for each game component Γ_i that occurs in the matrix M_r of Γ_r. The value map V from infinite real vectors to infinite real vectors is defined by taking the rth component of $V(W)$ as the value of $M_r(W)$, regarded as an ordinary matrix game. Let U_0 denote the infinite zero vector, and for $k = 1, 2, \ldots$ define $U_k = (u_1^k, u_2^k, \ldots)$ by $U_k = V(U_{k-1})$. Since clearly $U_1 \geq U_0$, it follows from Lemma 3 that

(a) $\bar{\Gamma}$ has a value given by the limit $u = (u_1, u_2, \ldots)$ of U_k as $k \to \infty$,

(b) U is a fixed point of the value map,

(c) the gunner has an ϵ-optimal semi-Markov strategy,

(d) the evader has an optimal stationary strategy.

A standard argument shows that the sequence U_k is increasing. Following closely the proof of [1, Lem. 3] it is easily established that

(i) for $r \geq 1$, $k \geq 1$, $1/3 \leq u_r^k \leq 1/2$ and $u_r^k \geq u_{r+1}^k$, and

(ii) for $r \geq 1$, $k \geq 2$, $u_r^k = u_{r-1}^{k-1}/(1 + 2u_{r-1}^{k-1} - u_r^{k-1} - u_{r+1}^{k-1})$, where for all $k \geq 2$ we define $u_0^{k-1} = 1/2$.

We also note that for $r \geq 1$, $u_r^1 = 1/3$. From (i) for each $r = 1, 2, \ldots$ the sequence u_r^k converges (say to u_r), and these limits satisfy the recurrence

$$u_r = u_{r-1}/(1 + 2u_{r-1} - u_r - u_{r+1}),$$

where $u_0 = 1/2$, which gives

$$u_{r+1} = 2u_{r-1} - u_r + 1 - (u_{r-1}/u_r).$$

Note that from (i) above the sequence u_r is decreasing. Clearly if we know u_1, this recurrence will define the sequence u_r.

However, in several ways this disguises an unsatisfactory situation. First, the iteration technique does not directly enable us to find sharp bounds on u_1. Second, it gives no method of determining the limit of u_r as $r \to \infty$. Third, we do not know whether Γ has a solution. Finally (c) provides only an ϵ-optimal strategy for the gunner that is far from being stationary. We will remedy these deficiencies in the subsequent sections. To do so we will need the following lemmas.

LEMMA 1. *In the bounded infinite recursive matrix game $\Lambda = (\Lambda_1, \Lambda_2, \ldots)$ let $Y \in \mathcal{S}_2$ and $W \geq 0$ be a real vector such that for all $X \in \mathcal{S}_1$ and all t, $E^t(X, Y; W) \leq W$, then for all $X \in \mathcal{S}_1$, $E(X, Y) \leq W$. [The ith component of $E^t(X, Y; W)$ is the expectation when the players use X_i^t and Y_i^t in the ordinary matrix game $M_i(W)$.]*

LEMMA 2. *Let $\Lambda = (\Lambda_1, \Lambda_2, \ldots)$ be a bounded infinite recursive matrix game. Let W be a real vector, and X^* a strategy in $\mathcal{S}_1$ such that for all $Y \in \mathcal{S}_2$ $E(X^*, Y) \geq W$. Then for all $Y \in \mathcal{H}_2$ $E(X^*, Y) \geq W$.*

We note Lemma 2 (together with the analogous result for player 2) implies that whenever a bounded infinite recursive matrix game Λ has a solution, $\bar{\Lambda}$ has the same solution.

LEMMA 3. *Let Γ be a bounded (possibly infinite) recursive finite matrix game satisfying $V(0) \geq 0$, where V is the value map. Then $\bar{\Gamma}$ has a value that is a fixed point of the value map and is the limit as $k \to \infty$ of $V^k(0)$. Player 1 has an ϵ-optimal semi-Markov strategy and player 2 has an optimal stationary strategy.*

Lemma 1 is an application of [5, Thm. 2] and Lemma 2 follows easily from [20, Prop. 9.1]. Lemma 3 is a special case of results in [14], note in particular Theorem 2 and Remark 1.

3. Evader strategies for the case $\mu = 0$. We first determine a stationary evader strategy, which, against any $X \in \mathcal{S}_1$, will hold the gunner's expectation down to at most u as defined under (a) in the previous section. Define the stationary strategy Y^* by

$$Y_r^* = (1 - u_r - u_{r+1}, u_{r-1}, u_{r-1})/(1 + 2u_{r-1} - u_r - u_{r+1}) \text{ for } r = 1, 2, \ldots,$$

where $u_0 = 1/2$. Now u is a fixed point of the value map, and in particular u_r is the value of the ordinary matrix game $M_r(u)$. It is easy to verify that the strategy Y_r^* is optimal for the minimizing player in the game $M_r(u)$, and then as a direct consequence of Lemma 1 the evader strategy Y^* will hold the gunner's expectation down to at most u against any $X \in \mathcal{S}_1$. Using the analogous result to Lemma 2 for the evader we see that Y^* is equally effective against any $X \in \mathcal{H}_1$.

We now prove the intuitively obvious result that $u_r \to 1/3$ as $r \to \infty$, which we will need in the next section. To do this we will produce a sequence $\alpha_0, \alpha_1, \ldots$, where for each $r = 0, 1, 2, \ldots$ the evader, when starting at A_r, is able to hold the gunner's expectation down to at most α_r. If the evader starts at the barrier A_0, then he is able to hold the gunner's expectation down to $1/2$, and we take $\alpha_0 = 1/2$. Suppose $Y(r) \in \mathcal{H}_2$ is a strategy for the evader, which, when he starts at A_r, holds down the gunner's expectation (using $X \in \mathcal{H}_1$) to at most α_r, where $1/3 < \alpha_r \leq 1/2$. Consider the strategy $Y^* \in \mathcal{H}_2$ defined as follows. If the evader starts at A_j with $j \neq r + 1$, he never moves. When the evader starts at A_{r+1} and finds himself at a point A_j, never having been at A_r, he moves to the points A_{j-1}, A_j, A_{j+1}, respectively, with probabilities $1 - 2p, p, p$, where $p = \sqrt{(\alpha_r/3)}$; however, should he arrive at A_r he subsequently employs $Y(r)$ as if he were starting at A_r at time $t = 1$. Note that $1/3 < p < 1/2$ and $\alpha_r > p$. Because $1 - 2p < p$, the theory of random walks shows that the probability that the evader is ever at the point A_r when starting at A_{r+1} is $(1 - 2p)/p$. Hence when the evader uses Y^*, the gunner's expectation is not more than $\alpha_r\big((1 - 2p)/p\big) + p(1 - (1 - 2p)/p) = 2\sqrt{(3\alpha_r)} - 2\alpha_r - 1$, which we define to be α_{r+1}. Now

(a) $\alpha_r - \alpha_{r+1} = (\sqrt{(3\alpha_r)} - 1)^2 > 0$, and

(b) $\alpha_{r+1} - 1/3 = 6(2/3 - \sqrt{(\alpha_r/3)})(\sqrt{(\alpha_r/3)} - 1/3) > 0$.

Thus $1/3 < \alpha_{r+1} \leq 1/2$, which means we have obtained our sequence $\alpha_0, \alpha_1, \ldots$. Since the sequence is decreasing and bounded below, it converges, say, to α. From (a) it follows that $\alpha = 1/3$. Since u is the value of the game $\bar{\Gamma}$ $\alpha_r \geq u_r (\geq 1/3)$, and it follows that $\lim_{r\to\infty} u_r = 1/3$.

4. Gunner strategies for the case $\mu = 0$. In this section we will obtain a stationary strategy for the gunner that ensures him an expectation of at least $u - \epsilon$ against any $Y \in \mathcal{H}_2$; to do this we find it convenient to define certain other recursive games. Let $N \geq 2$ be an integer and $\Gamma(N) = (\Gamma_1(N), \Gamma_2(N), \ldots, \Gamma_N(N))$ be the recursive game given by

$$\Gamma_1(N) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & \Gamma_1(N) & \Gamma_2(N) \end{pmatrix},$$

for $r = 1, 2, \ldots, N - 1$

$$\Gamma_r(N) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \Gamma_{r-1}(N) & \Gamma_r(N) & \Gamma_{r+1}(N) \end{pmatrix},$$

and

$$\Gamma_N(N) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \Gamma_{N-1}(N) & \Gamma_N(N) & 1/3 \end{pmatrix}.$$

The interpretation of the game $\Gamma_r(N)$ is as usual, that the evader starts at A_r and there is a barrier at A_0. There is, however, the boundary condition that should the evader, at some stage, be at the point A_N and choose to move to the point A_{N+1}, whilst the gunner chooses not to fire, then the game terminates with a payoff of $1/3$. If we iterate the zero vector k times under the standard value map for $\Gamma(N)$ to form a vector $(v_1^k(N), v_2^k(N), \ldots, v_N^k(N))$, then, because the game is recursive and all the number entries in all the game components are nonnegative, it follows [8] that for each $r = 1, 2, \ldots, N$ the sequence $v_r^k(N)$ is increasing and converges to the value of the game $\Gamma_r(N)$, say, $v_r(N)$.

Furthermore, following the proof of [1, Lem. 3] it is easily shown that

(i) For $r = 1, 2, \ldots, N$; $k \geq 1$, $1/3 \leq v_r^k(N) \leq 1/2$.

(ii) For $r = 1, 2, \ldots, N-1$; $k \geq 1$, $v_r^k(N) \geq v_{r+1}^k(N)$.

(iii) For $r = 1, 2, \ldots, N$; $k \geq 2$, $v_r^k(N) = v_{r-1}^{k-1}(N)/(1 + 2v_{r-1}^{k-1}(N) - v_r^{k-1} - v_{r+1}^{k-1}(N))$, where for all $k \geq 2$ we define $v_0^{k-1}(N) = 1/2$ and $v_{N+1}^{k-1}(N) = 1/3$.

We also note that for $r = 1, 2, \ldots, N$ $v_r^1(N) = 1/3$. Using the fact that when $x \leq y$

$$\operatorname{val}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ x_1 & x_2 & x_3 \end{pmatrix} \leq \operatorname{val}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ y_1 & y_2 & y_3 \end{pmatrix}$$

and the above results concerning u_r^k and $v_r^k(N)$, it is not hard to see that,

(a) for each r and all $N > r$, $u_r^k \geq v_r^k(N+1) \geq v_r^k(N)$, and

(b) provided $N \geq r + k$, $u_r^k = v_r^k(N)$.

Thus $v_r^k(N)$ is monotonic increasing as a function of N and $u_r \geq v_r(N)$ for all $N > r$; a simple reductio ad absurdum argument now establishes that $u_r = \lim_{N\to\infty} v_r(N)$. We proved in §3 that $u_r \to 1/3$ as $r \to \infty$. Thus given $\epsilon > 0$, let K be such that $u_r - 1/3 < \epsilon$ for $r > K$. Now for each $r = 1, 2, \ldots, K$ choose $K_r > r$ so that $u_r - v_r(N) < \epsilon/2$ whenever $N \geq K_r$, and let $M = \max\{K_1, K_2, \ldots, K_K\}$. We may take a stationary $(\epsilon/2)$-optimal strategy $X_{(M)}^* = X_1^*, X_2^*, \ldots, X_M^*$ for the gunner in the game $\Gamma(M)$ by using strategies pioneered by Everett [8]

Define the stationary strategy X^* for the gunner in Γ by, $X^* = X_1^*, X_2^*, \ldots$, where for each $i > M$ $X_i^* = (1, 1, 1, 0)/3$. If the gunner uses X^* and the evader starts at A_r, his expectation against any $Y \in \mathcal{S}_2$ is $1/3 > u_r - \epsilon$ when $r > M$ and at least $v_r(M) - \epsilon/2 > u_r - (\epsilon/2) - (\epsilon/2) = u_r - \epsilon$ when $r \leq M$. Lemma 2 now ensures that X^* maintains $u - \epsilon$ against any $Y \in \mathcal{H}_2$. The results of this and the previous section show that both Γ and $\bar{\Gamma}$ have a solution, with the gunner having a stationary ϵ-optimal strategy and the evader a stationary optimal strategy.

Note. Later we will need to ensure that in X_r^* the probability of not firing is strictly less than one. To see this, suppose the probability of not firing were one. Then, starting at A_r, the evader could clearly hold the gunner's expectation down to zero. However, $v_r(N)$ is at least $1/3$.

5. The game for $|\mu| < 1$. From the case $\mu = 0$ we now deduce that the value of the game Γ for $\mu \geq -1/2$, starting at A_r, is $\mu + (1 - \mu)u_r$. When starting at A_0

the result is trivial. With $\mu = 0$ we may clearly interpret the gunner's expectation as the probability of his hitting the evader. Given $\epsilon > 0$, suppose the gunner uses the strategy X^* of §4, which guarantees him $u_r - \epsilon$ when the evader starts at the point A_r. Let the evader adopt a strategy $Y(\in \mathcal{H}_2)$ and the game start at the point $A_r \neq A_0$. Let H_n, M_n, B_n, respectively, be the probabilities that up to time $t = n$ the gunner has hit the evader, fired and missed, or not fired. If E_n denotes the accumulated expectation up to $t = n$, then $E_n = H_n + \mu M_n$. Since $H_n + M_n + B_n = 1$, we have $M_n \leq 1 - H_n$ so that when $\mu \leq 0$ $E_n \geq H_n + \mu(1 - H_n) = (1 - \mu)H_n + \mu$. We may choose n_1 so that for all $n \geq n_1$ $H_n \geq u_r - \epsilon$, whence $E_n \geq \mu + (1 - \mu)u_r - \epsilon(1 - \mu)$.

Now consider the case $\mu \geq 0$. We have $E_n = H_n + (1 - H_n - B_n)\mu = (1 - \mu)H_n + (1 - B_n)\mu$. Since the game starts at A_r, there are at most $r + 2M$ points that can be visited by the evader up to and including time M. Under X^*, at each of these points the probability of not firing is strictly less than 1, according to the Note. Further, under X^*, the gunner fires at time $M + 1$. Hence there is an $\alpha < 1$ such that at each stage the probability of the gunner not firing is at most α. Hence $B_n \leq \alpha^n \to 0$ as $n \to \infty$. It now easily follows that for sufficiently large n,

$$E_n \geq (1 - \mu)(u_r - \epsilon) + (1 - \epsilon)\mu = \mu + (1 - \mu)u_r - \epsilon.$$

Let the evader use the strategy Y^* of §3, which holds the gunner's expectation down to u_r when he starts at the point A_r. Now $u_r + (1 - u_r)\mu \geq 0$, since $u_r \geq 1/3$ and we are taking $\mu \geq -1/2$. Since the expectation of a strategy pair (X, Y) in a matrix game $(\mu + (1 - \mu)a_{rs})$ is $\mu + \{(1 - \mu)$times the expectation of (X, Y) in the matrix game $(a_{rs})\}$, the conditions of Lemma 1 are satisfied by $Y = Y^*$ and $W = (\mu + (1 - \mu)u_r)$ in Γ^μ because they are satisfied by $Y = Y^*$ and $W = (u_r)$ in Γ. Hence the value of the game for $\mu \geq -1/2$ is the vector $(\mu + (1 - \mu)u_r)$.

We now turn our attention to the range $-1 < \mu < -1/2$. The methods of §2 go through in the present case, and it is easily seen that the value (w_r) of the game $\bar{\Gamma}^\mu$ satisfies the recurrence

$$w_r = \frac{\mu[w_{r-1} - w_r - w_{r+1}] + w_{r-1}}{1 - \mu + 2w_{r-1} - w_r - w_{r+1}}, \quad r = 1, 2 \ldots$$

and boundary condition $w_0 = (1 + \mu)/2$. A stationary strategy that guarantees $(w - \epsilon)$ (against any $Y \in \mathcal{H}_2$) for the gunner can now be obtained using the arguments of §4. In the description of $\Gamma(N)$ the zeros are replaced by μ's, the $1/2$ by $(1 + \mu)/2$, and the $1/3$ by 0. Lemma 3 shows that the corresponding sequence $v_r^k(N)$ converges to the value of the game. Further obvious modifications now lead to an appropriate strategy.

As in §3 the evader has a stationary strategy, which against any $X \in \mathcal{H}_1$, will hold the gunner's expectation down to (w_r). Hence the value of the game for $-1 < \mu < -1/2$ is (w_r). We may prove that $w_r \to 0$ as $r \to \infty$ in a manner similar to that for the case $\mu = 0$ by taking $p = -\mu/(1 - \mu)$, independent of the starting point A_r.

6. Calculation of u_1. For further information regarding the nature of the recurrence relation on u_r we go to the related map f given by $f(x, y) = (y, 2x - y + 1 - (x/y))$, $y \neq 0$. It is easy to see that the fixed points of f are precisely those points (x, x) with $x \neq 0$. The matrix of the derivative $df(x, y)$ of f at (x, y) is

$$\begin{bmatrix} 0 & 2 - (1/y) \\ 1 & (x/y^2) - 1 \end{bmatrix}$$

so that when p and q are small, at a fixed point we have,

$$f((x,x)+(p,q)) \simeq (x,x)+(p,q)\begin{bmatrix} 0 & 2-(1/x) \\ 1 & (1/x)-1 \end{bmatrix}.$$

The eigenvalues of $df(x,x)$ are given by $\lambda^2 - ((1/x)-1)\lambda + (1/x) - 2 = 0$, which yields $\lambda = 1$ and $(1/x)-2$. The eigenvalue 1 corresponds of course to the line of fixed points. Let $\lambda_0 = (1/x) - 2$, and suppose $x \neq 1/2$ so that $\lambda_0 \neq 0$. Then, provided $q = \lambda_0 p$, we have $f((x,x)+(p,q)) \simeq (x,x) + \lambda_0(p,q)$. With the exception of $x = 1/2$ and 0, corresponding to the fixed point (x,x), there is a stable or unstable manifold accordingly as $|\lambda_0| < 1$ or $|\lambda_0| > 1$. The tangent to the stable or unstable manifold through (x,x) has gradient λ_0. Our main concern is with the region $0 < x, y < 1/2$, where for $1/3 < x < 1/2$ the manifold is stable and for $0 < x < 1/3$ unstable. Numerical investigations show that in the region in which we are interested the phase plane appears to be foliated by a family of invariant curves as indicated in Fig. 1.

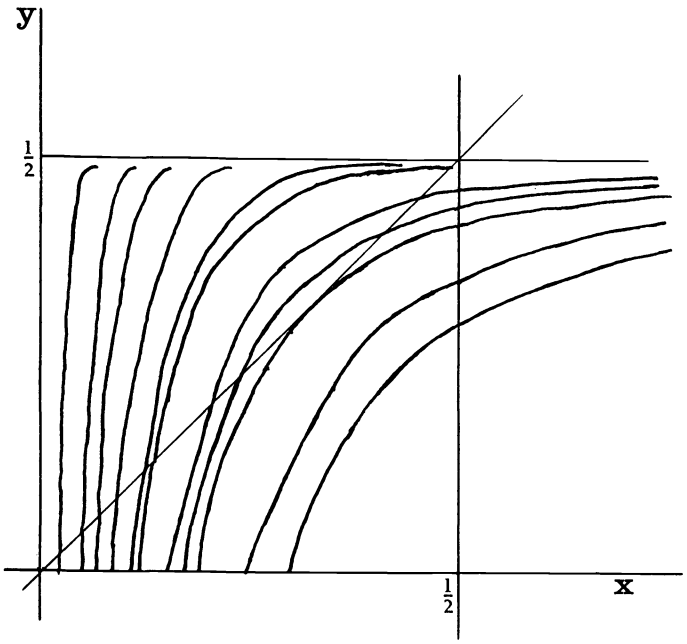


FIG. 1

In Fig. 1 the two right-hand curves are neither stable nor unstable manifolds as far as can be judged, but we put them forward as invariant curves under f and as particular members of a general family. Of particular interest for us is the curve that touches the line of fixed points at $x = 1/3$, where the eigenvalue $\lambda_0 = 1$. From the phase map standpoint it seems evident (but not conclusive) that this curve is the only one that touches the line of fixed points at $(1/3, 1/3)$. To see that this is in fact so in the region in which we are interested, we will prove that there is a unique solution to the recurrence relation on u_r with boundary conditions $u_0 = 1/2$, $1/3 < u_1 < 1/2$, and $\lim_{r\to\infty} u_r = 1/3$. First of all assume that the sequence (z_r) is another solution

with $z_1 > u_1$. Let $z_r = u_r + \epsilon_r$ so that $\epsilon_0 = 0$ and $\epsilon_1 > 0$. We show by induction that $\epsilon_t \geq \epsilon_{t-1}$ for all $t \geq 1$. The result certainly holds for $t = 1$. Suppose it holds for $t = r$. Then

$$\begin{aligned} z_{r+1} &= 2(u_{r-1} + \epsilon_{r-1}) - (u_r + \epsilon_r) + 1 - (u_{r-1} + \epsilon_{r-1})/(u_r + \epsilon_r) \\ &= u_{r+1} + 2\epsilon_{r-1} - \epsilon_r + (u_{r-1}/u_r) - (u_{r-1} + \epsilon_{r-1})/(u_r + \epsilon_r). \end{aligned}$$

So $\epsilon_{r+1} = 2\epsilon_{r-1} - \epsilon_r + (u_{r-1}\epsilon_r - u_r\epsilon_{r-1})/\{u_r(u_r + \epsilon_r)\}$

$$\begin{aligned} &\geq 2\epsilon_{r-1} + (\epsilon_r - \epsilon_{r-1})/(u_r + \epsilon_r) \qquad \text{since } u_{r-1} \geq u_r \\ &\geq 2\epsilon_{r-1} - \epsilon_r + 2\epsilon_r - 2\epsilon_{r-1} \qquad \text{since } u_r + \epsilon_r = z_r \leq 1/2 \text{ and } \epsilon_r \geq \epsilon_{r-1} \\ &= \epsilon_r. \end{aligned}$$

This completes the induction; it follows that $\epsilon_r \geq \epsilon_1$ for $r \geq 1$, and this contradicts $\lim_{r\to\infty} z_r = 1/3$. A similar method shows there is no solution with $z_r < u_1$, and hence (u_r) is the unique solution.

Hence there exists a unique number ρ with $1/3 \leq \rho \leq 1/2$ such that the orbit of the point $(1/2, \rho)$ converges to $(1/3, 1/3)$. Trial and error methods involving the phase map of the recurrence relation (with boundary conditions $u_0 = 1/2$ and $\lim_{r\to\infty} = 1/3$) have given the inequality, $0.423638 < u_1 < 0.42365$.

7. Conclusions. The sequence $u_0, u_1, \ldots$ can be thought of merely as stemming (uniquely) from the defining sequence u_r^k of §2. The sequence (u_r) was also easily seen to converge, and the determination of the limit could be regarded as a problem of pure mathematics. However, we were unable to prove in a direct manner that $\lim_{r\to\infty} u_r = 1/3$, and finally employed the game-theoretic argument used in this paper. It would be nice to have a simple direct proof of what is after all, from a game-theoretic standpoint, an intuitively obvious result. It may be of interest to note that a similar situation concerning boundary conditions arises in the general context of a simple random walk with one absorbing barrier. Here results are often obtained from a two-barrier analysis by a limiting procedure (see [6], for example). It is taken as obvious, presumably on the grounds of intuition, that this limiting process does in fact give the required probability.

When the gunner has j bullets there are two basic generalizations of Γ^μ. In one the gunner has a payoff of 1 for each hit and μ for each miss; for the special case $\mu = 0$ the gunner is effectively trying to maximize the total number of hits. In the other the gunner maximizes his chances of a single hit (this is in fact a generalization of the game Γ^0; a generalization in this direction with a general μ has little natural appeal). The methods of this paper can be used to deal with these problems by employing a dynamic programming technique. Note that, having obtained the value for one bullet, the case for two bullets can be modeled by component matrices having the same form as for one bullet (i.e., number (game component) entries where there were number (game component) entries), and so on.

In the case of a finite recursive matrix game there is always a solution in stationary strategies and we have found that to be true for our game. It would be interesting to know whether or not a bounded infinite recursive finite matrix game (with or without a bar!) that has a solution, always has one in stationary strategies. An example of a bounded infinite recursive infinite matrix game Λ such that $\bar{\Lambda}$ has a solution but Λ does not is given in [3]. We have also found an example of such a game Λ, where Λ has a solution but one of the players does not have an optimal or ϵ-optimal stationary strategy.

Acknowledgments. We would like to thank David Chillingworth and David Whitley for helpful conversations concerning §6. We would like to thank the referees

for pointing out that Lemmas 1–3 were already in the literature and for providing us with the appropriate references. One of the referees also pointed out that the game can be formulated as a dynamic game stopped at random time (depending on the position of the state), and the cost is incurred only at the final (stopping) time.

REFERENCES

[1] V. J. Baston and F. A. Bostock, *An evasion game with barriers*, SIAM J. Control Optim., 26 (1988), pp. 1099–1105.

[2] ———, *An evasion game on a finite tree*, SIAM J. Control Optim., 28 (1990), pp. 671–677.

[3] ———, *Infinite deterministic graphical games*, SIAM J. Control Optim., 31 (1993), pp. 1623–1629.

[4] P. Bernhard, A.-L Colomb, and G. P. Papavassilopoulos, *Rabbit and hunter game: Two discrete stochastic formulations*, Comput. Math. Appl., 13 (1987), pp. 205–225.

[5] D. Blackwell, *Positive dynamic programming*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, L. M. Le Cam and J. Neyman, eds., University of California Press, Berkeley and Los Angeles, 1967, pp. 415–418.

[6] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, Chapman & Hall, London, 1965.

[7] L. E. Dubins, *A discrete evasion game*, in Contributions to the Theory of Games III, M. Dresher, A. W. Tucker and P. Wolfe, eds., Ann. Math. Stud. Vol. 39, Princeton University Press, Princeton, NJ, 1957, pp. 231–255.

[8] H. Everett, *Recursive games*, in Contributions to the Theory of Games III, M. Dresher, A. W. Tucker and P. Wolfe, eds., Ann. Math. Stud. Vol. 39, Princeton University Press, Princeton, NJ, 1957, pp. 47–78.

[9] T. Ferguson, *On discrete evasion games with a two-move information lag*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, L. M. Le Cam and J. Neyman, eds., University of California Press, Berkeley and Los Angeles, 1967, pp. 453–462.

[10] A. Yu. Garnaev, *On a discrete game in a plane with delayed information*, Automat. Remote Control, 50 (1989), pp. 1480–1487.

[11] R. Isaacs, *The problem of aiming and evasion*, Naval Res. Logist. Quart., 2 (1955), pp. 47–67.

[12] S. Karlin, *An infinite move game with a lag*, in Contributions to the Theory of Games III, M. Dresher, A. W. Tucker and P. Wolfe, eds., Ann. Math. Stud. Vol. 39, Princeton University Press, Princeton, NJ, 1957, pp. 257–272.

[13] P. R. Kumar, *Optimal mixed strategies in a dynamic game*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 743–749.

[14] P. R. Kumar and T. H. Shiau, *Existence of value and randomized strategies in zero-sum discrete-time stochastic dynamic games*, SIAM J. Control Optim., 19 (1981), pp. 617–634.

[15] ———, *Zero-sum dynamic games*, in Control and Dynamic Systems, C. T. Leondes, ed., Academic Press, New York, 1981, pp. 345–378.

[16] K. T. Lee, *A firing game with a time lag*, J. Optim. Theory Appl., 41 (1983), pp. 547–558.

[17] ———, *An evasion game with a destination*, J. Optim. Theory Appl., 41 (1983), pp. 359–372.

[18] M. Orkin, *Recursive matrix games*, J. Appl. Probab., 9 (1972), pp. 813–820.

[19] M. Sakaguchi, *A pursuit game on a line with a bunker for the evader*, Math. Japonica, 31 (1986), pp. 449–461.

[20] S. E. Shreve and D. P. Bertsekas, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 480–500, March 1994

009

LINEAR PROGRAMMING AND AVERAGE OPTIMALITY OF MARKOV CONTROL PROCESSES ON BOREL SPACES—UNBOUNDED COSTS*

ONÉSIMO HERNÁNDEZ-LERMA† AND JEAN B. LASSERRE‡

Abstract. This paper is concerned with the linear programming formulation of Markov control processes with *Borel* state and action spaces, and the *average cost* (*AC*) criterion. The one-stage cost function may be *unbounded*. A linear program EP and its dual, EP^*, are introduced. Their values, inf EP and sup EP^*, bound the value (say, inf AC) of the AC problem, i.e., sup $EP^* \leq$ inf $AC \leq$ inf EP. Conditions are provided for the existence of *no duality gap*, viz., sup EP^*= inf EP, and also for *strong duality*, so that both EP and EP^* are solvable and their optimal values satisfy max EP^* = min EP. The latter implies (i) the existence of an AC-optimal control policy and that (ii) the AC optimality equation holds almost everywhere. These results are applied to a general vector-valued, additive-noise system with quadratic costs.

Key words. (discrete-time) Markov control processes, average cost criterion, linear programming (in general vector spaces), strong duality

AMS subject classifications. 93E20, 90C40

1. Introduction.

The linear programming (LP) formulation of Markov control problems—or stochastic dynamic programs—has been studied since the early 1960s, and it has proved to yield useful insight into some control problems. However, most of the related literature is concentrated on problems with denumerable (mainly finite) state and control spaces, which excludes of course many important applications (e.g., in engineering, economics, and operations research), where these spaces are nondenumerable (e.g., R^n).

The main objective of this paper is to study the LP formulation of *average cost* (*AC*) Markov control processes (MCPs) on *Borel* spaces, allowing *unbounded* one-stage costs. The idea is to introduce the linear program EP and its dual, EP^*, and to show that the value of the AC control problem, call it inf AC, satisfies

$$\sup EP^* \leq \inf AC \leq \inf EP. \tag{1.1}$$

(A precise formulation is given in §§4, 5.) Thus, if there is *no duality gap* for the LP problems, so that

$$\sup EP^* = \inf EP, \tag{1.2}$$

then their common value yields inf AC. Now, if (1.2) holds and, moreover, there are optimal solutions for EP and EP^* (so that inf EP = min EP and sup EP^* = max EP^*), then we have

$$\max EP^* = \min EP. \tag{1.3}$$

* Received by the editors March 26, 1992; accepted for publication (in revised form) December 9, 1992. This work is part of a joint research project sponsored by Consejo Nacional de Ciencia y Techología (México) and Centre National de Recherche Scientifique (France).

† Departemento de Matemáticas, Centro de Investigacion y de Estudios Aranzados-Instituto Politecnico Nacional, Apartado Postal 14–740, 07000 México D.F., México. The work of this author was supported by CONACYT grant 1332-E9206.

‡ Laboratoire d'Automatique et d'Analyse des Systems–Centre National de la Recherche Scientifique, 7 Avenue du Colonel Roche, 31077 Toulouse Cédex, France.

Therefore, in addition to getting the value inf AC, we also get an *optimal policy*. When (1.3) holds we say—following the terminology of [2]—that the *strong duality* condition holds. We illustrate (1.1)–(1.3) with a nonlinear, additive-noise, vector system of the form

$$x_{t+1} = G(x_t, a_t) + \xi_t, \quad t = 0, 1, \ldots, \tag{1.4}$$

with a quadratic cost function

$$c(x, a) = x'Hx + a'Ra \qquad (\text{"prime" denotes transpose}). \tag{1.5}$$

We also study the relation between AC, EP, EP^*, and a solution to the (so-called) AC *optimality equation* (OE). It is shown, e.g., that (1.3) implies that the OE is satisfied almost everywhere with respect to a probability measure related to an optimal solution of EP.

This paper is organized as follows: The basic MCP is introduced in §2, together with the usual AC optimality criterion and two stronger forms of optimality, namely, "strong optimality" and "F-strong optimality" (see Definition 2.3). In §3 we introduce the notion of "canonical triplet" and show that a "canonical policy" is AC-optimal, strong optimal, and F-strong optimal. This is done under condition (3.1)—weakened in §4—which allows a class of unbounded costs and thus extends the work of several authors [8], [27], [34]; see [3, §6] for a self-contained presentation of canonical triplets. Section 4 begins with the formulation of the linear programs EP and EP^*, following closely the approach of Anderson and Nash [2] for LP in general vector spaces. We also introduce the basic Assumptions 4.2 and 4.3, on which most of our results (1.1)–(1.3) are based. (It is worth noting that, in contrast to the "bounding" condition (3.1), Assumption 4.2 allows unbounded functions also in the control (or action) variable a.) Assumptions 4.2 and 4.3 yield, e.g., the standard LP results of consistency, weak duality, and complementary slackness (Proposition 4.1). For MCPs with Borel state space and *bounded* costs these results have also been obtained by other authors [11], [16], [32]. However, except for a previous report [23] dealing with *denumerable* state MCPs, the remainder of §4, as well as §5—where we give conditions for the strong duality (1.3)—seems to be new, at least in the context of Borel spaces/unbounded costs. Finally, in §6 we give conditions on (1.4)–(1.5) under which all the results in §§5 and 6 are valid.

There is an extensive literature on the LP approach to MCPs, as can be seen in the recent, excellent surveys [3], [28]; for earlier references see, e.g., [12], [22]; for constrained MCPs see, e.g., [1]. As already noted, though, virtually all of these works deal with finite or, perhaps, countably infinite spaces.

Remark 1.1.

Notation. Let S be a Borel space (i.e., a Borel subset of a complete and separable metric space). Then $\mathcal{B}(S)$ denotes the sigma-algebra of Borel subsets of S, and $\mathcal{P}(S)$ stands for the family of all probability measures on $\mathcal{B}(S)$. "Measurable" always means "Borel-measurable." If S and T are Borel spaces, a *stochastic kernel* [3], [14], [19] on S given T is a function $P(\cdot \mid \cdot)$ such that $P(\cdot \mid t)$ is a probability measure on S for each fixed $t \in T$, and $P(B \mid \cdot)$ is a measurable function on T for each fixed $B \in \mathcal{B}(S)$. The family of all stochastic kernels on S given T is denoted by $\mathcal{P}(S \mid T)$.

2. Problem definition.

The Markov control model. Let (X, A, Q, c) be a Markov control model with state space X, control (or action) set A, transition law Q, and one-stage cost function

c satisfying the following conditions. Both X and A are Borel spaces. To each $x \in X$ is associated a nonempty set $A(x) \in \mathcal{B}(A)$, which represents the set of admissible control actions when the system is in state x. The set

$$K := \{(x,a) \,|\, x \in X, \quad a \in A(x)\} \tag{2.1}$$

of admissible state-action pairs is assumed to be a Borel subset of $X \times A$. The transition law $Q(B \,|\, x,a)$, where $B \in \mathcal{B}(X)$ and $(x,a) \in K$, is a stochastic kernel on X given K, i.e., $Q \in \mathcal{P}(X \,|\, K)$ (see Remark 1.1). Finally, the one-stage cost c is a measurable function on K bounded from below. In fact, without loss of generality, we will assume that c is *nonnegative*.

The above Markov control model is standard [3], [14], [19].

DEFINITION 2.1. F denotes the set of all measurable functions $f : A \to X$ such that $f(x) \in A(x)$ for all $x \in X$, and Φ stands for the set of all stochastic kernels $\varphi \in \mathcal{P}(A \,|\, X)$ satisfying the constraint $\varphi(A(x) \,|\, x) = 1$ for all $x \in X$.

We will identify a function $f \in F$ with the stochastic kernel $\varphi \in \Phi$ for which, for every $x \in X$, $\varphi(\cdot \,|\, x)$ is the probability measure concentrated at $f(x)$. Thus we regard F as a subset of Φ. We will *assume* that F (hence Φ) is nonempty. (Equivalently, the set K in (2.1) contains the graph of a measurable map.) This condition holds, for instance, if the sets $A(x)$ are closed and the multifunction (or set-valued mapping) $x \to A(x)$ is measurable in the sense that the set $\{x \,|\, A(x) \cap C \neq \phi\}$ is in $\mathcal{B}(X)$ for every closed $C \subset A$. (See, e.g., [4] for a proof of this fact and related results.)

Consider the history spaces $H_0 := X$ and $H_t := K \times H_{t-1}$ if $t = 1, 2, \ldots$. An element h_t of H_t is a vector of the form $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$, where $(x_n, a_n) \in K$ for all $n = 0, \cdots, t-1$, and $x_t \in X$.

DEFINITION 2.2. A control policy is a sequence $\delta = \{\delta_t\}$ of stochastic kernels $\delta_t \in \mathcal{P}(A \,|\, H_t)$ satisfying the constraint

$$\delta_t\,(A(x_t) \,|\, h_t) = 1 \qquad \forall\, h_t \in H_t,\ t \geq 0.$$

The set of all policies is denoted by Δ. A control policy $\delta = \{\delta_t\}$ is said to be a *relaxed* (or *randomized stationary*) *policy* if there exists $\varphi \in \Phi$ such that

$$\delta_t(\cdot \,|\, h_t) = \varphi(\cdot \,|\, x_t) \qquad \forall\, h_t \in H_t, t \geq 0,$$

where Φ is the set in Definition 2.1. Finally, a policy $\delta = \{\delta_t\}$ is said to be a *nonrandomized stationary policy* (or briefly *stationary policy*) if there exists $f \in F$ such that $\delta_t(\cdot \,|\, h_t)$ is concentrated at $f(x_t)$ for all $h_t \in H_t$ and $t \geq 0$.

Following a standard convention, we will identify F (respectively, Φ) with the set of all stationary (respectively, relaxed) policies. Thus we may write $F \subset \Phi \subset \Delta$.

Let $(\Omega, \mathcal{F})$ be the measurable space that consists of the sample space $\Omega := (X \times A)^\infty$ and the corresponding product σ-algebra $\mathcal{F}$. Then for each policy δ and "initial distribution" $\nu \in \mathcal{P}(X)$, a probability P_ν^δ and a stochastic process $\{(x_t, a_t), \quad t = 0, 1, \ldots\}$ are defined on $(\Omega, \mathcal{F})$ in a canonical way [3], [8], [14], [19], where x_t and a_t represent the state and the control action at time t, respectively. The expectation operator with respect to P_ν^δ is denoted by E_ν^δ. If ν is the unit mass at (the initial state) x, then we write P_ν^δ and E_ν^δ as P_x^δ and E_x^δ, respectively.

Performance criteria. Given a policy $\delta \in \Delta$, an initial distribution ν, and a measurable function $h : X \to R$, define

$$J_0\,(\delta, \nu, h) := E_\nu^\delta h(x_0),$$

and for $n \geq 1$,

$$J_n(\delta, \nu, h) := E_\nu^\delta \left[\sum_{t=0}^{n-1} c(x_t, a_t) + h(x_n) \right], \tag{2.2}$$

assuming that the expectations are well defined (which will be the case under the assumptions given below). Equation (2.2) represents the total expected cost for an n-stage control problem with *terminal cost* function h. If $h(\cdot) \equiv 0$, then we write $J_n(\delta, \nu, h)$ as $J_n(\delta, \nu)$, i.e.,

$$J_n(\delta, \nu) := E_\nu^\delta \left[\sum_{t=0}^{n-1} c(x_t, a_t) \right]. \tag{2.3}$$

If the initial distribution ν is concentrated at a point x, we write the above functions as $J_n(\delta, x, h)$ and $J_n(\delta, x)$.

The average cost (or *AC*) problem. The main problem we are concerned with is the minimization of the long-run expected *average cost* (*AC*) per unit time defined as

$$J(\delta, \nu) := \limsup_n n^{-1} J_n(\delta, \nu). \tag{2.4}$$

Thus, denoting by J^* the *AC optimal value function*, i.e.,

$$J^*(\nu) = \inf_\delta J(\delta, \nu), \qquad \nu \in \mathcal{P}(X), \tag{2.5}$$

the *AC* problem is to find a policy δ^* such that

$$J(\delta^*, \nu) = J^*(\nu) \qquad \forall \nu \in \mathcal{P}(X), \tag{2.6a}$$

that is,

$$J(\delta^*, \cdot) \leq J(\delta, \cdot) \qquad \forall \delta \in \Delta. \tag{2.6b}$$

If (2.6) holds, then δ^* is said to be an (average cost) *optimal* policy. We also define the *value* $\inf_\Delta AC$ of the *AC* problem as

$$\inf_\Delta AC := \inf_\nu J^*(\nu) = \inf_\nu \inf_\delta J(\delta, \nu), \tag{2.7}$$

and a pair (δ^*, ν^*) consisting of a policy δ^* and an initial distribution ν^* is said to be a *minimum pair* [15], [21] if $J(\delta^*, \nu^*) = \inf_\Delta AC$.

In §§4 and 5 we study the *AC* problem via a dual pair of linear programs. First, however, in §3 we approach the problem using a so-called "canonical triplet" and two notions of strong *AC* optimality defined as follows.

DEFINITION 2.3. A policy δ^* is said to be

(a) *Strong optimal* (alias "asymptotically optimal" [8]) if

$$\limsup_{n\to\infty} n^{-1} J_n(\delta^*, \nu) \leq \liminf_{n\to\infty} n^{-1} J_n(\delta, \nu) \qquad \forall \delta \in \Delta, \nu \in \mathcal{P}(X).$$

(b) *F-strong optimal* (or strong optimal in the sense of Flynn [9]) if

$$\lim_{n\to\infty} n^{-1} [J_n(\delta^*, \nu) - J_n^*(\nu)] = 0 \qquad \forall \nu \in \mathcal{P}(X),$$

where $J_n^*(\cdot)$ is the optimal value function for the n-stage cost (2.3), i.e.,

$$J_n^*(\nu) := \inf_{\delta} J_n(\delta, \nu), \qquad \nu \in \mathcal{P}(X).$$

It is obvious that strong optimality (Definition 2.3(a)) implies optimality in the sense of (2.6), and so does F-strong optimality, since it implies that

$$\limsup_n n^{-1} J_n(\delta^*, \cdot) = \limsup_n n^{-1} J_n^*(\cdot).$$

Optimality, however, implies neither strong optimality nor F-strong optimality [8, Chap. 7], [9]. (In Theorem 3.3 we give conditions under which F-strong optimal $\Rightarrow$ strong optimal.)

Sufficient conditions for strong and/or F-strong optimality are given, e.g., in [9], [16], [3], [8], [10]. The conditions in the first two of these papers are based on the existence of a *bounded* solution to the optimality equation (OE) in (3.4) below. Our approach in the following section is somehow related to that in [9], [16], except that we allow a class of unbounded solutions to the OE. It should be noted, however, that there are other approaches. For instance, the optimal policy constructed in [6] is easily shown to be both strong and F-strong optimal, although the OE does *not* hold (in fact, the equality sign in (3.4) is replaced by strict inequality, $>$). The approach to F-strong optimality in [10], on the other hand, is quite different: it is based on an analysis of associated *discounted* cost problems (see also [9, Thm. 1]).

3. Canonical triplets. In this section we consider the so-called canonical triplets (or canonical systems [8], [27]) introduced by Yushkevich [34]. Our presentation below follows closely [3, §6], except that we allow unbounded one-stage costs satisfying the following condition (which will be weakened below: see Assumption 4.2).

(3.1) There is a number $m > 0$ and a positive measurable function b_1 on X such that, for all $(x, a) \in K$,

(i) $c(x, a) \leq m b_1(x)$, and

(ii) $\int b_1(y) Q(dy \mid x, a) \leq b_1(x)$.

If c is bounded, we may take $b_1(\cdot) \equiv 1$, and $m := \sup_{(x,a)} c(x, a)$.

Condition (3.1), as well as the following definition, are of common use in Markov control theory [17], [20], [24], [27].

DEFINITION 3.1. Let S be a Borel space, and $b : S \to R$ a positive measurable function. For any real-valued measurable function v on S, let

$$\| v \|_b := \sup_s | v(s) | b(s)^{-1},$$

and define $F(S, b)$ as the Banach space of all such functions v for which $\| v \|_b < \infty$. We call b a *bounding function* on S. ($\| \cdot \|_b$ is usually referred to as a *weighted supremum norm.*)

Among other important consequences, condition (3.1) guarantees the following lemma.

LEMMA 3.1. *Let $m, b_1(\cdot)$ and $F(X, b_1)$ be as in* (3.1) *and Definition* 3.1 (*with $S = X$*). *Then for any policy $\delta \in \Delta$ and any initial state $x \in X$.*

(a) $\{b_1(x_t), t \geq 0\}$ *is a P_x^δ-supermartingale, i.e.,*

$$E_x^\delta [b_1(x_{t+1}) \mid h_t] \leq b_1(x_t) \qquad \forall\ h_t \in H_t,\ t \geq 0,$$

which implies

(b) $E_x^\delta b_1(x_{t+1}) \le E_x^\delta b_1(x_t) \le \cdots \le b_1(x) \qquad \forall t \ge 0,$

and

(c) $J(\delta, x) \le m b_1(x).$

Moreover, for any function $h \in F(X, b_1)$,

(d) $\sup_\delta n^{-1} E_x^\delta |\, h(x_n)\,| \to 0$ *as* $n \to \infty$,

(e) $\limsup_n n^{-1} J_n(\delta, x, h) = \limsup_n n^{-1} J_n(\delta, x) \qquad (=: J(\delta, x)),$

and

$\liminf_n n^{-1} J_n(\delta, x, h) = \liminf_n n^{-1} J_n(\delta, x).$

Proof. By (3.1)(ii) and the "Markov property" (cf. [3], [14], [19]),

$$\begin{aligned} E_x^\delta [b_1(x_{t+1}) \,|\, h_t] &= \int_A \int_X b_1(y) Q(dy \,|\, x_t, a_t)\, \delta_t(da_t \,|\, h_t) \\ &\le \int_A b_1(x_t)\, \delta_t(da_t \,|\, h_t) \\ &= b_1(x_t). \end{aligned}$$

This proves (a), hence (b). Part (c) follows from (b), (3.1)(i), and (2.3)–(2.4). Also (d) follows from (b), since

$$E_x^\delta |\, h(x_n)\,| \le \|h\|_{b_1} E_x^\delta b_1(x_n) \le \|h\|_{b_1} b_1(x).$$

Finally, to obtain (e) note that, from (2.2)–(2.3),

$$J_n(\delta, x, h) = J_n(\delta, x) + E_x^\delta h(x_n). \qquad (3.2)$$

□

Canonical triplets. Let ρ and h be real-valued measurable functions on X, and let δ^* be a given policy. Then (ρ, h, δ^*) is said to be a canonical triplet if

$$J_n(\delta^*, x, h) = J_n^*(x, h) = n\rho(x) + h(x) \qquad \forall n \ge 0,\ x \in X, \qquad (3.3)$$

where

$$J_n^*(x, h) := \inf_\delta J_n(\delta, x, h)$$

is the optimal value function for the n-stage cost in (2.2).

The following theorem can be proved exactly as in the case of bounded costs [3, Thm. 6.2], [8, Chap. 7], [27], [34].

THEOREM 3.2. *Let* ρ *and* h *be real-valued measurable functions on* X, *and* $f^* \in F$ *a stationary policy* (*see Definitions* 2.1, 2.2). *Suppose that* (3.1) *holds and that* h *is in* $F(X, b_1)$. *Then* (ρ, h, f^*) *is a canonical triplet if and only if, for all* $x \in X$,

(a) $\rho(x) = \inf_{a \in A(x)} \int_X \rho(y) Q(dy \,|\, x, a)$,

(b) $\rho(x) + h(x) = \inf_{a \in A(x)} \left[c(x, a) + \int_X h(y) Q(dy \,|\, x, a) \right]$,

(c) $f^*(x) \in A(x)$ *attains the infimum in* (a) *and* (b), *i.e.*,

$$\rho(x) = \int \rho(y) Q(dy \,|\, x, f^*(x)),$$

$$\rho(x) + h(x) = c(x, f^*(x)) + \int h(y) Q(dy \,|\, x, f^*(x)).$$

The existence of a policy $f^* \in F$ satisfying (c) in Theorem 3.3 is typically ensured, under appropriate assumptions, by "measurable selection theorems" [3], [4], [8], [14], [19].

The equation in (b) is called the average cost *optimality equation* (OE). In §4 we relate two linear programs EP and EP^* to the OE with $\rho(\cdot) \equiv \rho^*$ a constant, i.e.,

$$\rho^* + h(x) = \inf_{a \in A(x)} \left[c(x,a) + \int h(y) Q(dy \mid x,a) \right]. \tag{3.4}$$

If h is such that (see (d) in Lemma 3.2)

$$\lim_{n \to \infty} n^{-1} E_x^{\delta} h(x_n) = 0 \qquad \forall\, \delta \in \Delta,\ x \in X, \tag{3.5}$$

then the policy $f^* \in F$ as in Theorem 3.3(c) is optimal with optimal value ρ^*, i.e.,

$$J^*(\cdot) = J(f^*, \cdot) = \rho^*; \tag{3.6}$$

moreover, (f^*, ν) is a minimum pair for any initial distribution ν, i.e.,

$$J(f^*, \nu) = \inf_{\Delta} AC \qquad \forall \nu \in \mathcal{P}(X).$$

We will next show, in the context of Theorem 3.3, that f^* is strong optimal and F-strong optimal (Definition 2.3).

THEOREM 3.3. *Suppose that* (3.1) *holds and let* (ρ, h, f^*) *be a canonical triplet with* $f^* \in F$ *and* $h \in F(X, b_1)$. *Then*

(a) $J(f^*, x) = \lim_n n^{-1} J_n(f^*, x) = \rho(x) \qquad \forall\, x \in X,$

(b) f^* *is* F*-strong optimal and strong optimal; in fact,* F*-strong optimality implies strong optimality (which implies optimality).*

Proof. (a) From (3.2)–(3.3), with $\delta = \delta^* = f^*$, we obtain

$$J_n(f^*, x, h) = J_n(f^*, x) + E_x^{f^*} h(x_n) = n\rho(x) + h(x). \tag{3.7}$$

Dividing by n and using Lemma 3.1(d), we obtain the second equality in (a). The first equality is obtained similarly, using now Lemma 3.1(e).

(b) First we show that f^* is F-strong optimal. From (3.3) with $\delta^* = f^*$, and (3.7),

$$J_n(f^*, x) = J_n^*(x, h) - E_x^{f^*} h(x_n).$$

On the other hand,

$$\begin{aligned} J_n^*(x,h) &:= \inf_{\delta} J_n(\delta, x, h) \le \inf_{\delta} J_n(\delta, x) + \sup_{\delta} E_x^{\delta} h(x_n) \\ &= J_n^*(x) + \sup_{\delta} E_x^{\delta} h(x_n). \end{aligned}$$

Thus

$$0 \le J_n(f^*, x) - J_n^*(x) \le \sup_{\delta} E_x^{\delta} h(x_n) - E_x^{f^*} h(x_n),$$

so that, by Lemma 3.1(d), f^* is F-strong optimal. Now, to show that the latter implies strong optimality, note that, by (3.3) and (3.7),

$$J_n(f^*, x) + E_x^{f^*} h(x_n) \le J_n(\delta, x) + E_x^{\delta} h(x_n) \qquad \forall\, \delta \in \Delta,\ x \in X,$$

and, therefore,

$$\liminf n^{-1} J_n(f^*, \cdot) \leq \liminf n^{-1} J_n(\delta, \cdot) \qquad \forall\, \delta.$$

This inequality, together with (a), yields that f^* is strong optimal. □

In brief, the main conclusion of Theorem 3.3 is that if the OE has a solution (ρ, h), with h satisfying (3.5), then a policy $f^* \in F$ determined by the OE (in the sense of Theorem 3.2(c)) is F-strong optimal. Moreover (from part (a) and the definition of F-strong optimal), the optimal average cost satisfies

$$J(f^*, x) = \lim_n n^{-1} J_n^*(x) = \rho(x),$$

i.e., the optimal long-run average cost is the limit of optimal average cost problems with a *finite* horizon. This was one of the main motivations for introducing the notion of F-strong optimality, in the first place [9].

4. The linear programming formulation. In this section we consider the linear programming (LP) formulation of the average cost (or AC) problem introduced in §2. We will use some basic facts on LP in general vector spaces for which our main source is [2, Chap. 3] (cf. also [20, Chap. 4]).

Dual pairs. Let X and Y be two vector spaces and let $\langle \cdot, \cdot \rangle$ be a bilinear form on $\mathsf{X} \times \mathsf{Y}$, i.e., a function from $\mathsf{X} \times \mathsf{Y}$ to R such that $\langle x, y \rangle$ is a linear function of x for each fixed $y \in \mathsf{Y}$, and a linear function in y for each fixed $x \in \mathsf{X}$.

DEFINITION 4.1. The pair of spaces (X, Y) is said to be a *dual pair* [2, p. 36] (alias a *separated duality* [20, p. 54]) if

(a) for each $x \neq 0$ in X there is some $y \in \mathsf{Y}$ with $\langle x, y \rangle \neq 0$, and

(b) for each $y \neq 0$ in Y there is some $x \in \mathsf{X}$ with $\langle x, y \rangle \neq 0$.

Let (X, Y) be a dual pair, and let $\sigma(\mathsf{X}, \mathsf{Y})$ denote the weak topology on X, i.e., the coarsest topology on X under which all the elements of Y are continuous when regarded as linear forms $\langle \cdot, y \rangle$ on X. If (X, Y) is a dual pair, then Y is the dual of X with the topology $\sigma(\mathsf{X}, \mathsf{Y})$.

To define the linear programs we are interested in, we now introduce two dual pairs (X, Y) and (Z, W) as follows.

Let K be the subset of $X \times A$ defined in (2.1), and let b be a bounding function on K (see Definition 3.1) given by

$$b(x, a) := c_0 + c(x, a) \quad \text{for some constant } c_0 > 0. \tag{4.1a}$$

(The role of c_0 is just to ensure that b is strictly positive.)

As our space Y we take $F(K, b)$, which we now write simply as $F(K)$, i.e., $F(K)$ is the Banach space of all measurable functions v from K to R such that

$$\|v\|_b := \sup_{(x,a)} |\, v(x, a) \,|\, b(x, a)^{-1} < \infty. \tag{4.1b}$$

We consider a function v on K to be extended to all of $X \times A$ in an arbitrary way as long as measurability and (4.1) are preserved. Now let $M(K)$ (this is our X) be the vector space consisting of all the finite signed measures μ on $X \times A$ concentrated on K such that

$$\int b\, d\,|\,\mu\,| < \infty; \quad \text{equivalently } \int c d|\mu| < \infty, \tag{4.2}$$

where $|\mu|$ denotes the total variation of μ. Note that $c \in F(K)$. Finally, if $v \in F(K)$ and $\mu \in M(K)$, we define

$$\langle \mu, v \rangle := \int v d\mu. \tag{4.3}$$

With this bilinear form, the pair $(M(K), F(K))$ is a dual pair.

Now let b_1 be a bounding function on X and let $F(X, b_1) =: F(X)$. Then we define $M(X)$ as the vector space that consists of all the finite signed measures ν on X such that

$$\int b_1 d|\nu| < \infty. \tag{4.4}$$

Finally, let (Z, W) be the dual pair defined by $\mathsf{Z} := R \times M(X)$, $\mathsf{W} := R \times F(X)$, and the bilinear form

$$\langle (r, \nu),\ (\rho, h) \rangle := r\rho + \int h d\nu. \tag{4.5}$$

Throughout the rest of this paper we suppose the following assumption to hold.

Assumption 4.2. The bounding functions b (on K) and b_1 (on X) are such that

(a) $b(x, a) \geq b_1(x) \quad \forall\ (x, a) \in K$;

(b) $\int b_1(y) Q(dy \mid \cdot)$ is in $F(K)$, i.e.,

$$\sup_{(x,a)} b(x, a)^{-1} \int b_1(y) Q(dy \mid x, a) < \infty.$$

(c) There is a policy $\delta \in \Delta$ such that, for every $\nu \in \mathcal{P}(X)$, $J(\delta, \nu) < \infty$ or, equivalently,

$$\limsup_n n^{-1} E_\nu^\delta \sum_{t=0}^{n-1} b(x_t, a_t) < \infty.$$

The set of all such policies is denoted by Δ_b.

Note that if $b(x, a) \geq b_1(x)\ \forall (x, a)$, then condition (3.1), if it holds with the bounding function b_1, implies (b) in Assumption 4.2. If, on the other hand, the one-stage cost is bounded, then Assumption 4.2 trivially holds with constants b and b_1, say $b = b_1 := \sup_{(x,a)} c(x, a)$, and $\Delta_b = \Delta$.

Linear programs. Let $L : M(K) \to R \times M(X)$ and $L^* : R \times F(X) \to F(K)$ be the linear maps defined as follows. For every $\mu \in M(K)$, let $L\mu := (\overline{\mu}, L_1\mu)$ be the element of $R \times M(X)$ given by

$$\overline{\mu} := \mu(K), \tag{4.6a}$$

$$L_1\mu(B) := \mu_1(B) - \int Q(B \mid x, a) \mu(d(x, a)), \qquad B \in \mathcal{B}(X), \tag{4.6b}$$

where μ_1 denotes the *marginal* (or projection) of μ on X, i.e.,

$$\mu_1(B) := \mu(B \times A) \qquad \forall\ B \in \mathcal{B}(X). \tag{4.7}$$

Now, if (ρ, h) is in $R \times F(X)$, we let $L^*(\rho, h)$ be the function on K defined as

$$L^*(\rho,h)(x,a) := \rho + h(x) - \int_X h(y)Q(dy \mid x,a).$$

Assumption 4.2 guarantees that both L and L^* are continuous maps and that L^* is the adjoint of L, i.e.,

$$\langle L\mu, (\rho,h)\rangle = \langle \mu, L^*(\rho,h)\rangle.$$

Finally, we consider two linear programs EP and EP^* (cf. [2, pp. 38–39]).

EP: minimize $\langle \mu, c\rangle$
subject to: $L\mu = (1,0)$, $\mu \geq 0$.
EP^*: maximize $\langle (1,0), (\rho,h)\rangle$
subject to: $L^*(\rho,h) \leq c$.

Equivalently, denoting by $M^1(K)$ the family of all *probability measures* μ in $M(K)$, so that $\overline{\mu} = 1$, we may rewrite EP as

EP: minimize $\int c d\mu$
subject to: $\mu \in M^1(K)$, and

$$\mu_1(B) - \int Q(B \mid x,a)\,\mu(d(x,a)) = 0, \qquad B \in \mathcal{B}(X). \tag{4.8}$$

Similarly, in a more explicit form, EP^* reads as follows.

EP^*: maximize ρ
subject to: $(\rho, h) \in R \times F(X)$, and

$$\rho + h(x) - \int h(y)Q(dy \mid x,a) \leq c(x,a) \qquad \forall\ (x,a) \in K. \tag{4.9}$$

In the terminology of [2, pp. 38–39], EP^* is the *dual* of the linear program EP.

A linear program is said to be *consistent* if it has a feasible solution, and *solvable* if it has an optimal solution. We will next give a condition under which EP is consistent (EP^* is always consistent!) and show how EP and EP^* relate to the AC problem in §2. The solvability question is postponed until §5.

If $\varphi \in \Phi$ is a relaxed policy (see Definitions 2.1 and 2.2), then we define, for all $x \in X$,

$$v(x,\varphi) := \int_A v(x,a)\varphi(da \mid x), \qquad v \in F(K), \tag{4.10}$$

$$Q(B \mid x,\varphi) := \int_A Q(B \mid x,a)\,\varphi(da \mid x), \qquad B \in \mathcal{B}(X). \tag{4.11}$$

Assumption 4.3. There exists a relaxed policy $\varphi \in \Phi$ such that $Q(\cdot \mid \cdot, \varphi)$ has an invariant probability measure $p^\varphi \in \mathcal{P}(X)$, i.e. (using the notation (4.10), (4.11)),

$$p^\varphi(B) = \int_X Q(B \mid x,\varphi)\,p^\varphi(dx) \qquad \forall\ B \in \mathcal{B}(X), \tag{4.12a}$$

and, furthermore,

$$\int b(x,\varphi)p^\varphi(dx) < \infty \quad \text{(equivalently, } \int c(x,\varphi)p^\varphi(dx) < \infty\text{)}. \tag{4.12b}$$

A relaxed policy that satisfies Assumption 4.3 is called *stable*. The invariant probability measure p^{φ} in (4.12a) is *not* required to be unique; if it is, then the transition kernel $Q(\cdot \mid \cdot, \varphi)$ is said to be *ergodic*. Sufficient conditions for (4.12a) are given in, e.g., [8], [14], [18], [25], [33]. (See also Lemma 6.1.) Notice, on the other hand, that Assumption 4.3 is in fact *equivalent* to EP's being consistent. (This fact is due to Lemma 4.4 and (4.15).)

Parts (b) and (c) in the following proposition are consequences, of course, of general LP results.

PROPOSITION 4.1. *Suppose that Assumptions* 4.2 *and* 4.3 *hold. Then*

(a) **Consistency:** *Both EP and EP^* are consistent; their values will be denoted by $\inf EP$ and $\sup EP^*$, respectively (if the programs are solvable, then we write $\inf$ as $\min$, and $\sup$ as $\max$)*;

(b) **Weak duality:** *For any feasible solutions μ of EP and (ρ, h) of EP^*, we have $\rho \leq \int c d\mu$; hence*

$$\sup EP^* \leq \inf EP; \tag{4.13}$$

(c) **Complementary slackness:** *If μ is feasible for EP, (ρ, h) is feasible for EP^*, and*

$$\langle \mu, c - L^*(\rho, h) \rangle = 0, \tag{4.14}$$

then μ is optimal for EP, (ρ, h) is optimal for EP^*, and equality holds in (4.13). (*The converse trivially holds.*)

Proof. (a) That EP^* is consistent is trivial: take, e.g., $\rho := \inf_{(x,a)} c(x,a)$ and $h(\cdot) \equiv 0$.

To see that EP is consistent, let φ be a stable relaxed policy, and let μ^{φ} be the probability measure on $X \times A$, concentrated on K, such that

$$\mu^{\varphi}(B \times C) := \int_B \varphi(C \mid x) p^{\varphi}(dx) \qquad \forall\, B \in \mathcal{B}(X),\ C \in \mathcal{B}(A). \tag{4.15}$$

Then $\mu_1^{\varphi}(B) := \mu^{\varphi}(B \times A) = p^{\varphi}(B) \quad \forall\ B \in \mathcal{B}(X)$, and so, from (4.12), μ^{φ} satisfies (4.8) and (4.2).

(b) To obtain (b), simply integrate (4.9) with respect to a measure $\mu \in M^1(K)$ that satisfies (4.8). This would yield

$$\langle \mu, L^*(\rho, h) \rangle \leq \langle \mu, c \rangle,$$

from which one can also deduce part (c). □

One can easily relate EP^* to the OE (3.4). Indeed, if (ρ^*, h) is a pair in $R \times F(X)$ that satisfies (3.4), then (ρ^*, h) is feasible for EP^* and, therefore, $\sup EP^* \geq \rho^*$. The reverse inequality is also obvious, since (4.9) implies that for any feasible solution (ρ, h) of EP^*

$$\rho + h(x) \leq \inf_{a \in A(x)} \left[c(x,a) + \int h(y) Q(dy \mid x, a) \right].$$

Thus, if such a pair (ρ^*, h) exists, then EP^* is solvable and

$$\max EP^* = \rho^*. \tag{4.16}$$

In other words, solving EP^* is equivalent to finding a solution to the OE when the latter exists.

To obtain further relations between the different problems, let us first prove the following (compare (4.17) with Lemma 3.1(d)).

LEMMA 4.2. (a) *If the Assumptions* 4.2(a), (c) *hold, and* $h \in F(X)$, *then*

$$\lim_n n^{-1} E_\nu^\delta [h(x_n)] = 0 \qquad \forall \delta \in \Delta_b, \ \nu \in \mathcal{P}(X). \tag{4.17}$$

(Δ_b *is as in Assumption* 4.2(c).)

(b) *If, in addition,* (ρ, h) *is a pair in* $R \times F(X)$ *that satisfies* (4.9), *then*

$$\rho \leq J(\delta, \nu) \qquad \forall \delta \in \Delta_b, \ \nu \in \mathcal{P}(X). \tag{4.18}$$

Proof. (a) Since, for *all* $\delta \in \Delta$ and $\nu \in \mathcal{P}(X)$,

$$E_\nu^\delta \, | \, h(x_n) \, | \leq \| \, h \, \|_{b_1} E_\nu^\delta b_1(x_n) \leq \| \, h \, \|_{b_1} E_\nu^\delta b(x_n, a_n),$$

part (a) follows from the Assumption 4.2(c), which implies that $\lim_n n^{-1} E_\nu^\delta b(x_n, a_n) = 0$ if $\delta \in \Delta_b$.

(b) This part is standard (see e.g. [3], [14], [29]). □

As a consequence of (4.18), we obtain, under Assumption 4.2(a) and (c),

$$\sup EP^* \leq \inf_\Delta AC, \tag{4.19}$$

where $\inf_\Delta AC$ is the value of the AC problem; see (2.7).

The corresponding result for the "primal" problem EP is as follows.

THEOREM 4.3. *Suppose that Assumption* 4.2 *holds and that* EP *is consistent. Then*

$$\inf_\Delta AC \leq \inf EP. \tag{4.20}$$

Proof. Suppose that $\mu \in M^1(K)$ satisfies (4.8). Then, by Lemma 4.4, there is a relaxed policy $\varphi \in \Phi$ satisfying (4.21). Thus, by the hypothesis on μ, the policy φ and $p^\varphi := \mu_1$ satisfy Assumption 4.3. Therefore, by the Individual Ergodic Theorem [33, p. 388], the average cost $J(\varphi, \mu_1)$ when using the policy φ and the initial distribution is μ_1 satisfies (cf. Remark 5.2)

$$\int c \, d\mu = J(\varphi, \mu_1) \geq \inf_\Delta AC.$$

Since μ was an arbitrary feasible solution of EP, (4.20) follows. □

LEMMA 4.4. *If* μ *is a probability measure on* $X \times A$ *concentrated on* K, *then there exists a relaxed policy* $\varphi \in \Phi$ *such that*

$$\mu(B \times C) = \int_B \varphi(C \, | \, x) \mu_1(dx) \qquad \forall B \in \mathcal{B}(X), \ C \in \mathcal{B}(A), \tag{4.21}$$

where μ_1 *is the marginal of* μ *on* X.

Proof. See, e.g., [8, p. 89, Thm. 2] or [19, p. 89, Cor. 12.7]. □

Equation (4.19) and Theorem 4.3 yield (4.22).

COROLLARY 4.5. *Suppose that Assumptions* 4.2 *and* 4.3 *hold. Then*

$$\sup EP^* \leq \inf_\Delta AC \leq \inf EP. \tag{4.22}$$

If, moreover, (ρ^*, h, f^*) *is a canonical triplet such that* ρ^* *is a constant and* $h \in F(X)$, *then (from* (3.6) *and* (4.16)) (ρ^*, h) *is optimal for* EP^*, f^* *is optimal for the AC problem and*

$$\rho^* = \max EP^* = \min_{\Delta} AC \leq \inf EP. \tag{4.23}$$

If a linear program and its dual have the same (finite) value, i.e.,

$$\sup EP^* = \inf EP, \tag{4.24}$$

it is then said that there is *no duality gap* for the problem [2, p. 52]. If there is no duality gap and the common value is achieved in each program (in which case we write sup EP^* as max EP^*, and inf EP as min EP), then the *strong duality* condition holds, i.e.,

$$\max EP^* = \min EP. \tag{4.25}$$

There are examples in which neither of these conditions hold (see, e.g., [6], [8], [23], [29]); in fact, both inequalities in (4.22) may be strict. We will give below conditions under which (4.24) and (4.25) hold. But first let us remark on a consequence of strong duality, when it holds.

Remark 4.6. Suppose that Assumptions 4.2 and 4.3 and the *strong duality condition* (4.25) hold, i.e., there exists an optimal solution $\mu^* \in M^1(K)$ for EP, an optimal solution $(\rho^*, h) \in R \times F(X)$ for EP^* and $\rho^* = \int c d\mu^*$. Then writing μ^* as in (4.21), i.e., $\mu^*(d(x,a)) = \varphi^*(da \,|\, x)\mu_1^*(dx)$, with $\varphi^* \in \Phi$, the complementary slackness equation (4.14) can be written as

$$\int_X \left[c(x, \varphi^*) - \rho^* - h(x) + \int h(y) Q(dy \,|\, x, \varphi^*) \right] \mu_1^*(dx) = 0. \tag{4.26}$$

Thus, by the Blackwell and Ryll–Nardzewski theorem, we conclude (as in [8, §3.2]) that there is a *nonrandomized* stationary policy $f^* \in F$ such that (using the abbreviation a.a.=almost all)

$$\begin{aligned} \rho^* + h(x) &= c(x, f^*(x)) + \int h(y) Q(dy \,|\, x, f^*(x)) \\ &= \min_{a \in A(x)} \left[c(x,a) + \int h(y) Q(dy|x,a) \right] \quad \text{for } \mu_1^*\text{–a.a. } x \in X. \end{aligned} \tag{4.27}$$

This is a "weaker" form of the OE (3.4), and one can derive by standard arguments [3], [8], [14], [29], etc., the corresponding "weak" form of (3.6):

$$J^*(x) = J(f^*, x) = \rho^* \qquad \text{for } \mu_1^*\text{–a.a. } x \in X. \tag{4.28}$$

That is to say, $f^* \in F$ is *optimal* μ_1^**-almost everywhere* (by which we mean of course (4.28)). Of course, integration with respect to μ_1^* in (4.28) shows that (f^*, μ_1^*) is a minimum pair. In conclusion, we have the following proposition.

PROPOSITION 4.7. *Suppose that Assumptions* 4.2 *and* 4.3 *are satisfied, and that there is strong duality. Then there exists a (nonrandomized) stationary policy* $f^* \in F$ *that is optimal* μ_1^**–a.e., where* μ_1^* *is the marginal on* X *of an optimal solution* μ^* *for the linear ("primal") problem* EP, *and* (f^*, μ_1^*) *is a minimum pair.*

In [23, Thm. 1], it is shown that if the state space is a *countable* set and control constraint sets $A(x)$ are *finite*, then the policy f^* obtained as in Proposition 4.7 is, in fact, *optimal*—as opposed to optimal μ_1^*-a.e., if there exists an ergodic stationary policy with finite average cost.

5. Solvability and absence of duality gap. We will now give sufficient conditions for the linear program EP to be solvable and for the absence of duality gap to hold. (These results are illustrated with examples in §6.) Throughout this section we suppose that both Assumptions 4.2 and 4.3 hold so that, in particular, EP is consistent.

Let $M^*(K)$ be the set of feasible solutions for EP, and let $M^+(K)$ be the positive cone in $M(K)$, i.e.,

$$M^*(K) := \{\mu \in M^1(K) \mid (4.8) \text{ holds}\}, \tag{5.1}$$
$$M^+(K) := \{\mu \in M(K) \mid \mu \geq 0\}.$$

Observe that $M^*(K) \subset M^1(K) \subset M^+(K) \subset M(K)$.

Solvability of EP requires Assumption 5.1 below, whereas the absence of duality gap requires the stronger Assumption 5.1′.

Notation. If S is a topological space, $C(S)$ denotes the space of real-valued, continuous and bounded functions on S.

Assumption 5.1. (a) The one-stage cost function c, hence the bounding function b, is lower semicontinuous (l.s.c.);

(b) The transition law Q is weakly continuous, i.e., $\int u(y)Q(dy \mid x, a)$ is a continuous and bounded function on K for every $u \in C(X)$;

(c) The set $M_r := \{\mu \in M^*(K) \mid \int c d\mu \leq r\}$ is tight for every number $r \geq 0$.

Assumption 5.1′ This assumption is the same as Assumption 5.1 except that (c) is replaced by

(c′) The set $M_r' := \{\mu \in M^1(K) \mid \int c d\mu \leq r\}$ is tight for every $r \geq 0$.

See Remark 5.6 for information concerning the tightness assumptions (c) and (c′).

Remark 5.1. A relaxed policy $\varphi \in \Phi$ that satisfies Assumption 4.3 is said to be *stable*. If $\varphi \in \Phi$ is stable and p^φ is as in Assumption 4.3, then the average cost $J(\varphi, p^\varphi)$ when using the policy φ with initial distribution p^φ satisfies

$$J(\varphi, p^\varphi) = \int J(\varphi, x) p^\varphi(dx) = \int c(x, \varphi) p^\varphi(dx) = \int c d\mu^\varphi, \tag{5.2}$$

where $\mu^\varphi \in M^*(K)$ is as in (4.15). Indeed, by (4.12b), the integral $\int c(x, \varphi) p^\varphi(dx)$ is finite and, therefore, (5.2) follows from the Individual Ergodic Theorem [33, p. 388].

THEOREM 5.2. *If Assumptions* 4.2, 4.3, *and* 5.1 *hold, then EP is solvable, i.e., there is a feasible solution μ^* for EP such that*

$$\min EP = \int c d\mu^*. \tag{5.3}$$

Moreover, decomposing μ^* as in Proposition 4.7 *we obtain a minimum pair (φ^*, μ_1^*) for the average cost (or AC) problem and*

$$\min EP = \inf_{\Delta} AC = J(\varphi^*, \mu_1^*). \tag{5.4}$$

Proof. 1°. By a standard argument using Assumptions 5.1(a) and (c) (see, e.g., [15], [21]), for any policy $\delta \in \Delta$ and any initial distribution ν on X for which the average cost $J(\delta, \nu)$ is finite, there exists a stable relaxed policy φ such that

$$J(\delta, \nu) \geq J(\varphi, p^\varphi) = \int c d\mu^\varphi.$$

(See (5.2).) In other words, if δ and ν are such that $J(\delta,\nu) < \infty$, then there is a feasible solution μ for EP such that

$$J(\delta,\nu) \geq \int c d\mu \; (\geq \inf EP). \tag{5.5}$$

This inequality and (4.20) imply (5.3) *if* EP is solvable.

2^o. Now, to show that EP is solvable, let us write $\rho^* := \inf EP$ and let $\{\varepsilon_n\}$ be a sequence of numbers such that $\varepsilon_n \downarrow 0$. For each n, let $\mu^n \in M^*(K)$ be such that

$$\rho^* \leq \int c d\mu^n < \rho^* + \varepsilon_n. \tag{5.6}$$

By the tightness assumption, Assumption 5.1(c), and Prohorov's theorem [5, p. 37], there is a subsequence $\{\mu^{n_i}\}$ of $\{\mu^n\}$ and a probability measure μ^* such that μ^{n_i} converges weakly (in the sense of probability measures [5]) to μ^*, i.e.,

$$\lim_i \int v d\mu^{n_i} = \int v d\mu^* \qquad \forall \; v \in C(K). \tag{5.7}$$

This weak convergence and the lower semicontinuity assumption, Assumption 5.1(a), yield

$$\int c d\mu^* \leq \liminf_i \int c d\mu^{n_i}, \tag{5.8}$$

so that, from (5.6),

$$\int c d\mu^* = \rho^* := \inf EP. \tag{5.9}$$

Thus, to conclude that μ^* is an optimal solution for EP, it only remains to show that μ^* is indeed feasible for EP, i.e., that it satisfies (4.2) and (4.8).

3^o. The fact that μ^* satisfies (4.2) follows from (5.6) and (5.9). Finally, that μ^* satisfies (4.8) follows from (5.7) and the weak continuity assumption, Assumption 5.1(b); namely, (5.7) implies the weak convergence of the marginals $\mu_1^{n_i} \to \mu_1^*$ on X and, therefore, for any function $u \in C(X)$

$$\begin{aligned} \int_X u(y)\mu_1^*(dy) &= \lim_i \int_X u(y)\mu_1^{n_i}(dy) \\ &= \lim_i \int_K \int_X u(y)Q(dy \,|\, k)\mu^{n_i}(dk) \qquad \text{[by (4.8)]} \\ &= \int_K \int_X u(y)Q(dy \,|\, k)\mu^*(dk), \end{aligned} \tag{5.10}$$

where the latter equality is due to (5.7) and Assumption 5.1(b). This yields (4.8). We have thus shown that $\mu^* \in M^*(K)$, which combined with (5.9) completes the proof of (5.3). Finally, to obtain (5.4) note that [cf. (5.2)]

$$\int c d\mu^* = \int c(x,\varphi^*)\mu_1^*(dx) = J(\varphi^*,\mu_1^*). \qquad \square$$

Sufficient conditions for the absence of duality gap (4.24) in general linear programs are given in [2, Chap. 3]. Here we choose one based on the closedness of the subset H of $(R \times M(X)) \times R$ defined as

$$H := \left\{ (L\mu, \langle \mu, c \rangle + r) \,|\, \mu \in M^+(K), \; r \geq 0 \right\}, \tag{5.11}$$

where $L\mu := (\overline{\mu}, L_1\mu)$ is the pair defined in (4.6); $\langle \mu, c\rangle = \int c d\mu$, as in (4.3), $M^+(K)$ is the positive cone in $M(K)$ and r a positive scalar; see (5.1). Explicitly, Theorem 3.9 in [2, p. 52] yields the following.

LEMMA 5.3. *If EP has a finite value and the set H in* (5.11) *is closed, then there is no duality gap for EP.*

THEOREM 5.4. *If Assumptions* 4.2, 4.3, *and* 5.1′ *hold, then there is no duality gap for EP, which combined with* (5.4) *yields*

$$\sup EP^* = \min EP = \inf_{\Delta} AC. \tag{5.12}$$

Proof. Assumption 5.1′ implies Assumption 5.1, since M_r' contains M_r. Hence, by Theorem 5.2 and Lemma 5.3, we only need to prove that H is closed, i.e., if $\{(\mu^n, r^n)\}$ is a sequence in $M^+(K) \times R^+$ such that

$$((\overline{\mu}^n, L_1\mu^n), \langle \mu^n, c\rangle + r^n) \to ((r_*, \nu_*), \rho_*) \in (R \times M(X)) \times R, \tag{5.13}$$

then $((r_*, \nu_*), \rho_*)$ is in H. The latter means of course that, for some $\mu \in M^+(K)$ and $r \in R^+$,

$$r_* = \lim \overline{\mu}^n = \overline{\mu}, \tag{5.14a}$$

$$\nu_* = \lim L_1\mu^n = L_1\mu, \tag{5.14b}$$

$$\rho_* = \lim \int c d\mu^n + r^n = \int c d\mu + r, \tag{5.14c}$$

where (see (4.5)) the convergence in (a)–(b) is in the sense that for all (ρ, h) in $R \times F(X)$,

$$r_*\rho + \int h d\nu_* = \lim \left[\overline{\mu}^n \rho + \int h d(L_1\mu^n)\right]. \tag{5.15}$$

To begin, note that if $r_* = \lim \overline{\mu}^n = 0$, then we are done, for (5.14) trivially holds taking $\mu(\cdot) \equiv 0$, the null measure on Kand $r = \rho_*$. On the other hand, if $r_* > 0$, then $\overline{\mu}^n := \mu^n(K)$ is positive for all n sufficiently large. Thus (dividing $\mu^n(\cdot)$ by $\overline{\mu}^n$ if necessary) we may, and will, assume, without loss of generality, that the μ^n in (5.13)–(5.15) are *probability* measures, i.e., $\overline{\mu}^n = r_* = 1$ for all n.

Now, taking $\rho = 0$, (5.15) and the definition of L_1 in (4.6) yield

$$\lim_n \left[\int_X h d\mu_1^n - \int_K \int_X h(y) Q(dy|k)\, \mu^n(dk)\right] = \int h d\nu_* \qquad \forall h \in F(x). \tag{5.16}$$

On the other hand, by (5.14c), for any given $\varepsilon > 0$, there exists $n(\varepsilon)$ such that

$$\int c d\mu^n < \rho_* + \varepsilon \qquad \forall\, n \geq n(\varepsilon). \tag{5.17}$$

Therefore, by Assumption 5.1′(c′), there is subsequence $\{\mu^{n_i}\}$ of $\{\mu^n\}$ and a probability measure μ such that (5.17) holds for all $n = n_i \geq n(\varepsilon)$ and μ^{n_i} converges weakly to μ, i.e.,

$$\int v d\mu^{n_i} \to \int v d\mu \qquad \forall\, v \in C(K). \tag{5.18}$$

Thus, the same argument that gave (5.8) now gives

$$\int c d\mu \leq \liminf_i \int c d\mu^{n_i},$$

which in turn gives that μ satisfies (4.2), i.e., μ is in $M^1(K)$. Moreover, as in the proof of (5.10), we obtain that, for all $u \in C(X)$,

$$\lim_i \int u d\mu_1^{n_i} = \int u d\mu_1, \tag{5.19a}$$

and

$$\lim_i \int \int u(y) Q(dy \mid k) \mu^{n_i}(dk) = \int \int u(y) Q(dy \mid k) \mu(dk). \tag{5.19b}$$

Therefore $l_1 \mu = \nu^*$. In other words, we have found a measure $\mu \in M^1(K)$ that satisfies (5.14a,b) and

$$\int c d\mu \leq \rho_*.$$

Finally, by taking $r = \rho_* - \int c d\mu \geq 0$ there is a measure $\mu \in M^1(K) \subset M^+(K)$ and a positive scalar r that satisfies (5.14), i.e., H is closed. This completes the proof of the theorem. □

A trivial case in which (all) the assumptions of Lemma 5.3 and Theorem 5.4 are satisfied occurs when both X and the control set A are finite. In the nonfinite case, we know of only three papers that have obtained results similar to ours: in [23] X is a *denumerable* set and the one-stage cost c is unbounded; in [32] X is a compact subset of R^n, and in [16] X is a Borel space, but in these two papers c is *bounded*. In §6 we shall consider a general class of systems for which Lemma 5.3 and Theorem 5.4 are valid.

Remark 5.5. General sufficient conditions for tightness of, say, a subset M of $M^1(K)$ are well known [5]; in particular, for Markov control/decision processes see, e.g., [18] and references therein. On the other hand, a sufficient condition that is particularly useful when dealing with unbounded one-stage costs is the following (cf. [13], [15], [25]): $\mathsf{M} \subset M^1(K)$ is tight if

(5.20) There is a moment function v such that $\sup\{\int v d\mu \mid \mu \in \mathsf{M}\} < \infty$. (By definition, a nonnegative measurable function v on Kis a *moment* if there exists a sequence of compact sets $K_n \uparrow K$ such that $\inf\{v(x,a) \mid (x,a) \notin K_n\} \to \infty$.)

The interpretation of a moment as a *Lyapunov* function is well known; see, e.g., [13], [25] and references therein. In [15] and [23], (5.20) has been used with $v = c$, that is, the moment function v is taken as the one-stage cost itself. A similar choice is done for the system in §6.

6. An application: Additive-noise systems. To illustrate the results of §§4 and 5, we introduce below a class of nonlinear additive-noise (i.e., autoregressive-like) systems, which are common in fields such as engineering and economics. Without going into too many details the idea is to give *sufficient* conditions for Assumptions 4.2, 4.3, and 5.1/5.1′ to hold, so that, in particular, *the conclusions of Propositions* 4.1, 4.7, *and Theorems* 5.2, 5.4 *are all valid.*

The state space and the control set are $X = R^p$ and $A = R^q$, respectively, and the control system evolves according to the equation

$$x_{t+1} = G(x_t, a_t) + \xi_t, \quad t = 0, 1, \ldots; \quad x_0 \in X \text{ given}, \tag{6.1}$$

where $x_t \in X$, $a_t \in A(x_t) \subset A$, and $\{\xi_t\}$ is a sequence of independent and identically distributed random p-vectors. The sets $A(x)$ are assumed to be closed such that K is convex. If the initial state x_0 is random, then it is assumed to be independent of $\{\xi_t\}$. The one-stage cost is the quadratic function on K (the set defined in (2.1)) given by

$$c(x, a) := x'Hx + a'Ra, \quad (\text{"prime" denotes "transpose"}), \tag{6.2}$$

where H and R are symmetric and positive definite matrices. If, in particular, G is linear in both x and a, and $A(\cdot) \equiv A$, then (6.1)–(6.2) reduces to the familiar LQ (or linear-quadratic) system.

Let $c_0 \geq c_1 > 0$ be two given constants, and define the "bounding" functions

$$b(x, a) := c_0 + c(x, a);$$
$$b_1(x) := c_1 + x'Hx \quad \left(\leq c_1 + \inf_{a \in A(x)} c(x, a) \right). \tag{6.3}$$

Let us suppose the following (basically continuity and "growth") conditions.

Assumption 6.1. (a) $G(x, a) : K \to X$ is continuous;

(b) $G(x, \varphi) := \int G(x, a)\varphi(da|x)$ is locally bounded for every $\varphi \in \Phi$;

(c) For some constant $m > 0$, $G(x, a)'HG(x, a) \leq mc(x, a)$ $\forall (x, a) \in K$;

(d) The random vector ξ_t are absolutely continuous with a density γ such that:

(d_1) γ is positive λ-almost everywhere ($\lambda :=$ Lebesgue measure) and, moreover,

(d_2) the ξ_t have zero mean and $E\,|\,\xi_0\,|^2 < \infty$;

(e) There exists a relaxed policy $\varphi^* \in \Phi$ for which the following holds: there are positive constants $\rho < 1$, k_1, k_2 such that

(e_1) $E|G(x, \varphi^*) + \xi|^2 \leq \rho|x|^2 \quad \forall |x| \geq k_1$, and

(e_2) $\int (a'Ra)\varphi^*(da|x) \leq k_2|x|^2$ $\forall x$.

The key fact to be noted about Assumption 6.1 is that it allows us to use results on ergodicity of time series [7], [26], [31]. For instance, [7, Prop. 4] or [26, Prop. 3] yields the following.

LEMMA 6.1. *Suppose that Assumptions* 6.1(b) *and* 6.1(d_1) *hold. If, moreover,* $\varphi^* \in \Phi$ *satisfies Assumption* 6.1(e_1), *then, when using the policy* φ^*, *the state (Markov) process* $\{x_t\}$ *is geometrically ergodic and its (unique) stationary distribution* p^{φ^*} *is such that* $\lambda << p^{\varphi^*}$ *(in words: the Lebesque measure* λ *is absolutely continuous with respect to* p^{φ^*}), *and it has a finite second moment, i.e.,* $\int |x|^2 p^{\varphi^*}(dx) < \infty$.

The following theorem is the main result in this section.

THEOREM 6.2. *Consider the system* (6.1) *with one-stage cost* (6.2). *Then Assumption* 6.1 *implies the Assumptions* 4.2, 4.3, *and* 5.1′ (*hence* 5.1).

Proof. We first verify Assumption 4.2: Assumption 4.2(a) is trivially satisfied by the definition of b and b_1 in (6.3). With respect to 4.2(b), we have

$$\begin{aligned} \int b_1(y)Q(dy|x, a) &= E\left[b_1(x_{t+1}) \,|\, x_t = x,\, a_t = a\right] \\ &= c_1 + E\left[(G(x, a) + \xi)'\, H(G(x, a) + \xi)\right] \\ &= c_1 + G(x, a)'HG(x, a) + E(\xi'H\xi), \end{aligned}$$

where ξ stands for a generic random vector with density γ (see Assumption 6.1(d)). Thus, from Assumptions 6.1(c), (d_2), and the definition of b, we obtain

$$\int b_1(y)Q(dy|x,a) \leq \text{constant} \cdot b(x,a) \qquad \forall\,(x,a) \in K,$$

i.e., Assumption 4.2(b) holds.

To verify Assumption 4.2(c), as well as Assumption 4.3, let us first note that Assumptions 6.1(b) and $6.1(d_1)$ imply that the state (Markov) process $\{x_t\}$ when using *any* relaxed policy φ is λ-Harris recurrent [7], [26], [31]. Thus if φ^* is the "ergodic" stable policy in Assumption 6.1(e) (see Lemma 6.1), then the Strong Law of Large Numbers for functionals of Markov chains [30] yields that, for any initial distribution ν,

$$\begin{aligned}\limsup_n n^{-1}E_\nu^{\varphi^*}\sum_{t=0}^{n-1} b(x_t,a_t) &= c_0 + J(\varphi^*,\nu)\\ &= c_0 + \int c(x,\varphi^*)p^{\varphi^*}(dx) < \infty,\end{aligned}$$

where the latter inequality comes from (6.2), Assumption 6.1 (e_2), and Lemma 6.1; i.e., from some constant H_1,

$$\begin{aligned}\int c(x,\varphi^*)p^{\varphi^*}(dx) &= \int (x'Hx)\,p^{\varphi^*}(dx) + \int\int (a'Ra)\,\varphi^*(da|x)p^{\varphi^*}(dx)\\ &\leq H_1\int |x|^2 p^{\varphi^*}(dx) + k_2\int |x|^2 p^{\varphi^*}(dx) < \infty.\end{aligned}$$

It only remains to verify Assumption 5.1′. The lower semicontinuity (in fact, continuity) of c is obvious: see (6.2). Assumption 5.1(b) follows from Assumption 6.1(a) and the Dominated Convergence Theorem: they yield that for any function $u \in C(X)$, the function

$$\int u(y)Q(dy|x,a) = E\,[u(x_{t+1})\,|\,x_t = x,\ a_t = a] = \int u\,[G(x,a)+s]\,\gamma(s)ds$$

is continuous and bounded in (x,a). Finally, the tightness condition in Assumption 5.1(c) (or (c′)) follows from (5.20), since clearly the function c in (6.2) is a moment:

$$c(x,a) \geq x'Hx \to \infty \text{ as } |x| \to \infty,\ \forall a \in A(x).$$

This completes the proof of the theorem. □

We have shown that Assumption 6.1 ensures the conclusions of Theorems 5.2 and 5.4, for example, but one could try for the problem (6.1)–(6.2), as well as for the general Markov control process in §2, alternative sufficient conditions (cf. [18], [25]). On the other hand, an important problem left open is how can one compute (or "approximate") the optimal value in, say, (5.3)–(5.4). For example, for *finite* state average cost problem the relation between policy iteration and the simplex algorithm of linear programming is well known. Does it hold an analogous relation in more general spaces?

REFERENCES

[1] E. ALTMAN AND A. SHWARTZ, *Markov decision problems and state-action frequencies*, SIAM J. Control Optim., 29 (1991), pp. 786–809.

[2] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces,* Wiley, Chichester, England, 1987.

[3] A. ARAPOSTATHIS, V. S. BORKAR, E. FERNÁNDEZ–GAUCHERAND, M. K. GHOSH, AND S. I. MARCUS, *Discrete–Time Controlled Markov Processes with Average Cost Criterion: A Survey,* Working Paper #91–032, Systems and Industrial Engineering Department, University of Arizona, Tuscon, AZ, 1991.

[4] J. P. AUBIN AND A. CELLINA, *Differential Inclusions,* Springer–Verlag, Berlin, 1984.

[5] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley, New York, 1968.

[6] R. CAVAZOS-CADENA, *A counterexample on the optimality equation in Markov decision chains with the average cost criterion*, Systems Control Lett., 16 (1991), pp. 387–392.

[7] J. DIEBOLT AND D. GUÉGAN, *Probabilistic Properties of the General Nonlinear Markovian Process of Order One and Applications to Time Series Modelling*, Rapport Technique #125, Laboratoire de Statistique Théorique et Appliquée, CNRS–URA 1321, Université Paris VI, 1990.

[8] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Processes*, Springer–Verlag, Berlin, 1979.

[9] J. FLYNN, *On optimality criteria for dynamic programs with long finite horizons*, J. Math. Anal. Appl., 76 (1980), pp. 202–208.

[10] M. K. GHOSH AND S. I. MARCUS, *On strong average optimality of Markov decision processes with unbounded costs*, Oper. Res. Lett., 11 (1992), pp. 99–104.

[11] W.-R. HEILMANN, *Generalized linear programming in Markovian decision problems*, Bonner Math. Schriften, 98 (1977), pp. 33–39.

[12] ———, *Solving stochastic dynamic programming problems by linear programming—An annotated bibliography*, Z. Oper. Res., 22 (1978), pp. 43–53.

[13] O. HERNÁNDEZ-LERMA, *Lyapunov criteria for stability of differential equations with Markov parameters*, Boletín Soc. Mat. Mexicana, 24 (1979), pp. 27–48.

[14] ———, *Adaptive Markov Control Processes*, Springer–Verlag, New York, 1989.

[15] ———, *Existence of average optimal policies in Markov control processes with strictly unbounded costs*, Kybernetika (Prague), 29 (1993), pp. 1–17.

[16] O. HERNÁNDEZ-LERMA, J. C. HENNET, AND J. B. LASSERRE, *Average cost Markov decision processes: Optimality conditions*, J. Math. Anal. Appl., 158 (1991), pp. 396–406.

[17] O. HERNÁNDEZ-LERMA AND J. B. LASSERRE, *Value iteration and rolling plans for Markov control processes with unbounded rewards*, J. Math. Anal. Appl., 177 (1993), pp. 38–55.

[18] O. HERNÁNDEZ-LERMA, R. MONTES DE OCA, AND R. CAVAZOS-CADENA, *Recurrence conditions for Markov decision processes with Borel state space: A survey*, Ann. Oper. Res., 28 (1991), pp. 29–46.

[19] K. HINDERER, *Foundations of Non-Stationary Dynamic Programming with Discrete-Time Parameter*, Lecture Notes Oper. Res. 33, Springer–Verlag, New York, 1970.

[20] W. K. KLEIN HANEVELD, *Duality in Stochastic Linear and Dynamic Programming*, Lecture Notes in Econom. Math. Systems, 274, Springer-Verlag, Berlin, 1986.

[21] M. KURANO, *The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin*, SIAM J. Control Optim., 27 (1989), pp. 296–307.

[22] H. J. KUSHNER AND A. J. KLEINMAN, *Mathematical programming and the control of Markov chains*, Internat. J. Control, 13 (1971), pp. 801–820.

[23] J. B. LASSERRE, *Average Optimal Stationary Policies and Linear Programming in Countable State Markov Decision Processes*, Rapport LAAS No. 91311, LAAS–CNRS, Toulouse, France 1991; J. Math. Anal. Appl., to appear.

[24] S. A. LIPPMAN, *On dynamic programming with unbounded rewards*, Management Sci., 21 (1975), pp. 1225–1233. To appear in J. Math. Anal. Appl.

[25] S. P. MEYN, *Ergodic theorems for discrete time stochastic systems using a stochastic Lyapunov function*, SIAM J. Control Optim., 27 (1989), pp. 1409–1439.

[26] A. MOKKADEM, *Sur un modèle autorégressif nonlinéaire. Ergodicité et ergodicité géométrique*, J. Time Series Anal., 8 (1987), pp. 195–204.

[27] A. V. PIUNOVSKI, *General Markov models with the infinite horizon*, Probl. Control Inform. Theory, 18 (1989), pp. 169–182.

[28] M. L. PUTERMAN, *Markov decision processes*, in Handbooks in Operations Research and Mathematical Sciences, Vol. 2, D. P. Heyman and M. J. Sobel, eds., North–Holland, Amsterdam, 1990, pp. 331–434.

[29] S. M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.
[30] D. Revuz, *Markov Chains*, 2nd ed., North–Holland, Amsterdam, 1984.
[31] D. Tjostheim, *Non–linear time series and Markov chains*, Adv. Appl. Probab., 22 (1990), pp. 587–611.
[32] K. Yamada, *Duality theorem in Markovian decision problems*, J. Math. Anal. Appl., 50 (1975), pp. 579–595.
[33] K. Yosida, *Functional Analysis*, 5th ed., Springer-Verlag, Berlin, 1978.
[34] A. A. Yushkevich, *On a class of strategies in general Markov decision models*, Theory Probab. Appl., 18 (1973), pp. 777–779.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 501–515, March 1994

010

A SIMPLE FREE BOUNDARY PROBLEM IN $\mathbf{R}^d$*

GUILLERMO FERREYRA† AND OMAR HIJAB‡

Abstract. A multidimensional deterministic singular control problem is posed and solved. The interest here is the explicitness of the result and the novelty of the gradient constraint.

Key words. singular control, free boundary problem, viscosity solution, Bellman equation

AMS subject classifications. 35F30, 49L20, 49L25

Introduction. We start with the linear system

$$\dot{x} = -a(t)x, \qquad x(0) = x \in \mathbf{R}^d, \tag{0.1}$$

where $a(t) \geq 0$ is a measurable function of time—the control—valued in the space of nonnegative $d \times d$ matrices, we define

$$v^a(x) = \int_0^\infty e^{-t}[\langle b, x(t)\rangle + \operatorname{trace}(a(t))]dt, \tag{0.2}$$

where $b \in \mathbf{R}^d \setminus 0$ is fixed throughout, and we set

$$v(x) = \inf\{v^a(x) \colon a(\cdot) \geq 0\}. \tag{0.3}$$

The problem we address in this paper is the analysis of the so-called *value function* v of the variational problem (0.1), (0.2), (0.3).

The interest in this problem is twofold: First, the question is a prototype of a class of multidimensional *singular control* problems involving linear dynamics and linear cost; as such, this class of problems should be applicable in a wide setting. In fact, we were led to the above model while searching for a multidimensional analog of a one-dimensional advertising model due to Vidale and Wolfe [3], [8], [11], [15]. A two-dimensional generalization, in a different direction, is given in [4].

Second, we show that v is the unique viscosity solution of the *free boundary problem*

$$\max(u - \langle b, x\rangle, \lambda(x, \nabla u) - 2) = 0, \qquad x \in \mathbf{R}^d, \tag{0.4}$$

in the class of solutions growing at most linearly; here $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product on $\mathbf{R}^d$ and

$$\lambda(x, p) = \langle x, p\rangle + |x||p|. \tag{0.5}$$

*Received by the editors June 10, 1992; accepted for publication (in revised form) December 14, 1992. This research was supported by the Louisiana Education Quality Support Fund.

† Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana 70803.

‡ Department of Mathematics, Temple University, Philadelphia, Pennsylvania 19122. This author's research was supported by National Science Foundation grant DMS-9121317.

We shall see that the free boundary—the hypersurface where both terms in (0.4) equal zero—is the paraboloid in $\mathbf{R}^d$ with axis through b, vertex at $b/|b|^2$, opening in the direction of $-b$, and defining equation

$$\lambda(x, b) = 2. \tag{0.6}$$

Using the method of characteristics, we also construct, off a portion L of the axis of the paraboloid (0.6), a classical $C^{1,1}$ solution u of (0.4). We then establish $u = v$.

We also address the question of the *optimal* choice of $a(\cdot)$, i.e., for each x how should we choose $a(\cdot)$ so that $v^a(x) = v(x)$? Since the answer depends on the initial state x, it is preferable to exhibit the solution in *feedback* form, i.e., to exhibit a single matrix-valued function $\mathbf{a}(x)$ such that for each starting x the optimal choice satisfies $a(t) = \mathbf{a}(x(t))$, $t \geq 0$, where $x(\cdot)$ satisfies (0.1).

Because of the singular nature of the above variational problem, one also expects the optimal control $a(\cdot)$ to be extreme, i.e., to equal zero or infinity. Below we show that, off L, the optimal feedback is given by

$$\mathbf{a}(x) \times \begin{cases} 0 & \text{if} \quad v(x) = \langle b, x\rangle, \\ \infty & \text{if} \quad \lambda(x, \nabla v(x)) = 2, \end{cases}$$

where $\mathbf{a}(x) = \xi(x) \otimes \xi(x)$ and

$$\xi(x) = \frac{|x|\nabla v(x) + |\nabla v(x)|x}{\sqrt{2|x||\nabla v(x)|}}. \tag{0.7}$$

More explicitly, we show that if $v(x) = \langle b, x\rangle$ the optimal feedback is zero, while if $\lambda(x, \nabla v(x)) = 2$ the optimal feedback is infinite so as to cause an instantaneous jump sending x to the intersection point x_0 of the free boundary with the integral curve through x of the vector field $-\mathbf{a}(x)x = -\langle x, \xi(x)\rangle\xi(x)$ corresponding to (0.7). In general terms, this is the kind of behavior expected for these types of problems [6], [7], [10], [12], [13], [14]. The interest here is, of course, the explicitness of our results.

In fact, it turns out that for each initial state x the optimal choice is

$$a(t) = a^*\delta(t), t \geq 0, \tag{0.8}$$

where $\delta(\cdot)$ is the Dirac impulse at time zero and the constant matrix $a^* = 0$ if $v(x) = \langle b, x\rangle$, while if $\lambda(x, \nabla v(x)) = 2$, a^* equals the (time-ordered) integral of $\mathbf{a}(x)$ along the integral curve of the vector field $-\mathbf{a}(x)x$ from x to x_0.

Finally we obtain another representation $v = w$, where

$$w(x) = \inf\{\langle b, qx\rangle - \log\det(q) : 0 < q \leq 1, \lambda(qx, b) \leq 2\}, \tag{0.9}$$

exhibiting the concave function v as a (restricted) conjugate of the strictly concave function $\log\det(q)$ on the space of positive $d \times d$ matrices q. In fact, w is obtained by restricting the infimum in (0.3) to controls of the form $a(\cdot) = a\delta(\cdot)$ with $a \geq 0$ constant and satisfying $\lambda(e^{-a}x, b) \leq 2$.

In particular, since L does not intersect the hypersurface $\lambda(x, b) = 2$ (§2), we obtain the C^1 regularity of v across the free boundary, and thus the C^1 "principle of smooth fit" holds in our situation.

If $\langle b, x\rangle$ is replaced by a C^∞ function $f(x)$ in (0.2), then we obtain a broader class of variational problems. We expect, under appropriate assumptions on f, the above to continue to hold with $\langle b, \cdot\rangle$ replaced by f throughout and with the free boundary now given by $\lambda(x, \nabla f(x)) = 2$; we lose, however, the explicitness of the results.

In the context of elliptic theory, (0.4) can be viewed as a rough approximation to

$$\max(v - \epsilon\Delta v - f(x), \lambda(x, \nabla v) - 2) = 0. \tag{0.10}$$

However, because our situation is degenerate ($\epsilon = 0$), we do not expect the C^2 "principle of smooth fit" [1], [12], [13] to hold. Indeed, for (0.10) with $\epsilon = 0$, a simple check reveals that v is not C^2 at any point where the free boundary is a C^1 hypersurface. In particular, for (0.4) v is never C^2 on the free boundary.

In §1 we show that v solves the Bellman equation (0.4) and prove uniqueness, in §2 we construct u, and in §3 we show $u = v = w$ and we verify the optimality of the choice (0.7), (0.8).

1. Derivation of the Bellman Equation. For background on viscosity solutions, see [2]. For background on viscosity solutions and singular control, see [9]. Strictly speaking, all we need from these references is the definition of a "viscosity solution" in terms of local extrema or equivalently strict local extrema. Nevertheless these references provide for a fuller understanding of the point of view taken here.

For $\epsilon > 0$ set

$$v_\epsilon(x) = \inf\{v^a(x) : I \geq \epsilon a(\cdot) \geq 0\}.$$

Then an easy approximation argument shows

$$v(x) = \inf_{\epsilon > 0} v_\epsilon(x). \tag{1.1}$$

Throughout this paper $|\cdot|$ denotes Euclidean length in $\mathbf{R}^d$. We assume $d \geq 2$, and $a \geq b$ for symmetric matrices a, b means $a - b \geq 0$, i.e., the eigenvalues of $a - b$ are nonnegative.

LEMMA 1. *$v_\epsilon \to v$ locally uniformly on $\mathbf{R}^d$ as $\epsilon \downarrow 0$ and v, v_ϵ are Lipschitz on $\mathbf{R}^d$ with Lipschitz constant $|b|$ for all $\epsilon > 0$. Moreover, $v(x)$ and $v_\epsilon(x)$ are bounded above by $\langle b, x\rangle$ and below by $-|b||x|$.*

Proof. Since $d/dt|x(t)|^2 = -2\langle x(t), a(t)x(t)\rangle \leq 0$, we have $|x(t)| \leq |x|$ for every initial state x. Since

$$\langle \nabla v^a(x), \xi\rangle = \int_0^\infty e^{-t}\langle b, \xi(t)\rangle dt,$$

where $\dot{\xi}(t) = -a(t)\xi(t)$, $\xi(0) = \xi$, it follows that $|\nabla v^a(x)| \leq |b|$. Since v and v_ϵ are infima of v^a, they are Lipschitz with constant $|b|$.

Let $a(\cdot)$ be a bounded control and suppose $x_\epsilon \to x$. Then

$$v^a(x) = \limsup_{\epsilon \downarrow 0} v^a(x_\epsilon) \geq \limsup_{\epsilon \downarrow 0} v_\epsilon(x_\epsilon)$$

and so $v(x) \geq \limsup_{\epsilon \downarrow 0} v_\epsilon(x_\epsilon)$ by (1.1). Moreover, we have

$$\liminf_{\epsilon \downarrow 0} v_\epsilon(x_\epsilon) \geq \liminf_{\epsilon \downarrow 0} v(x_\epsilon) = v(x);$$

the local uniform convergence follows. The last part follows from the fact that $v^0(x) = \langle b, x\rangle$ for all x, the Lipschitz nature of v_ϵ, v, and from $v_\epsilon(0) = v(0) = 0$. □

Given a symmetric matrix a, we wish to define $a^+ \geq 0$. This can be defined by the functional calculus or more specifically as follows. If a is diagonal, we let a^+ be obtained from a by replacing the negative diagonal elements by zero. For general symmetric a we extend this definition by insisting that $(rar^t)^+ = ra^+r^t$ for all rotations r. Similarly, we define $a^- \geq 0$. Then $a = a^+ - a^-$.

Given two vectors ξ, η in $\mathbf{R}^d$, let $\xi \otimes \eta$ denote the unique symmetric matrix satisfying $\langle a\xi, \eta\rangle = \text{trace}(a(\xi \otimes \eta))$ for all symmetric matrices a.

We now derive the Bellman equation satisfied by v_ϵ.

LEMMA 2. *For all $\epsilon > 0$, v_ϵ is a viscosity solution of*

$$\frac{1}{\epsilon}H(x, \nabla v_\epsilon) + v_\epsilon - \langle b, x\rangle = 0, \qquad x \in \mathbf{R}^d, \tag{1.2}$$

where

$$H(x,p) = \sup\{\langle p, ax\rangle - \textit{trace}(a) : 0 \leq a \leq I\} = \textit{trace}((x \otimes p - I)^+). \tag{1.3}$$

Proof. We begin by recalling the dynamic programming principle, which states that

$$v_\epsilon(x) = \inf\left\{\int_0^T e^{-t}[\langle b, x(t)\rangle + \text{trace}(a(t))]dt + e^{-T}v_\epsilon(x(T))\right\}, \tag{1.4}$$

where the infimum is over all controls $a(\cdot)$ satisfying $0 \leq \epsilon a(\cdot) \leq I$, and $T > 0$ is fixed.

Now suppose $\phi \in C^1(\mathbf{R}^d)$ and $x \in \mathbf{R}^d$ with $\phi(x) = v_\epsilon(x)$ and $v_\epsilon - \phi \leq 0$ near x. Then (1.4) yields, for all $a(\cdot) = a$ constant and satisfying $0 \leq \epsilon a \leq I$,

$$\phi(x) \leq \int_0^T e^{-t}[\langle b, x(t)\rangle + \text{trace}(a)]dt + e^{-T}\phi(x(T))$$

which implies

$$0 \leq \frac{1}{T}\int_0^T e^{-t}[\langle b, x(t)\rangle + \text{trace}(a) - \phi(x(t)) - \langle \nabla\phi(x(t)), ax(t)\rangle]dt;$$

letting $T \downarrow 0$ and taking the supremum over a, we obtain

$$\frac{1}{\epsilon}H(x, \nabla\phi(x)) + \phi(x) - \langle b, x\rangle \leq 0.$$

On the other hand, suppose x and $\phi \in C^1(\mathbf{R}^d)$ are such that $\phi(x) = v_\epsilon(x)$ and $v_\epsilon - \phi \geq 0$ near x. Since

$$\sup_{I \geq \epsilon a(\cdot) \geq 0}\left(\sup_{0 \leq t \leq T}|x(t) - x|\right) \to 0 \quad \text{as} \quad T \downarrow 0, \tag{1.5}$$

(1.4) implies, for $T > 0$ sufficiently small,

$$0 \geq \inf_{a(\cdot)}\left(\frac{1}{T}\int_0^T e^{-t}[\langle b, x(t)\rangle + \text{trace}(a(t)) - \phi(x(t)) - \langle \nabla\phi(x(t)), a(t)x(t)\rangle]dt\right).$$

Now (1.5) allows us to pass to the limit $T \downarrow 0$ and obtain

$$\frac{1}{\epsilon}H(x, \nabla\phi(x)) + \phi(x) - \langle b, x\rangle \geq 0.$$

To derive (1.3), let $c = x \otimes p - I$. Then $H(x,p)$ is the sup of $\operatorname{trace}(ac)$ over $0 \leq a \leq I$. But $\operatorname{trace}(ac) = \operatorname{trace}(c^+)$ if we choose a to be the orthogonal projection onto the positive eigenspaces of c and otherwise

$$\operatorname{trace}(ac) = \operatorname{trace}(ac^+) - \operatorname{trace}(ac^-) \leq \operatorname{trace}(ac^+) \leq \operatorname{trace}(c^+). \qquad \square$$

We note that (1.5) fails for $\epsilon = 0$.

For any $\xi \in \mathbf{R}^d \setminus 0$ we set $\xi' = \xi/|\xi|$. If ξ, η are two vectors in $\mathbf{R}^d$, we always take the angle between them to be $\theta = \cos^{-1}(\langle \xi', \eta'\rangle) \in [0, \pi]$.

The wedge product $\xi \wedge \eta$ is a vector in $\wedge^2\mathbf{R}^d$ and we have

$$|\xi \wedge \eta|^2 = \sum_{i<j}(\xi_i\eta_j - \xi_j\eta_i)^2.$$

An easy computation shows that $|\xi \wedge \eta|^2 + |\langle \xi, \eta\rangle|^2 = |\xi|^2|\eta|^2$ and hence $|\xi \wedge \eta| = |\xi||\eta| \sin\theta$, where θ is the angle between ξ and η.

We now turn to the analysis of the eigenvalues of $I - x \otimes p$.

LEMMA 3. *The eigenvalues of $I - x \otimes p$ are 1 with multiplicity $d-2$ and $\lambda_\pm$, where $\lambda_- \leq 1 \leq \lambda_+$ and*

$$\lambda_\pm = 1 + \frac{1}{2}(-\langle x, p\rangle \pm |x||p|). \tag{1.6}$$

If both x and p are nonzero, a unit eigenvector η for the eigenvalue λ_- is

$$\eta = \frac{|x|p + |p|x}{\sqrt{2|x||p|\lambda(x,p)}}, \tag{1.7}$$

where $\lambda(x,p)$ is given by (0.5). *Moreover,*

$$H(x,p) = \max(0, -\lambda_-) = \frac{1}{2}\max(0, \lambda - 2) = \frac{1}{\lambda_+}\max(0, -\det(I - x \otimes p)), \tag{1.8}$$

$$\det(I - x \otimes p) = 1 - \langle x, p\rangle - \frac{1}{4}|x \wedge p|^2, \tag{1.9}$$

and $\lambda(x,p) = 2$ iff $\lambda_-(x,p) = 0$ iff $\det(I - x \otimes p) = 0$ iff $I - x \otimes p$ has a nullspace.

Proof. Let $e_1, \ldots, e_d$ denote the standard basis in $\mathbf{R}^d$. Since the range of the linear transformation $x \otimes p : \mathbf{R}^d \to \mathbf{R}^d$ is at most two-dimensional, the linear transformation $\wedge^k(x \otimes p)$ induced on the basis of k-vectors $e_{i_1} \wedge \cdots \wedge e_{i_k} \in \wedge^k\mathbf{R}^d$ equals zero if $k \geq 3$. Thus the eigenpolynomial of $x \otimes p$ is given by

$$\lambda^d - \operatorname{trace}(x \otimes p)\lambda^{d-1} + \operatorname{trace}(\wedge^2(x \otimes p))\lambda^{d-2}.$$

Since traces can be computed by applying linear transformations to basis elements e_i and $e_i \wedge e_j$, we have

$$\operatorname{trace}(x \otimes p) = \langle x, p\rangle,$$

$$\operatorname{trace}(\wedge^2(x \otimes p)) = -\frac{1}{4}\sum_{i<j}(x_ip_j - x_jp_i)^2 = -\frac{1}{4}|x \wedge p|^2.$$

Thus the eigenpolynomial of $I - x \otimes p$ equals

$$(\lambda - 1)^{d-2}\left((\lambda - 1)^2 + \langle x, p\rangle(\lambda - 1) - \frac{1}{4}|x \wedge p|^2\right).$$

Solving the quadratic and using the fact that $\langle x, p\rangle = |x||p|\cos\theta$, $|x \wedge p| = |x||p|\sin\theta$, the results for $\lambda_\pm$ follow. Formula (1.7) follows by direct computation and (1.9) follows since $\det(I - x \otimes p) = \lambda_+\lambda_-$. $\square$

THEOREM 1. *v is a Lipschitz viscosity solution of* (0.4).

Proof. Lemma 1 states that v is Lipschitz. Let x and $\phi \in C^1$ be such that $v - \phi$ has a strict local maximum at x. Since $v_\epsilon \to v$ locally uniformly, this implies that there exists $x_\epsilon \to x$ such that $v_\epsilon - \phi$ has a local maximum at x_ϵ. This implies, by (1.2),

$$\frac{1}{\epsilon} H(x_\epsilon, \nabla\phi(x_\epsilon)) + v_\epsilon(x_\epsilon) - \langle b, x_\epsilon \rangle \leq 0.$$

Since $H \geq 0$, we obtain $v_\epsilon(x_\epsilon) - \langle b, x_\epsilon \rangle \leq 0$; letting $\epsilon \downarrow 0$ yields $v(x) - \langle b, x \rangle \leq 0$. Also, multiplying by ϵ and sending $\epsilon \downarrow 0$ yields $H(x, \nabla\phi(x)) \leq 0$. By Lemma 3, we obtain $\lambda(x, \nabla\phi(x)) - 2 \leq 0$. Thus v is a subsolution of (0.4).

Let x and $\phi \in C^1$ be such that $v - \phi$ has a strict local minimum at x. Choose $x_\epsilon \to x$ such that $v_\epsilon - \phi$ has a local minimum at x_ϵ. Then by (1.2)

$$\frac{1}{\epsilon} H(x_\epsilon, \nabla\phi(x_\epsilon)) + v_\epsilon(x_\epsilon) - \langle b, x_\epsilon \rangle \geq 0.$$

Now if $v(x) - \langle b, x \rangle \geq 0$, then v is a supersolution of (0.4). If not, then it follows that $H(x_\epsilon, \nabla\phi(x_\epsilon)) > 0$, which by Lemma 3 implies $\lambda(x_\epsilon, \nabla\phi(x_\epsilon)) - 2 > 0$, which yields in the limit $\lambda(x, \nabla\phi(x)) - 2 \geq 0$. Thus v is a supersolution of (0.4). □

The proof of Theorem 2 below is standard [2], [16]; the only new twist is the observation, needed below, that the constraint $\lambda(x, p)$ satisfies $\lambda(x, p + tx) \geq \lambda(x, p)$ for $t \geq 0$.

THEOREM 2. *Suppose $f(x)$ is Lipschitz. Then there is at most one continuous viscosity solution to*

$$\max(u - f(x), \lambda(x, \nabla u) - 2) = 0, \qquad x \in \mathbf{R}^d, \tag{1.10}$$

having at most linear growth.

Proof. The result is an immediate consequence of the corresponding comparison result: If u, v are sub- and supersolutions of (1.10), respectively, with at most linear growth, then $u \leq v$ on $\mathbf{R}^d$. To establish this, assume $|f(x) - f(y)| \leq C|x - y|$, $|f(x)| \leq C(1 + |x|)$.

It is enough to show that

$$(1 - \epsilon)u(x) - v(x) - \frac{\epsilon}{2}|x|^2 \leq \epsilon\left(C + \frac{1}{2}C^2\right), \qquad 0 < \epsilon < 1, x \in \mathbf{R}^d, \tag{1.11}$$

for then sending $\epsilon \downarrow 0$ we conclude. Now (1.11) follows from

$$\Phi(x, y) = (1 - \epsilon)u(x) - v(y) - \frac{\epsilon}{2}|x|^2 - \frac{\alpha}{2}|x - y|^2 \leq \epsilon\left(C + \frac{1}{2}C^2\right) + \frac{C^2}{2\alpha},$$
$$0 < \epsilon < 1, \alpha > 1, x, y \in \mathbf{R}^d, \tag{1.12}$$

for then taking $x = y$ and sending $\alpha \uparrow \infty$ we conclude.

To establish (1.12), let $(\hat{x}, \hat{y})$ be a point at which $\Phi(x, y)$ is maximized; such a point exists since u, v have at most linear growth. Then $\Phi(\cdot, \hat{y})/(1 - \epsilon)$ is maximized at $x = \hat{x}$; since u is a subsolution, this yields $u(\hat{x}) \leq f(\hat{x})$ and $\lambda(\hat{x}, \hat{p} + \epsilon\hat{x}) \leq 2(1 - \epsilon)$, where $\hat{p} = \alpha(\hat{x} - \hat{y})$. Here we have used the fact that $\lambda(x, p)$ is first-order homogeneous

in p. Similarly, since v is a supersolution, we have two cases: either $v(\hat{y}) \geq f(\hat{y})$ or $\lambda(\hat{y}, \hat{p}) \geq 2$. In the first case we obtain

$$\begin{aligned}\Phi(\hat{x}, \hat{y}) &\leq (1-\epsilon)f(\hat{x}) - f(\hat{y}) - \frac{\epsilon}{2}|\hat{x}|^2 - \frac{\alpha}{2}|\hat{x}-\hat{y}|^2 \\ &\leq C\epsilon(1+|\hat{x}|) + C|\hat{x}-\hat{y}| - \frac{\epsilon}{2}|\hat{x}|^2 - \frac{\alpha}{2}|\hat{x}-\hat{y}|^2,\end{aligned}$$

which is bounded by the right side of (1.12). Thus in the first case the result follows.

In the second case we have, using $\lambda(x, p+tx) \geq \lambda(x,p)$ for $t \geq 0$,

$$\lambda(\hat{x}, \hat{p}) \leq \lambda(\hat{x}, \hat{p} + \epsilon\hat{x}) \leq 2(1-\epsilon) < 2,$$

which yields $\lambda(\hat{x}, \hat{p}) - \lambda(\hat{y}, \hat{p}) < 0$. But the triangle inequality shows that the left side of this last inequality is nonnegative and we conclude. □

We remark that the development of this section can be easily modified to yield the existence and uniqueness of a solution to (0.10) with at most linear growth, when $f(x)$ is Lipschitz.

2. Solution of the free boundary problem. To motivate the solution, suppose first that u is a global C^1 solution of (0.4) and let $G = \{x : \lambda(x, \nabla u(x)) < 2\}$. Then it follows that $u(x) = \langle b, x\rangle$ on G and so $G \subset \{x : \lambda(x,b) < 2\}$ since G is open. Also if $x \in \partial G$, then $\lambda(x, \nabla u(x)) = 2$ and $u(x) = \langle b, x\rangle$. Since $u - \langle b, \cdot\rangle$ has a maximum at x, it follows that $\nabla u(x) = b$, which yields $\lambda(x,b) = 2$, which yields $\partial G \subset \{x : \lambda(x,b) = 2\}$. Thus we expect $G = \{x : \lambda(x,b) < 2\}$, $\partial G = \{x : \lambda(x,b) = 2\}$.

Throughout $C = \{x : \det(I - x \otimes b) \leq 0\} = \{x : \lambda(x,b) \geq 2\}$ is the complement of G and we refer to $\partial C = \{x : \det(I - x \otimes b) = 0\} = \{x : \lambda(x,b) = 2\}$ as the "free boundary." By (1.9), the free boundary is given by

$$0 = 1 - \langle x, b\rangle - \frac{1}{4}|x \wedge b|^2.$$

It follows that the free boundary is the paraboloid in $\mathbf{R}^d$ with axis along the line through the vector b, opening in the direction of $-b$, and vertex at $b/|b|^2$ (Fig. 1).

Let $L = \{tb/|b|^2 : t \geq e^2\} \subset \text{int}(C)$ be the infinite portion of the ray through b and starting at the point $e^2 b/|b|^2$.

Let A be an arbitrary subset of Euclidean space. We say a function f is C^∞ on A if for each x in A there is a C^∞ function g defined on an open neighborhood B of x such that f and g agree on $A \cap B$.

We point out that the *closed* free boundary lies in the *open* $\mathbf{R}^d \setminus L$. In this section we establish the following result.

THEOREM 3. *There is a Lipschitz function u on $\mathbf{R}^d$ such that*

(1) *u is $C^{1,1}$ on $\mathbf{R}^d \setminus L$,*
(2) *u is C^∞ on $C \setminus L$,*
(3) *$\lambda(x, \nabla u) = 2$ on $C \setminus L$,*
(4) *$u(x) < \langle b, x\rangle$ on int(C),*
(5) *$u(x) = \langle b, x\rangle$ on the complement of C,*
(6) *$\lambda(x, \nabla u) < 2$ on the complement of C;*

in particular u is a classical solution of (0.4) *on $\mathbf{R}^d \setminus L$ and $I - x \otimes \nabla u \geq 0$ on $\mathbf{R}^d \setminus L$.*

We begin by recalling the method of characteristics, following [5; §35.1].

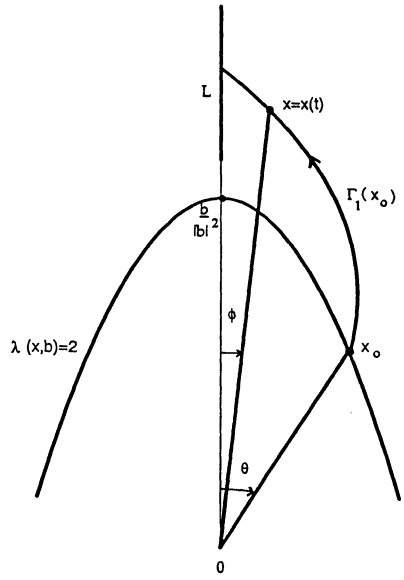


FIG. 1

Since $\lambda \in C^\infty((C \setminus L) \times (\mathbf{R}^d \setminus 0))$, the flow α_t of the Hamiltonian vector field

$$\begin{aligned} X_\lambda &= \langle \nabla_p \lambda, \nabla_x \rangle - \langle \nabla_x \lambda, \nabla_p \rangle \\ &= \langle x + p'|x|, \nabla_x \rangle - \langle p + x'|p|, \nabla_p \rangle \end{aligned}$$

is well defined (recall $\xi' = \xi/|\xi|$).

A submanifold $\Gamma \subset (C \setminus L) \times \mathbf{R}^d$ (of any dimension) is *Lagrangian* if the *symplectic form* $\langle dp \wedge dx \rangle = \sum_{i=1}^d dp_i \wedge dx_i$ vanishes on Γ. For example, let $f : C \setminus L \to \mathbf{R}^d$ be C^∞, and let $\Gamma = \{(x,p) : p = f(x)\}$ be the graph of f. Then Γ is Lagrangian if and only if $\langle dp \wedge dx \rangle = \langle df \wedge dx \rangle = d\langle f, dx \rangle = 0$; this happens iff the differential form $\langle f, dx \rangle = \sum_{i=1}^d f_i(x) dx_i$ is closed or, equivalently, the gradient $\nabla f(x)$ is a symmetric $d \times d$ matrix for all $x \in C \setminus L$. Since $C \setminus L$ is simply connected, this happens if and only if $\langle f, dx \rangle = du$ is exact or, equivalently, $f = \nabla u$ for some $u \in C^\infty(C \setminus L)$.

Given $x_0 \in \partial C$, let θ denote the angle between x_0 and b (Fig. 1), $\cos(\theta) = \langle x_0', b' \rangle$. Note $\theta < \pi$ always, since $\lambda(x_0, b) = 2$. For each $x_0 \in \partial C$ let $\Gamma(x_0)$ denote the Hamiltonian trajectory segment $\Gamma(x_0) = \{\alpha_t(x_0, b) : 0 \le t < \theta/\sin\theta\}$, where we define $\sin 0/0 = 1$. These are curves in phase (x,p)-space whose projections onto position x-space are drawn in Fig. 1. Although the Hamiltonian trajectory segments $\Gamma(x_0)$, being integral curves of the C^∞ vector field X_λ, cannot intersect, their projections $\Gamma_1(x_0)$ onto x-space—the characteristics—can and do in fact intersect. As we shall see below, the locus of points of intersections of the closures of $\Gamma_1(x_0)$ is precisely the "line of caustics" L.

Recall that the *Poisson bracket* of λ and $\beta = \beta(x,p)$

$$\{\lambda, \beta\} = X_\lambda(\beta) = \langle \nabla_p \lambda, \nabla_x \beta \rangle - \langle \nabla_x \lambda, \nabla_p \beta \rangle$$

vanishes if and only if the function β is a constant of the motion. In particular, λ is a constant of the motion and hence $\lambda(x,p) = 2$ on $\Gamma(x_0)$ for all $x_0 \in \partial C$.

Let

$$\Gamma = \bigcup_{x_0 \in \partial C} \Gamma(x_0).$$

Since $\{(x_0, b) : x_0 \in \partial C\} = \partial C \times \{b\}$ is a $(d-1)$-dimensional Lagrangian submanifold, it is a standard consequence that Γ is Lagrangian wherever it is a (d-dimensional) manifold. In particular, if we show that Γ is the graph of a C^∞ function $f : C \setminus L \to \mathbf{R}^d$, then Γ is a Lagrangian graph and hence $\lambda(x, \nabla u(x)) = \lambda(x, f(x)) = \lambda(x, p) = 2$ for $x \in C \setminus L$, since $\lambda(x, p) = 2$ on Γ.

To construct f we shall solve for the trajectories explicitly. Fix $x_0 \in \partial C$. Then the trajectory $(x(t), p(t)) = \alpha_t(x_0, b)$ starting from (x_0, b) satisfies

$$\begin{aligned} \dot{x}(t) &= x(t) + |x(t)|p(t)' \qquad x(0) = x_0, \\ \dot{p}(t) &= -|p(t)|x(t)' - p(t) \qquad p(0) = b. \end{aligned}$$

This implies ($I = I_d$ is the $d \times d$ identity matrix)

$$\frac{d}{dt}\begin{pmatrix} x(t)' \\ p(t)' \end{pmatrix} = \begin{pmatrix} -\cos\theta I & I \\ -I & \cos\theta I \end{pmatrix}\begin{pmatrix} x(t)' \\ p(t)' \end{pmatrix}, \tag{2.1}$$

$$\frac{d}{dt}|x(t)| = (1 + \cos\theta)|x(t)|, \qquad \frac{d}{dt}|p(t)| = -(1 + \cos\theta)|p(t)|. \tag{2.2}$$

Here $\theta = \theta(t)$ is the angle between $x(t)$ and $p(t)$.

Since the Poisson bracket of λ and $\beta(x, p) = \langle x, p\rangle$ vanishes, it follows that $\langle x, p\rangle$ and $|x||p|$ are constant along the trajectories $(x(t), p(t))$. In particular, θ does not depend on t and equals the angle between x_0 and b, which is always strictly less than π (Fig. 1).

If Q denotes the $2d \times 2d$ matrix appearing in (2.1), then $Q^2 = -\sin^2\theta I_{2d}$. Writing out the exponential series for e^{Qt} it follows that, when $\theta > 0$,

$$\begin{aligned} \sin\theta\, x(T)' &= \sin(\theta - T\sin\theta)\, x_0' + \sin(T\sin\theta)\, b', \\ \sin\theta\, p(T)' &= -\sin(T\sin\theta)\, x_0' + \sin(\theta + T\sin\theta)\, b'. \end{aligned} \tag{2.3}$$

Now take the inner product of the first of the pair (2.3) with b'. Let $\phi(t)$ denote the angle between $x(t)$ and b. We obtain $\cos\phi(T) = \cos(\theta - T\sin\theta)$ when $\theta > 0$ and hence

$$\phi(T) = \theta - T\sin\theta, \qquad 0 \le T < \theta/\sin\theta, \tag{2.4}$$

i.e., on Γ. It can be verified separately that (2.4) also holds when $\theta = 0$. Note that $T = \theta/\sin\theta$ is precisely when the characteristics intersect, at which point $\phi(T) = 0$, i.e., we are on the axis through b.

To see that we are actually in L, solve the first of the pair (2.2) to obtain

$$|x(T)||b| = |x_0||b|\exp(T(1 + \cos\theta)) \ge |x_0||b|, 0 \le T < \theta/\sin\theta. \tag{2.5}$$

Inserting $T = \theta/\sin\theta$ in (2.5) and noting that $\lambda(x_0, b) = |x_0||b|(1 + \cos\theta) = 2$, we obtain

$$|x(\theta/\sin\theta)||b| = \sec^2(\theta/2)\exp(\theta\cot(\theta/2)) \equiv g(\theta).$$

Since g is increasing and $g(0) = e^2$, we conclude that $x(\theta/\sin\theta) \in L$. Therefore, characteristics do not intersect the axis (which is itself the characteristic corresponding to $\theta = 0$) except in L.

The fact that we should have well-behaved (i.e., nonintersecting) characteristics *near* ∂C, even along the axis, is a reflection of the Cauchy–Kovalevska theorem; this is crucial because it is the existence of the segment of intersection of $C \setminus L$ with the axis that allows $C \setminus L$ to be simply connected.

Define $\Phi : C \setminus L \to \mathbf{R}^2$ by setting $\Phi(x) = (\mu, \phi/2)$, where $\mu = |x||b|$ and ϕ is the angle between x and b. Since $\lambda(x, b) \geq 2$ we have (Fig. 2)

$$\begin{aligned}\Phi(C \setminus L) &= \Phi(C) \setminus \Phi(L) \\ &= \{(\mu, \alpha) : \mu \geq 1, 0 \leq \alpha \leq \cos^{-1}(1/\sqrt{\mu})\} \setminus \{(\mu, 0) : \mu \geq e^2\} \subset \mathbf{R}^2.\end{aligned}$$

Moreover, $\Phi(\text{int}(C \setminus L)) = \text{int}(\Phi(C \setminus L))$ and Φ is C^∞ on $C \setminus L$.

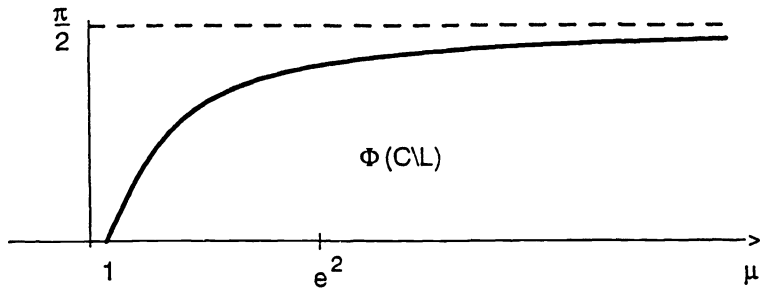


FIG. 2

We now show that θ, $T < \theta/\sin\theta$, and x_0 can be solved for uniquely in terms of $x = x(T) \in C \setminus L$, $\phi = \phi(T)$. Solving the first of the pair (2.3) for x_0' yields

$$x_0' = \frac{\sin\theta x' - \sin(\theta - \phi)b'}{\sin\phi}. \tag{2.6}$$

Integrating the second of the pair (2.2) yields

$$|p(T)| = |b| \exp(-T(1 + \cos\theta)) \leq |b|. \tag{2.7}$$

Since $\lambda(x_0, b) = |x_0||b|(1 + \cos\theta) = 2$ (2.5) implies

$$(1 + \cos\theta)|b||x| = 2\exp(T(1 + \cos\theta)). \tag{2.8}$$

Eliminating T in (2.4), (2.8) we arrive at

$$h\left(|x||b|, \frac{\theta}{2}\right) = \frac{\phi}{2}, \tag{2.9}$$

where $h(\mu,\cdot):[0,\cos^{-1}(1/\sqrt{\mu})]\to[0,\cos^{-1}(1/\sqrt{\mu})]$ is given by

$$h(\mu,\alpha)=\alpha-\frac{1}{2}\tan\alpha\log(\mu\cos^2\alpha) \tag{2.10}$$

for each $\mu=|x||b|\geq 1$.

There are two cases, depending on whether $1\leq\mu<e^2$ or $\mu\geq e^2$ (Fig. 3). In the first case, for each $\phi/2\in[0,\cos^{-1}(1/\sqrt{\mu})]$ there is a unique $\theta/2\in[0,\cos^{-1}(1/\sqrt{\mu})]$, $\mu=|x||b|$, such that (2.9) holds and so $h(\mu,\cdot)$ is a homeomorphism of $[0,\cos^{-1}(1/\sqrt{\mu})]$ onto itself. In the second case, for each $\phi/2\in(0,\cos^{-1}(1/\sqrt{\mu})]$ there is a unique $\theta/2\in(\alpha^*(\mu),\cos^{-1}(1/\sqrt{\mu})]$, $\mu=|x||b|$, such that (2.9) holds and so $h(\mu,\cdot)$ is a homeomorphism of $[\alpha^*(\mu),\cos^{-1}(1/\sqrt{\mu})]$ onto $[0,\cos^{-1}(1/\sqrt{\mu})]$; here $\alpha^*(\mu)$ is the largest zero of $h(\mu,\cdot)$ in $[0,\cos^{-1}(1/\sqrt{\mu})]$. We conclude that for each $x\in C\setminus L$ there is a unique $\theta/2\in[0,\cos^{-1}(1/\sqrt{\mu})]$, $\mu=|x||b|$, such that (2.9) holds. Defining T by (2.8), x_0 by (2.5), (2.6), we obtain $p=p(T)$ as a uniquely determined function f of $x=x(T)$ by the second of the pair (2.3) and (2.7), and hence Γ is a graph over $C\setminus L$.

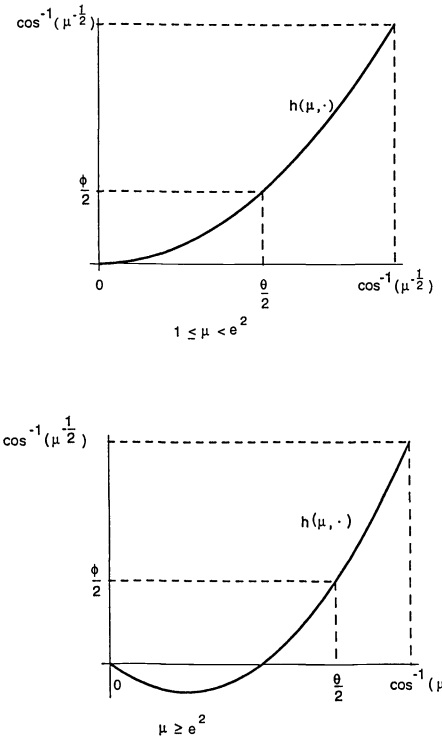


FIG. 3

Moreover, it is easy to see that $\partial h(\mu,\theta/2)/\partial\alpha\neq 0$ for all $(\mu,\phi/2)\in\Phi(C\setminus L)$ and hence, by the implicit function theorem, $\theta=\theta(\mu,\phi)$ given by (2.9) is a C^∞ function on $\Phi(C\setminus L)$. It follows that $\theta=\theta(x)$ is a C^∞ function of $x\in C\setminus L$. Thus Γ is a C^∞

graph over $C \setminus L$ and the differential form $\langle p, dx\rangle = \langle f(x), dx\rangle$ is closed and hence exact on $C \setminus L$. Setting

$$u(x) = \langle b, x_0\rangle + \int_{x_0}^{x} \langle p, dx\rangle, \tag{2.11}$$

yields a C^∞ function u satisfying $p = \nabla u(x)$, $\lambda(x, \nabla u) = 2$ on $C \setminus L$, whose value does not depend on the choice of x_0 in ∂C, since $p = b$ on ∂C.

Since $p = \nabla u(x)$, it follows from (2.7) that this solution u is Lipschitz with Lipschitz constant $|b|$ and hence extends uniquely as a Lipschitz function across L. Set $u(x) = \langle b, x\rangle$ if $x \notin C$. It then follows that u is Lipschitz on $\mathbf{R}^d$ and $u \in C^{1,1}(\mathbf{R}^d \setminus L)$, since $u(x) = \langle b, x\rangle$, $\nabla u = b$ on ∂C. Moreover, since $\nabla u = b$ on the complement of C, we have $\lambda(x, \nabla u) < 2$ on the complement of C.

To complete the proof of Theorem 3, it remains to establish (4), i.e., $u(x) < \langle b, x\rangle$ on $\operatorname{int}(C)$. We first show $u(x) \le \langle b, x\rangle$ on C.

By (2.11) and

$$\langle b, x\rangle = \langle b, x_0\rangle + \int_{x_0}^{x} \langle b, dx\rangle, \tag{2.12}$$

it is enough to show $2 = \langle p, \dot{x}\rangle \le \langle b, \dot{x}\rangle$ along $\Gamma_1(x_0)$, i.e., it is enough to show $g(t) \ge 2$ on $[0, \theta/\sin\theta]$, where

$$g(t) = \langle b, \dot{x}(t)\rangle = \langle b, x(t) + p(t)'|x(t)|\rangle.$$

To this end, note that $g(0) = 2$ and $g(\theta/\sin\theta) = |b||x|(1 + \cos\theta) = \frac{|x|}{|x_0|}|b||x_0|(1 + \cos\theta) = 2|x|/|x_0| \ge 2$ by (2.5). Thus if $g(t)$ were strictly less than 2 somewhere, it would have an interior minimum t. At this time t we would then have

$$\dot{g}(t) = 2\langle p(t)', b\rangle|x(t)|(1 + \cos\theta) = 0$$

and

$$\ddot{g}(t) = \dot{g}(t)(1 + 2\cos\theta) - 2\langle x(t), b\rangle(1 + \cos\theta) \ge 0.$$

Since θ is never equal to π, this would imply that the angle between $p(t)$ and b equals $\pi/2$ and the angle $\phi(t)$ between $x(t)$ and b is $\ge \pi/2$. But it follows from the second of the pair (2.3) that the angle between $p(t)$ and b equals $\theta - \phi(t)$. Since θ is strictly less than π, this cannot equal $\pi/2$. This contradiction establishes $g(t) \ge 2$ on the entire interval $[0, \theta/\sin\theta]$ and so $u(x) \le \langle b, x\rangle$ on C.

Since $g(t) \ge 2$ on the interval, repeating the same argument shows that $g(t) > 2$ on the interior of the interval. Thus $\langle b, x\rangle - u(x)$ is a strictly increasing function along $\Gamma_1(x_0)$ from x_0 to L; therefore $u(x) < \langle b, x\rangle$ on $\Gamma_1(x_0) \setminus x_0$ and hence on $\operatorname{int}(C)$.

Since $I - x \otimes \nabla u(x) \ge 0$ on $\mathbf{R}^d \setminus L$ follows from Lemma 3 and the above, this completes the proof of Theorem 3. □

3. Equality of u, v, and w.

THEOREM 4. *$u = v = w$ on $\mathbf{R}^d$.*

We begin by showing $u = v$. We first show that $v(x) \ge u(x)$ for all $x \in \mathbf{R}^d$. To this end it enough to establish $v^a(x) \ge u(x)$ for all bounded controls $a(\cdot)$. In fact, if we fix a bounded control $a(\cdot)$, it is enough to establish $v^a(x) \ge u(x)$ almost everywhere on $\mathbf{R}^d$, since both sides are Lipschitz.

Let $x^a(\cdot;x)$ denote the solution trajectory of (0.1) starting from x; since $x \mapsto x^a(t;x)$ is a diffeomorphism and L has measure zero, for each $t \geq 0$ we see that the set $\{x : x^a(t;x) \in L\}$ has measure zero. By Fubini, this implies that there is a null set N (depending on $a(\cdot)$) such that the set $\{t \geq 0 : x(t) = x^a(t;x) \in L\}$ has measure zero for all $x \notin N$.

Now let $x = x(0) \notin N$; then $t \mapsto e^{-t}u(x(t))$ equals the integral of its derivative and hence

$$e^{-T}u(x(T)) = u(x) + \int_0^T e^{-t}\left(-u(x(t)) - \langle \nabla u(x(t)), a(t)x(t)\rangle\right) dt$$

for all $T > 0$. Since u has linear growth, $|\nabla u| \leq |b|$, and $a(\cdot)$, $x(\cdot)$ are bounded letting $T \uparrow \infty$ yields

$$0 = u(x) + \int_0^\infty e^{-t}\left(-u(x(t)) - \langle \nabla u(x(t)), a(t)x(t)\rangle\right) dt. \tag{3.1}$$

Combining (0.2) and (3.1) yields

$$v^a(x) = u(x) + \int_0^\infty e^{-t}[\langle b, x\rangle - u(x) + \operatorname{trace}(a(I - x \otimes \nabla u))]dt. \tag{3.2}$$

Since $a(\cdot) \geq 0$, $I - x \otimes \nabla u \geq 0$, $\langle b, x\rangle - u(x) \geq 0$ on $\mathbf{R}^d \setminus L$, we obtain $v^a(x) \geq u(x)$. By the remarks above, this establishes $v \geq u$ on $\mathbf{R}^d$.

Since $v^0(x) = \langle b, x\rangle$, we have $u = v$ on the complement of $\operatorname{int}(C)$. Now define, for $x \in C \setminus L$,

$$\mathbf{a}(x) = \xi(x) \otimes \xi(x),$$

where $\xi(x)$ and $\eta(x)$ are given by

$$\xi(x) = \sqrt{2}\eta(x) = \frac{|x|\nabla u(x) + |\nabla u(x)|x}{\sqrt{2|x||\nabla u(x)|}}$$

and check that

$$\mathbf{a}(x)x = \langle x, \xi(x)\rangle\xi(x) = \nabla_p\lambda(x, \nabla u(x)) = \nabla_p\lambda(x,p); \tag{3.3}$$

here we have used $\lambda(x, \nabla u) = 2$ on $C \setminus L$.

Fix $x \in \operatorname{int}(C \setminus L)$ and let $x_1(t)$, $t \geq 0$, be the integral curve of the vector field $-\mathbf{a}(x)x$ (note the minus!) starting at x at time zero. Setting $p_1(t) = \nabla u(x_1(t))$, differentiating $\lambda(x, \nabla u(x)) = 2$, and using (3.3) shows

$$\begin{aligned} \dot{x}_1 &= -\nabla_p\lambda(x_1, p_1), \\ \dot{p}_1 &= +\nabla_x\lambda(x_1, p_1). \end{aligned}$$

Thus $(x_1(t), p_1(t))$ is the integral curve of $-X_\lambda$ through $(x, \nabla u(x))$ and, as a result, $(x_1(t), p_1(t)) = \alpha_{T-t}(x_0, b) = (x(T-t), p(T-t))$, where T, x_0, $(x(t), p(t))$ are as in §2.

We can now attempt to define a control $a_1(\cdot)$ satisfying $v^{a_1}(x) = u(x)$ for the fixed $x \in \operatorname{int}(C \setminus L)$: Set $a_1(t) = \mathbf{a}(x_1(t))$, $0 \leq t < T$, $a_1(t) = 0$, $t \geq T$. It follows

that the unique solution of (0.1) corresponding to $a_1(\cdot)$ equals $x_1(t)$, if $0 \leq t \leq T$, and equals x_0 if $t \geq T$.

Now note that $\lambda(x_1(t), \nabla u(x_1(t))) = 2$ and hence (Lemma 3) the matrix $I - x_1(t) \otimes \nabla u(x_1(t))$ has a nullspace for each $0 \leq t \leq T$. In fact, by (1.7), our control $a_1(t) = \xi(t) \otimes \xi(t)$ equals twice the orthogonal projection $\eta(t) \otimes \eta(t)$ onto the aforementioned nullspace and thus

$$\text{trace}(a_1(t)(I - x_1(t) \otimes \nabla u(x_1(t)))) = 0, 0 \leq t \leq T!$$

Armed with this information, we compute the corresponding cost $v^{a_1}(x)$ using (3.2) and obtain

$$v^{a_1}(x) = u(x) + \int_0^T e^{-t}(\langle b, x_1(t)\rangle - u(x_1(t)))dt. \tag{3.4}$$

Hence our choice $a_1(\cdot)$ does *not* satisfy $v^{a_1}(x) = u(x)$, since $u(x) < \langle b, x\rangle$ in the interior of C.

In fact, a little thought shows that no control $a(\cdot)$ can satisfy $v^a(x) = u(x)$ for any $x \in \text{int}(C \setminus L)$, since $x(\cdot)$ must spend some time in the interior of $C \setminus L$ and hence by (3.2) $v^a(x)$ is strictly greater than $u(x)$.

But now we know what to do: If we rescale time and run along the trajectory $x_1(t)$ at a very high speed, we spend less time in $C \setminus L$ and hence lower the simple running cost in (3.4). More precisely, let

$$a_\epsilon(t) = \frac{1}{\epsilon} a_1\left(\frac{t}{\epsilon}\right), t \geq 0. \tag{3.5}$$

Repeating the above steps, we obtain

$$v^{a_\epsilon}(x) = u(x) + \epsilon \int_0^T e^{-\epsilon t}(\langle b, x_1(t)\rangle - u(x_1(t)))dt.$$

Letting $\epsilon \downarrow 0$, we obtain $v(x) \leq \lim_{\epsilon \downarrow 0} v^{a_\epsilon}(x) = u(x)$. This shows $v = u$ on $\mathbf{R}^d$. Since we now know $u = v$, the claim made in §0 relating to (0.7) follows.

Now let $w^q(x)$ denote the quantity whose infimum is taken in (0.9). To establish $v = w$, fix $x \in \mathbf{R}^d$ and let $0 < q \leq 1$ be such that $\lambda(qx, b) \leq 2$. Let $a = -\log q \geq 0$, and set $a_\epsilon(t) = a/\epsilon$, if $0 \leq t < \epsilon$, and equal zero otherwise. Since $-\log \det(q) = \text{trace}(a)$, computing $v^{a_\epsilon}(x)$ according to (0.2) and passing to the limit yields

$$w^q(x) = \langle b, qx\rangle - \log \det(q) = \lim_{\epsilon \downarrow 0} v^{a_\epsilon}(x) \geq v(x) \tag{3.6}$$

and thus $w \geq v$ on $\mathbf{R}^d$.

On the other hand, if $x \notin \text{int}(C)$, $q = 1$ shows $w(x) = u(x) = v(x)$, while if $x \in C \setminus L$ choosing $a_\epsilon(\cdot)$ according to (3.5) and evaluating $v^{a_\epsilon}(x)$ according to (0.2) yields

$$w^{q^*}(x) = \langle b, q^* x\rangle - \log \det(q^*) = \lim_{\epsilon \downarrow 0} v^{a_\epsilon}(x) = v(x), \tag{3.7}$$

where q^* equals the fundamental solution of $\dot{x} = -a_1(t)x$ evaluated at time T. Thus $w = v$ on $\mathbf{R}^d \setminus L$.

It can be shown, using the strict concavity of $q \mapsto \log\det(q)$ on the space of positive matrices q and the convexity of the set of matrices q satisfying $0 < q \leq 1$, $\lambda(qx, b) \leq 2$, that the infimum in (0.9) is achieved at a unique point for each x. From this it follows that w is continuous on $\mathbf{R}^d$ and hence $w = v$ on $\mathbf{R}^d$. This completes the proof of Theorem 4.

The claim regarding (0.8) is clear if the initial state $x \notin C$, so fix an $x \in C$ and set $a^* = -\log q^*$, by definition the "time-ordered" integral of $\mathbf{a}(x_1(t))$ over $[0, T]$. Then it follows from (3.7) that we obtain equality in (3.6), which establishes the claim made in §0 regarding (0.8) for this x and hence all $x \in \mathbf{R}^d$.

REFERENCES

[1] V. E. Benes, L. A. Shepp, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.

[2] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bulletin Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[3] J. R. Dorroh and Guillermo Ferreyra, *Optimal Advertising in Exponentially Decaying Markets*, Lecture Notes in Functional Analysis and PDEs, LSU, 1991.

[4] ———, *A Multi-State, Multi-Control Problem With Unbounded Controls*, preprint.

[5] B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern Geometry I*, Graduate Texts in Mathematics #93, Springer-Verlag, New York, 1984.

[6] N. El-Karoui and I. Karatzas, *Integration of the Optimal Stopping Time Problem and a New Approach to the Skorokhod Problem*, preprint.

[7] L. C. Evans, *A second order elliptic equation with gradient constraint*, Comm. Partial Differential Equations, 4 (1979), pp. 555–572.

[8] G. Ferreyra, *The Optimal Control Problem for the Vidale–Wolfe Advertising Model Revisited*, preprint.

[9] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.

[10] J. L. Menaldi & M. Robin, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.

[11] S. P. Sethi, *Optimal control of the Vidale-Wolfe advertising model*, Oper. Res., 21 (1973), pp. 998–1013.

[12] S. E. Shreve and H. M. Soner, *Regularity of the value function for a 2 dimensional singular stochastic control problem*, SIAM J. Control Optim., (1989), pp. 876–907.

[13] ———, *A free boundary problem related to singular stochastic control: The parabolic case*, Comm. Partial Differential Equations, 16 (1991), pp. 373–424.

[14] P. L. J. Van Moerbeke, *On optimal stopping & free boundary problems*, Arch. Rational Mech. Anal., 60 (1976), pp. 101–148.

[15] M. L. Vidale and H. B. Wolfe, *An operations research study of sales response to advertising*, Oper. Res., 5 (1957), pp. 370–381.

[16] H. Zhu, *Dynamic Programming and Variational Inequalities in Singular Stochastic Control*, Ph.D. thesis, Brown University, 1991.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 516–537, March 1994

011

ON THE GRADIENT PROJECTION METHOD FOR OPTIMAL CONTROL PROBLEMS WITH NONNEGATIVE $\mathcal{L}^2$ INPUTS*

T. TIAN† AND J. C. DUNN†

Abstract. Local convergence and active constraint identification theorems are proved for gradient-projection iterates in the cone of nonnegative $\mathcal{L}^2$ functions on $[0,1]$. The theorems are based on recently established infinite-dimensional extensions of the Kuhn–Tucker sufficient conditions and are directly applicable to a large class of continuous-time optimal control problems with smooth nonconvex nonquadratic objective functions and Hamiltonians that are quadratic in the control input u.

Key words. gradient projection, infinite-dimensional programs, nonnegativity constraints, nonconvex objectives, $\mathcal{L}^\infty$-local convergence, $\mathcal{L}^2$-local convergence, active constraint identification, optimal control

AMS subject classifications. 49M07, 49M10, 49K15, 65K10, 90C06

1. Introduction.

Reference [1] investigates the local convergence properties of an unscaled gradient projection (GP) algorithm for minimizing nonconvex C^2 real functions over the nonnegative orthant in $\mathbb{R}^n$. Here we consider the same gradient-projection scheme for analogous infinite-dimensional problems,

$$\min_{u\in\Omega} J(u), \tag{1A}$$

$$\Omega = \{u \in \mathcal{L}^2(0,1) : u(t) \geq 0 \text{ a.e. in } [0,1]\}. \tag{1B}$$

A comprehensive local convergence theory already exists for GP algorithms and constrained minimization problems with *convex* objective functions and closed convex feasible sets in a general real Hilbert space [2]. Portions of this theory have also been extended to problems with smooth nonconvex objective functions and feasible sets with embedded open facets [3], [4], or to feasible sets prescribed by a *finite* number of smooth inequality constraints [5]–[7]; however, further extensions are not readily made for sets of the form

$$\Omega = \{u \in \mathcal{L}^p([0,1],\mathbb{R}^m) : u(t) \overset{a.e}{\in} U\}, \tag{2}$$

where U is a closed convex set prescribed by finitely many smooth inequality constraints in $\mathbb{R}^m$. An improved understanding of algorithm convergence behavior in (2) is needed because GP schemes are readily implemented for many continuous-time optimal control problems with feasible sets of this kind [1], [8], [9], because infinite-dimensional problems are typically richer in good and bad solution types and algorithm behavior than their finite-dimensional counterparts [10], [11], and because convergence features that are characteristically infinite-dimensional are nevertheless often *almost* present in large-scale algorithm implementations in approximating finite-dimensional spaces [12], [13]. We are interested in problem (1) because its feasible

* Received by the editors July 7, 1992; accepted for publication (in revised form) November 30, 1992. This research was supported by National Science Foundation research grants DMS-9002848 and DMS-9205240.

† Mathematics and Computer Science Department, Fort Hayes State University, Hayes, Kansas 67601, and Mathematics Department, Box 8205, North Carolina State University, Raleigh, North Carolina 27695-8205.

set is a prototype for the class of closed convex sets (2), and because nonnegativity constraints are important in their own right in many applications.

The analysis in [1] is fundamentally tied to the equivalence of the Euclidean and Chebychev norms in $\mathbb{R}^n$, and to the Kuhn–Tucker second-order sufficient conditions for local optimality in the nonnegative orthant. On the other hand, the $\mathcal{L}^2$ and $\mathcal{L}^\infty$ norms are not equivalent; the formal extension of the Kuhn–Tucker sufficient conditions in Ω are generally not sufficient even for $\mathcal{L}^\infty$- local optimality [14], [15]; and the proof strategies in [1] are therefore not applicable to problems (1) with general unstructured C^2 objective functions. Nevertheless, we are able to develop results similar to those in [1] for nonconvex C^2 functionals J with Hessians and gradients that satisfy the following conditions at each u in $\mathcal{L}^2(0,1)$: For all v in $\mathcal{L}^2(0,1)$

$$\left[\nabla^2 J(u)v\right](t) \overset{a.e.}{=} S(u)(t)v(t) + \int_0^1 K(u)(t,s)v(s)ds, \tag{3A}$$

where

$$S(u) \in \mathcal{L}^\infty(0,1), \tag{3B}$$

$$K(u) \in \mathcal{L}^2\left([0,1]\times[0,1]\right), \tag{3C}$$

$$K(u)(t,s) \overset{a.e.}{=} K(u)(s,t), \tag{3D}$$

$$\lim_{\|v-u\|_2\to 0} \|S(v) - S(u)\|_\infty = 0, \tag{3E}$$

$$\lim_{\|v-u\|_2\to 0} \|K(v) - K(u)\|_2 = 0, \tag{3F}$$

and

$$\nabla J(u)(t) \overset{a.e.}{=} \phi(u)(t) + S(u)(t)u(t), \tag{4A}$$

with

$$\phi(u) \in \mathcal{L}^\infty(0,1), \tag{4B}$$

$$\lim_{\|v-u\|_2\to 0} \|\phi(v) - \phi(u)\|_\infty = 0. \tag{4C}$$

Conditions (3) and (4) are met by an important class of continuous-time optimal control problems with Bolza objective functions and associated Hamiltonians that are quadratic in the control input $u(t)$. These problems are described further in §6. Extensions of the Kuhn–Tucker sufficient conditions for $\mathcal{L}^\infty$-local optimality in Ω are proved in [15] for J's that satisfy (3) and mild topological restrictions on $S(u)(\cdot)$ and the null set for u. Sufficient conditions for $\mathcal{L}^2$-local optimality are also established in [15] when (3C) is replaced by the stronger condition

$$K(u) \in \mathcal{L}^\infty\left([0,1]\times[0,1]\right). \tag{3C}'$$

These results are summarized in §2. In §§3 and 4, we use the Kuhn–Tucker conditions, the gradient representation formulas (4) and proof techniques from [15] and [16] to establish $\mathcal{L}^\infty$ and $\mathcal{L}^2$ local convergence theorems and convergence rate estimates for the projected gradient iteration

$$u \to G(u), \tag{5A}$$

where

$$G(u) = g\left(a(u), u\right), \tag{5B}$$

$$g(a,u) = P_\Omega\left(u - a\nabla J(u)\right), \tag{5C}$$

$$\left(P_\Omega v\right)(t) \overset{a.e.}{=} P_{[0,\infty)}v(t) = \max\{0, v(t)\}, \tag{5D}$$

and $a(\cdot)$ is a positive real-valued function defined by Bertsekas' modification of the Armijo step length rule [17], [1], i.e.,

$$a(u) = \min\ a \tag{5E}$$

subject to

$$a \in \{\overline{a}, \overline{a}\beta, \overline{a}\beta^2, \cdots\} \tag{5F}$$

and

$$J(u) - J\left(g(a,u)\right) \geq \sigma\langle \nabla J(u), u - g(a,u)\rangle \tag{5G}$$

with $\overline{a}$ fixed in $(0,\infty)$, and σ and β fixed in $(0,1)$.

Our development parallels [16], with the following exceptions. In [16], it is assumed that $K(u)$ is *bounded* on the square $[0,1]\times[0,1]$, and that

$$\lim_{\|v-u\|_2 \to 0} \sup_{[0,1]\times[0,1]} |K(v)(s,t) - K(u)(s,t)| = 0. \tag{3F)$'}$$

This stronger alternative to the $\mathcal{L}^2$ continuity condition (3F) secures the identity

$$\nabla J(v)(t) - \nabla J(u)(t) \overset{a.e.}{=} \int_0^1 \left[\nabla^2 J\left(u + \tau(v-u)\right)(v-u)\right](t)d\tau, \tag{6}$$

which, in [16], takes the place of the present formulas (4); however, our modified proof technique based on (4) yields theorems that apply to a somewhat larger class of optimal control problems. In addition, the $\mathcal{L}^2$-local convergence theorem in §4 establishes a uniform bound on the number of iterations that must elapse before geometric convergence ensues, and the proof given here does *not* require an essentially bounded starting point $u^{(0)}$ for (5). We also provide a new example in §3, a new active constraint identification result in §5, and a treatment of optimal control problems with nonquadratic Bolza objective functions in §6.

2. Sufficient conditions. We note that if $\langle u, v\rangle$ is the standard $\mathcal{L}^2$ inner product $\int_0^1 u(t)v(t)dt$, and if J has a second Gâteaux differential,

$$d^2J(u; v, w) = \langle v, \nabla^2 J(u)w\rangle$$

with a bounded linear operator $\nabla^2 J(u) : \mathcal{L}^2(0,1) \to \mathcal{L}^2(0,1)$ satisfying (3), then J is twice continuously Fréchet differentiable on $\mathcal{L}^2(0,1)$, $\nabla^2 J(u)$ is the Hessian of J at u, and consequently

$$J(v) - J(u) = \langle \nabla J(u), v - u\rangle + \frac{1}{2}\langle v - u, \nabla^2 J(u)(v - u)\rangle + o(\|v - u\|_2^2). \tag{7}$$

This is observed in [15] and proved in [16]. From here onward, we assume that conditions (3) hold. Our goal is to base local convergence proofs for GP sequences in the set Ω in (1B) on (3), (4), and the sufficient conditions for local optimality described below.

For each u in $\mathcal{L}^2(0,1)$ and $\epsilon > 0$, let

$$B_2(u, \epsilon) = \{v \in \mathcal{L}^2(0,1) : \|v - u\|_2 < \epsilon\},$$

$$B_\infty(u, \epsilon) = \{v \in \mathcal{L}^2(0,1) : \|v - u\|_\infty < \epsilon\},$$

where

$$\|u\|_2 = \langle u, u\rangle^{\frac{1}{2}},$$
$$\|u\|_\infty = \operatorname{ess\,sup}_{[0,1]} |u(t)|.$$

We say that u^* is an *$\mathcal{L}^2$-local minimizer* for J in Ω if and only if

$$\exists \delta > 0 \ \forall u \quad (u \in B_2(u^*, \epsilon) \cap \Omega \Rightarrow J(u) \geq J(u^*)).$$

Similarly, u^* is an *$\mathcal{L}^\infty$-local minimizer* for J in Ω if and only if

$$\exists \delta > 0 \ \forall u \quad (u \in B_\infty(u^*, \epsilon) \cap \Omega \Rightarrow J(u) \geq J(u^*)).$$

Every $\mathcal{L}^2$-local minimizer u^* is a *stationary point* for J in Ω, i.e.,

$$\forall u \in \Omega \quad \langle \nabla J(u^*), u - u^*\rangle \geq 0;$$

moreover, every $\mathcal{L}^\infty$-local minimizer u^* is also stationary [15]. We note that u^* is stationary if and only if for all $a > 0$, u^* is a fixed point of the map $g(a, \cdot)$ in (5) [3].

For each u in Ω, let

$$\alpha(u) = \{t \in [0,1] : u(t) = 0\},$$
$$T(u) = \{v \in \mathcal{L}^2(0,1) : v(t) = 0 \text{ a.e. in } \alpha(u)\}.$$

The set $\alpha(u)$ and the corresponding closed subspace $T(u)$ are analogous to the active constraint index sets and associated tangent spaces for the nonnegative orthant in $\mathbb{R}^n$ [1], and the following requirements can be seen as a formal extension of the Kuhn–Tucker sufficient conditions from the finite-dimensional orthant to the nonnegative cone Ω in $\mathcal{L}^2(0,1)$.

$$\nabla J(u^*)(t) = 0 \quad \text{a.e. in } \alpha(u^*)^c, \tag{8A}$$

(8B) $$\nabla J(u^*)(t) \geq 0 \quad \text{a.e. in } \alpha(u^*),$$

(8C) $$\forall \beta \subset INT\, \alpha(u^*)\ (\beta \text{ closed} \Rightarrow \exists c_1 > 0,\ \nabla J(u^*)(t) \geq c_1 \quad \text{a.e. in } \beta),$$

(8D) $$\exists c_2 > 0\ \forall v \in T(u^*), \quad \langle v, \nabla^2 J(u^*)v\rangle \geq 2c_2\|v\|_2^2,$$

where $\alpha(u^*)^c = [0,1] \setminus \alpha(u^*)$ = the complement of $\alpha(u^*)$ in $[0,1]$. We note that (8C) is considerably weaker than the obvious formal extension of strict complementarity in $\mathbb{R}^n_+$, namely, that $\nabla J(u^*)(t)$ be bounded away from 0 almost everywhere in $\alpha(u^*)$. The latter condition *never* holds in the commonly encountered case where u^* and $\nabla J(u^*)$ are continuous, and the sets $\alpha(u^*)$ and $\alpha(u^*)^c$ have positive measure. On the other hand, condition (8C) is not incompatible with continuity of u^* and $\nabla J(u^*)$, since this condition allows $\nabla J(u^*)(t)$ to approach 0 as t approaches the frontier of $\alpha(u^*)$ within $\alpha(u^*)$ (see §3, Example 1).

The formal sufficient conditions (8) actually *are* sufficient for $\mathcal{L}^\infty$-local optimality in Ω when (3) holds, $\alpha(u^*)$ is closed, and $S(u^*)(\cdot)$ is continuous on the frontier of $\alpha(u^*)$ in $[0,1]$; moreover, if (3C)′ is also satisfied and $S(u^*)(\cdot)$ is positive and bounded away from 0 almost everywhere on [0,1], then conditions (8) become sufficient for $\mathcal{L}^2$-local optimality in Ω [15]. The $\mathcal{L}^\infty$-local optimality proof in [15] first shows that when (8D) holds, $S(u^*)(t)$ is positive and bounded away from 0 on $\alpha(u^*)^c$, and by continuous extension, on a larger set $\mathcal{O}_o$ that is open in $[0,1]$ and contains $\overline{\alpha(u^*)^c}$ (= the closure of $\alpha(u^*)^c$). Because of (3), it then follows that a coercivity condition like (8D) also holds on a larger subspace

(9) $$T_\alpha = \{v \in \mathcal{L}^2(0,1) : v(t) = 0 \text{ a.e. in } \alpha\} \supset T(u^*)$$

with $\alpha = \mathcal{O}^c$ for some open set $\mathcal{O}$ such that $\mathcal{O}_o \supset \mathcal{O} \supset \overline{\alpha(u^*)^c}$. According to (8A)–(8C), $\nabla J(u^*)(t)$ is positive and bounded away from 0 almost everywhere in α, and the first- and second-order terms in Taylor's formula (7) will now combine to produce $\mathcal{L}^2$- quadratic growth for J, uniformly in some $\mathcal{L}^\infty$ neighborhood of u^* in Ω, i.e.,

(10) $$\exists \epsilon > 0\ \exists c > 0\ \forall u \quad \left(u \in B_\infty(u^*, \epsilon) \cap \Omega \Rightarrow J(u) - J(u^*) \geq c\|u - u^*\|_2^2\right).$$

(Two-metric local optimality results of this kind have also been proved in [18] under different hypotheses for J.) Finally, when (3C)′ holds at u^* and $S(u^*)(t)$ is positive and bounded away from 0 on $\alpha(u^*)$ as well as $\alpha(u^*)^c$, the $\mathcal{L}^2$-local optimality proof in [15] uses Taylor's formula and (3) to establish the $\mathcal{L}^2$ counterpart of (10), i.e.,

(11) $$\exists \epsilon > 0\ \exists c > 0\ \forall u \quad \left(u \in B_2(u^*, \epsilon) \cap \Omega \Rightarrow J(u) - J(u^*) \geq c\|u - u^*\|_2^2\right).$$

(In this connection, we note that u^* is an $\mathcal{L}^2$-local minimizer of J in Ω *only if* $S(u^*)(t) \geq 0$ almost everywhere in [0,1] [15].)

The following theorem collects the sufficiency results needed in the convergence analysis of §§3 and 4; these results are immediate corollaries of Lemmas 1–3 and Theorem 4 in [15].

THEOREM 1. *Let J be a twice continuously Fréchet differentiable real function on $\mathcal{L}^2(0,1)$ satisfying conditions* (3). *Suppose that the formal Kuhn–Tucker sufficient conditions* (8) *hold at u^* in the nonnegative cone Ω, that the set $\alpha(u^*)$ is closed, and that $S(u^*)$ is continuous on the frontier of $\alpha(u^*)$ in* $[0,1]$. *Then $S(u^*)(t)$ is positive and bounded away from* 0 *almost everywhere in some open neighborhood of $\overline{\alpha(u^*)^c}$*

in $[0,1]$, *there are positive numbers* c_1 *and* c_2, *a closed set* $\alpha \subset INT\alpha(u^*)$, *and a corresponding closed subspace* T_α *in* (9) *such that*

$$\nabla J(u^*)(t) = 0 \quad a.e.\ in\ \alpha(u^*)^c, \tag{12A}$$

$$\nabla J(u^*)(t) \geq 0 \quad a.e.\ in\ \alpha(u^*), \tag{12B}$$

$$\nabla J(u^*)(t) \geq c_1 \quad a.e.\ in\ \alpha, \tag{12C}$$

$$\forall v \in T_\alpha \quad \langle v, \nabla^2 J(u^*)v\rangle \geq c_2\|v\|_2^2, \tag{12D}$$

and consequently, u^* *is a strict* $\mathcal{L}^\infty$*-local minimizer of* J *in* Ω *satisfying* (10). *Furthermore, if condition* (3C)′ *also holds, and if* $S(u^*)$ *is positive and bounded away from* 0 *almost everywhere on the* entire *interval* $[0,1]$, *then* u^* *is a strict* $\mathcal{L}^2$*-local minimizer satisfying* (11).

3. $\mathcal{L}^\infty$-local convergence. To motivate our convergence proof strategy, we first outline the local convergence analysis for gradient projection iterates in [1] and indicate where this analysis fails in the infinite-dimensional setting of problem (1). If u^* satisfies the Kuhn–Tucker sufficient conditions in the nonnegative orthant $\mathbb{R}^n_+$, then iterates of the GP counterpart of (5) that begin sufficiently near u^* are eventually confined to a region of the subspace tangent to $\mathbb{R}^n_+$ at u^* where the GP algorithm reduces to a convergent unconstrained Armijo steepest descent iteration for the restriction of J to the tangent space. The proof of this result in [1] rests on the following facts for C^2 objective functions:

(i) The bounded positive step lengths $a(u)$ in (5) are bounded away from 0 in sufficiently small neighborhoods of stationary points u^*.

(ii) If the strict complementarity condition

$$\inf_{\{i:u_i^*=0\}} \frac{\partial J}{\partial u_i}(u^*) > 0$$

holds at a stationary point u^* in $\mathbb{R}^n_+$, then

$$\lim_{\substack{u\to u^*\\ u\in\mathbb{R}^n_+}} \sup_{\{i:u_i^*=0\}} \left(u_i - a(u)\frac{\partial J}{\partial u_i}(u)\right) < 0$$

and

$$\lim_{\substack{u\to u^*\\ u\in\mathbb{R}^n_+}} \inf_{\{i:u_i^*>0\}} \left(u_i - a(u)\frac{\partial J}{\partial u_i}(u)\right) > 0.$$

(iii) Every nonsingular unconstrained local minimizer u^* is a *stable* fixed point of the steepest descent map with Armijo steplengths, i.e., iterate sequences $\{u^{(k)}\}$ generated by this map will remain in any specified arbitrarily small neighborhood of u^*, provided $u^{(0)}$ lies in a sufficiently small neighborhood of u^*.

(iv) Unconstrained Armijo steepest descent iterates that remain sufficiently near a nonsingular minimizer u^* must converge to u^*.

In (ii), the limits can be taken in any norm, since all norms are equivalent in $\mathbb{R}^n$.

Assertion (i) can be demonstrated for the general version of (5) in *arbitrary* nonempty closed convex subsets of a real Hilbert space [5]. Similarly, Assertions (iii) and (iv) are true for nonsingular minimizers in $\mathbb{R}^n$ [19], and more generally for uniformly proper local minimizers that are also uniformly isolated stationary points in closed convex subsets of a Hilbert space [3], [7]. In particular, if u^* satisfies (8A) and (8D) in the formal Kuhn–Tucker sufficient conditions, then u^* is a nonsingular local minimizer for the restriction $J|_{T(u^*)}$ in the closed subspace $T(u^*)$, and hence a stable local attractor for the corresponding Armijo steepest descent iterates in $T(u^*)$. On the other hand, $\mathcal{L}^2$ continuity of $\nabla J(\cdot)$ and the sufficient conditions in Theorem 1 do *not* imply $\mathcal{L}^2$ or $\mathcal{L}^\infty$ analogs of Assertions (ii), since the quotients $\|w\|_\infty/\|w\|_2$ are unbounded on $\mathcal{L}^\infty(0,1) \setminus \{0\}$, and the values $\nabla J(u^*)(t)$ and $u^*(t)$ are typically not bounded away from 0 on the infinite sets $\alpha(u^*)$ and $\alpha(u^*)^c$, respectively. As a consequence, the map G in (5) need not send small $\mathcal{L}^2$ (or $\mathcal{L}^\infty$) neighborhoods of u^* into $T(u^*)$, and need not reduce to the Armijo steepest descent map for $J|_{T(u^*)}$ in small $\mathcal{L}^2$ (or $\mathcal{L}^\infty$) neighborhoods of u^* in $T(u^*)$. The proof strategy in [1] therefore fails for problem (1), even when J is a C^2 objective function satisfying (3). In this section and the next, we develop modified proof schemes for J's that satisfy (3) *and* the gradient representation formulas (4), and for u^* that meet the $\mathcal{L}^\infty$ or $\mathcal{L}^2$ sufficient conditions in Theorem 1. The modified proofs establish circumstances under which the iterates of (5) are eventually confined not to $T(u^*)$, but to a *larger* subspace T, or more precisely, to a closed convex G-invariant subset of T in which $J|_T$ is convex and (5) reduces to a convergent GP iteration for $J|_T$.

We begin with an $\mathcal{L}^\infty$-$\mathcal{L}^2$ variant of the general Hilbert space stability definition in [3], and then prove associated stability and convergence results for (5) and (1).

DEFINITION 1. *A stationary point u^* is $\mathcal{L}^\infty$-$\mathcal{L}^2$ stable for the GP algorithm* (5) *if and only if for each $\epsilon > 0$, there exists a $\delta > 0$ such that for all sequences $\{u^{(k)}\}$ generated by* (5),

$$u^{(0)} \in B_\infty(u^*, \delta) \cap \Omega \Rightarrow \left(\forall k \geq 0 \quad u^{(k)} \in B_2(u^*, \epsilon) \cap \Omega\right).$$

THEOREM 2. *Let $J : \mathcal{L}^2(0,1) \to \mathbb{R}^1$ be a twice continuously Fréchet differentiable function satisfying conditions* (3) *and* (4). *Suppose that the sufficient conditions for $\mathcal{L}^\infty$-local optimality in Theorem* 1 *hold at the point u^* in the nonnegative $\mathcal{L}^2$ cone Ω, and let T_α and G be the closed subspace in Theorem* 1 *and the gradient projection map in the algorithm* (5). *Then for every $\epsilon > 0$, there is a $\delta > 0$ and a G-invariant set, $\mathcal{I} \subset \overline{B_2(u^*, \epsilon)} \cap \Omega \cap T_\alpha$, such that $u^* \in \mathcal{I}$, and for all sequences $\{u^{(k)}\}$ generated by* (5),

$$u^{(0)} \in B_\infty(u^*, \delta) \cap \Omega \Rightarrow \left(\forall k \geq 1 \quad u^{(k)} \in \mathcal{I}\right).$$

It follows that u^* is $\mathcal{L}^\infty$-$\mathcal{L}^2$ stable for (5).

Proof. Let α, c_1, and c_2 be the measurable set and positive numbers in Theorem 1. For $\epsilon > 0$, define

$$\mathcal{L}_\epsilon = \{u \in \Omega : J(u) - J(u^*) \leq \tfrac{1}{4} c_2 \epsilon^2\}$$

and

$$\mathcal{I}_\epsilon = \overline{B_2(u^*, \epsilon)} \cap T_\alpha \cap \mathcal{L}_\epsilon.$$

We will prove the theorem by first showing that for ϵ sufficiently small,

$$G[\mathcal{I}_\epsilon] \subset \mathcal{I}_\epsilon \tag{13}$$

and then showing that for every $\epsilon > 0$ there is a $\delta > 0$ for which

$$G[B_\infty(u^*, \delta) \cap \Omega] \subset \mathcal{I}_\epsilon. \tag{14}$$

Since u^* is stationary and $u^* \in T(u^*) \subset T_\alpha$, condition (12D) and Taylor's formula (7) establish that for some $\epsilon_0 > 0$ and for all u,

$$u \in B_2(u^*, \epsilon_0) \cap \Omega \cap T_\alpha \Rightarrow J(u) - J(u^*) \geq \tfrac{1}{4} c_2 \|u - u^*\|_2^2. \tag{15}$$

Condition (4) implies that for all u,

$$|\nabla J(u)(t) - \nabla J(u^*)(t)| \overset{a.e.}{\leq} |\phi(u)(t) - \phi(u^*)(t)| + |S(u)(t)u(t) - S(u^*)(t)u^*(t)| \tag{16}$$

and therefore

$$u \in T_\alpha \Rightarrow |\nabla J(u)(t) - \nabla J(u^*)(t)| \leq |\phi(u)(t) - \phi(u^*)(t)| \text{ a.e. in } \alpha \subset \alpha(u^*). \tag{17}$$

According to (4), (12C), and (17), there is an $\epsilon_1 \in (0, \epsilon_0]$ such that for all u

$$\begin{aligned} u \in B_2(u^*, \epsilon_1) \cap \Omega \cap T_\alpha &\Rightarrow \nabla J(u)(t) \geq 0 \text{ a.e. in } \alpha \subset \alpha(u^*) \\ &\Rightarrow G(u)(t) = 0 \text{ a.e. in } \alpha \\ &\Rightarrow G(u) \in T_\alpha. \end{aligned}$$

Since u^* is stationary and $\nabla J(\cdot)$ is $\mathcal{L}^2$-continuous, it follows that $G(u^*) = u^*$, G is $\mathcal{L}^2$- continuous at u^* [3], and there is an $\epsilon_2 \in (0, \epsilon_1]$ such that for all u,

$$u \in B_2(u^*, \epsilon_2) \cap \Omega \cap T_\alpha \Rightarrow G(u) \in B_2(u^*, \epsilon_0) \cap \Omega \cap T_\alpha.$$

Since G also has the descent property

$$\forall u \in \Omega \quad J(G(u)) \leq J(u),$$

we therefore find that for all ϵ in $(0, \epsilon_2)$ and all u,

$$\begin{aligned} u \in \mathcal{I}_\epsilon \Rightarrow G(u) \in B_2(u^*, \epsilon_0) \cap \Omega \cap T_\alpha \text{ and} \\ \tfrac{1}{4} c_2 \|G(u) - u^*\|_2^2 \leq J(G(u)) - J(u^*) \leq \tfrac{1}{4} c_2 \epsilon^2 \\ \Rightarrow G(u) \in \mathcal{I}_\epsilon. \end{aligned} \tag{18}$$

This proves (13) and shows that $\mathcal{I}_\epsilon$ is a G-invariant set for all ϵ in $(0, \epsilon_2)$.

To prove (14), we note that J is $\mathcal{L}^2$-continuous, and hence for every $\epsilon > 0$ there is a $\rho > 0$ such that

$$B_2(u^*, \rho) \cap \Omega \cap T_\alpha \subset \mathcal{I}_\epsilon. \tag{19}$$

With a slight modification of the proof of in [5, Lem. A2], it can be shown that there are positive numbers $\underline{a}$ and δ_0 such that for all u,

$$u \in B_2(u^*, \delta_0) \cap \Omega \Rightarrow a(u) \geq \underline{a}. \tag{20}$$

By (4), (12C), and (16), there is a $\delta_1 \in (0, \delta_0]$ such that for all u,

$$\begin{aligned} u \in B_\infty(u^*, \delta_1) \cap \Omega &\Rightarrow |\nabla J(u)(t) - \nabla J(u^*)(t)| \overset{a.e.}{\leq} \tfrac{1}{2} c_1 \\ &\Rightarrow \nabla J(u)(t) \geq \tfrac{1}{2} c_1 \text{ a.e. in } \alpha. \end{aligned}$$

Let $\delta_2 = \min\{\delta_1, \frac{1}{2}\underline{a}c_1\}$. Since $B_\infty(u^*, \delta_2) \subset B_2(u^*, \delta_2)$, it follows that for all u,

$$\begin{aligned} u \in B_\infty(u^*, \delta_2) \cap \Omega &\Rightarrow u(t) - a(u)\nabla J(u)(t) \leq 0 \text{ a.e. in } \alpha \\ &\Rightarrow G(u)(t) = 0 \text{ a.e. in } \alpha \\ &\Rightarrow G(u) \in \Omega \cap T_\alpha. \end{aligned}$$

We have already seen that $G(u^*) = u^*$ and G is $\mathcal{L}^2$-continuous at u^*. Therefore, for every $\rho > 0$ there is $\delta \in (0, \delta_2]$ such that for all u,

$$u \in B_\infty(u^*, \delta) \cap \Omega \Rightarrow G(u) \in B_2(u^*, \rho) \cap \Omega \cap T_\alpha. \tag{21}$$

In view of (19), this proves (14). □

THEOREM 3. *Suppose that J and u^* satisfy the hypotheses of Theorem 2. Then for some $\delta > 0$ and $\lambda \in [0, 1)$, and all sequences $\{u^{(k)}\}$ generated by (5),*

$$u^{(0)} \in B_\infty(u^*, \delta) \cap \Omega \Rightarrow \left(\forall k \geq 1 \quad J(u^{(k+1)}) - J(u^*) \leq \lambda(J(u^{(k)}) - J(u^*))\right).$$

Furthermore, for $k \geq 1$ the norms $\|u^{(k)} - u^*\|_2$ are bounded above by a real sequence that converges to 0 geometrically, with ratio $\lambda^{\frac{1}{2}}$.

Proof. We will show that for sufficiently small $\epsilon > 0$, $J|_{T_\alpha}$ is convex and satisfies the growth condition (15) on $\overline{B_2(u^*, \epsilon)} \cap \Omega \cap T_\alpha$, that the G-invariant set $\mathcal{I}_\epsilon$ in Theorem 2 is closed and convex, and that for all u in $\mathcal{I}_\epsilon$,

$$J|_{T_\alpha}(u) - J|_{T_\alpha}(G(u)) \geq \sigma\langle \nabla(J|_{T_\alpha})(u), u - G(u)\rangle, \tag{22}$$

$$G(u) = P_{\mathcal{I}_\epsilon}\left(u - a(u)\nabla\left(J|_{T_\alpha}\right)(u)\right) \tag{23}$$

with $a(u)$ determined by (5). Our assertions then follow from Theorem 2, the local steplength bound (20), and a result in [2] for GP algorithms and convex programs in Hilbert spaces.

Let α, c_1, c_2, and $\underline{a}$ be the measurable set and positive numbers in Theorem 1 and condition (20). In view of (12D), (15), (18), (20), and the continuity of $\nabla^2 J(\cdot)$, there is an $\epsilon > 0$ such that $\mathcal{I}_\epsilon$ is G-invariant, and for all u and v,

$$u \in \overline{B_2(u^*, \epsilon)} \cap T_\alpha \text{ and } v \in T_\alpha \Rightarrow \langle v, \nabla^2\left(J|_{T_\alpha}\right)(u)v\rangle = \langle v, \nabla^2 J(u)v\rangle \geq \tfrac{1}{2}c_2\|v\|_2^2, \tag{24}$$

$$u \in \mathcal{I}_\epsilon \Rightarrow J|_{T_\alpha}(u) - J|_{T_\alpha}(u^*) \geq \tfrac{1}{4}c_2\|u - u^*\|_2^2, \tag{25}$$

$$u \in \mathcal{I}_\epsilon \Rightarrow a(u) \geq \underline{a}. \tag{26}$$

According to (24), $J|_{T_\alpha}$ is convex on $\overline{B_2(u^*, \epsilon)} \cap T_\alpha$, and therefore $\mathcal{I}_\epsilon$ is closed and convex.

Since $\mathcal{I}_\epsilon$ is G-invariant, the difference $u - G(u)$ lies in T_α for all u in $\mathcal{I}_\epsilon$, and therefore

$$\begin{aligned}\langle \nabla J(u), u - G(u)\rangle &= \langle \nabla J(u), P_{T_\alpha}(u - G(u))\rangle \\ &= \langle P_{T_\alpha}\nabla J(u), u - G(u)\rangle \\ &= \langle \nabla (J|_{T_\alpha})(u), u - G(u)\rangle\end{aligned}$$

for all u in $\mathcal{I}_\epsilon$. Thus, (22) is implied by (5).

Since $G[\mathcal{I}_\epsilon] \subset \mathcal{I}_\epsilon$, it can be seen that

$$\min_{v\in\mathcal{I}_\epsilon} \|u - a(u)\nabla J(u) - v\|_2 \le \|u - a(u)\nabla J(u) - G(u)\|_2$$

for all u in $\mathcal{I}_\epsilon$. On the other hand, since $\mathcal{I}_\epsilon \subset \Omega$, conditions (5) imply that

$$\begin{aligned}\|u - a(u)\nabla J(u) - G(u)\|_2 &= \min_{v\in\Omega} \|u - a(u)\nabla J(u) - v\|_2 \\ &\le \min_{v\in\mathcal{I}_\epsilon} \|u - a(u)\nabla J(u) - v\|_2\end{aligned}$$

for u in $\mathcal{I}_\epsilon$. Consequently, for all u,

$$u \in \mathcal{I}_\epsilon \Rightarrow G(u) = P_{\mathcal{I}_\epsilon}(u - a(u)\nabla J(u)).$$

Furthermore, since $\mathcal{I}_\epsilon \subset T_\alpha$, we have for all z in $\mathcal{L}^2(0,1)$,

$$\begin{aligned}\|P_{\mathcal{I}_\epsilon}z - P_{T_\alpha}z\|_2^2 + \|P_{T_\alpha^\perp}z\|_2^2 &= \|P_{\mathcal{I}_\epsilon}z - z\|_2^2 \\ &= \min_{v\in\mathcal{I}_\epsilon} \|v - z\|_2^2 \\ &= \min_{v\in\mathcal{I}_\epsilon} \|v - P_{T_\alpha}z\|_2^2 + \|P_{T_\alpha^\perp}z\|_2^2.\end{aligned}$$

It follows that

$$\|P_{\mathcal{I}_\epsilon}z - P_{T_\alpha}z\|_2^2 = \min_{v\in\mathcal{I}_\epsilon} \|v - P_{T_\alpha}z\|_2^2$$

and therefore

$$P_{\mathcal{I}_\epsilon}z = P_{\mathcal{I}_\epsilon}(P_{T_\alpha}z)$$

for z in $\mathcal{L}^2(0,1)$. Thus, for all u in $\mathcal{I}_\epsilon$,

$$\begin{aligned}G(u) &= P_{\mathcal{I}_\epsilon}(u - a(u)\nabla J(u)) \\ &= P_{\mathcal{I}_\epsilon}(P_{T_\alpha}(u - a(u)\nabla J(u))) \\ &= P_{\mathcal{I}_\epsilon}(u - a(u)\nabla (J|_{T_\alpha})(u)),\end{aligned}$$

as claimed in (23).

Finally, let ρ and δ be positive numbers satisfying (19) and (21) with ϵ. If the sequence $\{u^{(k)}\}$ is generated by (5) with $u^{(0)} \in B_\infty(u^*, \delta) \cap \Omega$, then for $k \ge 1$, $u^{(k)}$ is confined to the closed convex G-invariant set $\mathcal{I}_\epsilon \subset T_\alpha$, where $J|_{T_\alpha}$ is convex and continuously Fréchet differentiable and (5) reduces to an iteration of the GP map,

$$u \to P_{\mathcal{I}_\epsilon}(u - a(u)\nabla (J|_{T_\alpha})(u))$$

with steplengths $a(u)$ satisfying (22)–(23) and (26). Our convergence claims now follow at once from the growth condition (25) and [2, Thm. 4.3]. □

Example 2 in [15] shows that C^2 objective functions J satisfying (3) and (4) can have $\mathcal{L}^\infty$-local minimizers u^* that are not $\mathcal{L}^2$-local minimizers in the non-negative $\mathcal{L}^2$ cone Ω. Every $\mathcal{L}^2$ neighborhood of such a u^* contains a point $u \in \Omega$ at which $J(u) < J(u^*)$. Since J is $\mathcal{L}^2$ continuous and (5) does not increase J, we see that GP iterations (5) beginning at u can not converge to u^* in the $\mathcal{L}^2$ norm. The following special case of Example 2 in [15] demonstrates that u^* can also be *unstable* for (5) in the conventional $\mathcal{L}^2$ sense.

Example 1. For $u \in \mathcal{L}^2(0,1)$, put

$$J(u) = \int_0^1 \left[r(t)u(t) + \frac{1}{2} S(t)u(t)^2 \right] dt$$

and

$$u^*(t) = \max\{0, -r(t)\}, \quad t \in [0,1],$$

where

$$r(t) = 1 - 2t, \quad t \in [0,1]$$

and

$$S(t) = \begin{cases} -1, & t \in [0, \frac{1}{2} - \Delta) \\ 1, & t \in [\frac{1}{2} - \Delta, 1] \end{cases}$$

with Δ fixed in $(0, \frac{1}{2})$. Then u^* is an $\mathcal{L}^\infty$-local minimizer of J in the nonnegative cone Ω, but u^* is *not* an $\mathcal{L}^2$-local minimizer in Ω [15]; moreover, for every δ, no matter how small, there is a v in $B_2(u^*, \delta) \cap \Omega$ such that

$$\lim_{k \to \infty} \|u^{(k)} - u^*\|_2 = \infty$$

for the GP sequence generated by (5) and beginning at v. To prove the latter assertion, we first note that J is quadratic, with $K(u)(t,s) = 0$, $S(u)(t) = S(t)$, and $\phi(u)(t) = 1 - 2t$ in (3)–(4). The gradient map $\nabla J(\cdot)$ is also $\mathcal{L}^2$-Lipschitz continuous (with Lipschitz constant 1), and this implies that the steplengths $a(u)$ in (5) are *globally* bounded away from 0 on Ω by some $\underline{a} > 0$ [2]. Furthermore, given $\delta > 0$, there are measurable sets $\theta \subset (0, \frac{1}{2} - \Delta)$ and functions $v \in \Omega$ such that,

$$\mu(\theta) > 0,$$

$$\forall t \in \theta \quad h(t) \stackrel{def}{=} v(t) - (1 - 2t) > 0,$$

$$\|v - u^*\|_2 \leq \delta.$$

If $\{u^{(k)}\}$ is generated by (5), with $u^{(0)} = v$, then a simple induction yields

$$\forall t \in \theta \; \forall k \geq 1 \quad u^{(k)}(t) \geq v(t) + k\underline{a}h(t),$$

$$\|u^{(k)} - u^*\|_2^2 \geq \int_\theta \left(u^{(k)}\right)^2 dt \geq k^2 \underline{a}^2 \int_\theta h(t)^2 dt.$$

Therefore,

$$\forall t \in \theta \quad \lim_{k\to\infty} |u^{(k)}(t) - u^*(t)| = \infty$$

and

$$\lim_{k\to\infty} \|u^{(k)} - u^*\|_2 = \infty.$$

4. $\mathcal{L}^2$-local convergence. The Hilbert space stability formulation in [3] has the following expression for (1) and (5).

DEFINITION 2. *A stationary point u^* is $(\mathcal{L}^2)$ stable for the GP algorithm* (5) *if and only if for each $\epsilon > 0$, there is a $\delta \in (0, \epsilon]$ such that for all sequences $\{u^{(k)}\}$ generated by* (5),

$$u^{(0)} \in B_2(u^*, \delta) \cap \Omega \Rightarrow \left(\forall k \geq 0 \quad u^{(k)} \in B_2(u^*, \epsilon) \cap \Omega\right).$$

As a corollary of Lemma 2.1 in [3], we find that a stationary point u^* is stable for (5) if u^* is a uniformly proper local minimizer of J in the nonnegative cone Ω, and in particular, if u^* satisfies (11). Thus, the sufficient conditions for $\mathcal{L}^2$-local optimality in Theorem 1 imply that u^* is stable. We will now prove a related $\mathcal{L}^2$-local convergence theorem for (5).

THEOREM 4. *Let $J : \mathcal{L}^2(0,1) \to \mathbb{R}^1$ be a twice continuously Fréchet differentiable function satisfying conditions* (3) *and* (4), *and suppose that the sufficient conditions for $\mathcal{L}^2$-local optimality in Theorem* 1 *hold at the point u^* in the nonnegative $\mathcal{L}^2$ cone Ω. Then there are positive numbers $\delta > 0$ and n, and a number λ in $[0,1)$, such that for all sequences $\{u^{(k)}\}$ generated by the GP algorithm* (5),

$$u^{(0)} \in B_2(u^*, \delta) \cap \Omega \Rightarrow \forall k > n \quad J(u^{(k+1)}) - J(u^*) \leq \lambda(J(u^{(k)}) - J(u^*)).$$

Furthermore, for $k > n$ the norms $\|u^{(k)} - u^*\|_2$ are bounded above by a real sequence that converges to 0 *geometrically, with ratio $\lambda^{\frac{1}{2}}$.*

Proof. Let α be the measurable set in Theorem 1. As noted earlier, u^* is stable and the iterates $\{u^{(k)}\}$ of (5) can be confined to any specified $\mathcal{L}^2$ neighborhood $\overline{B_2(u^*, \epsilon)} \cap \Omega$ by restricting $u^{(0)}$ to a suitable $\mathcal{L}^2$ subneighborhood $B_2(u^*, \delta)$. We will show that if ϵ is sufficiently small, then iterates confined to $\overline{B_2(u^*, \epsilon)} \cap \Omega$ will eventually satisfy the condition,

$$u^{(k)}(t) = 0, \text{ a.e. in } \alpha \setminus \theta$$

or equivalently,

$$u^{(k)} \in T_{\alpha\setminus\theta}$$

for some subset $\theta \subset \alpha$ with Lebesgue measure $\mu(\theta)$ so small that a coercivity condition like (24) holds on the larger closed subspace $T_{\alpha\setminus\theta} \supset T_\alpha$. Our claims are then established by applying the results of [2] once again, this time to $J|_{T_{\alpha\setminus\theta}}$ in the set $\overline{B_2(u^*, \epsilon)} \cap \Omega \cap T_{\alpha\setminus\theta}$.

Let $\underline{a}$, c, c_1, and c_2 be the positive numbers in Theorem 1 and conditions (11) and (20), let c_3 be a positive number that bounds $S(u^*)(t)$ away from 0 almost everywhere in $[0,1]$, and put

$$d = \min\{c_2, c_3\}.$$

Condition (4A) implies that

$$\nabla J(u)(t) \overset{a.e.}{\geq} \nabla J(u^*)(t) - |\phi(u)(t) - \phi(u^*)(t)| + S(u)(t)u(t) - S(u^*)(t)u^*(t).$$

Therefore, in view of (3E), (3F), (4C), (11), (20), and the continuity of $\nabla^2 J(\cdot)$, there is an $\epsilon > 0$ such that for all u in $\overline{B_2(u^*, \epsilon)} \cap \Omega$, and for all v,

$$\nabla J(u)(t) \geq \tfrac{1}{2}(c_1 + c_3 u(t)) \geq \tfrac{1}{2}c_1 \quad \text{a.e. in } \alpha, \tag{27}$$

$$a(u) \geq \underline{a}, \tag{28}$$

$$J(u) - J(u^*) \geq c\|u - u^*\|_2^2, \tag{29}$$

$$v \in T_\alpha \Rightarrow \langle v, \nabla^2 J(u)v\rangle \geq \tfrac{1}{2}c_2\|v\|_2^2, \tag{30}$$

$$S(u)(t) \overset{a.e.}{\geq} \tfrac{1}{2}c_3, \tag{31}$$

$$\|K(u) - K(u^*)\|_2 \leq \tfrac{1}{8}d. \tag{32}$$

Moreover, there is a $\rho > 0$ such that for all measurable sets $\omega \subset [0,1] \times [0,1]$,

$$\mu(\omega) \leq \rho \Rightarrow \left(\int\int_\omega K(u^*)^2(t,s)dtds\right)^{\frac{1}{2}} \leq \tfrac{1}{8}d. \tag{33}$$

Now let $\{u^{(k)}\}$ be a GP iterate sequence with range in $\overline{B_2(u^*, \epsilon)}$, and for $l > 0$ put

$$\theta_l = \{t \in \alpha : u^{(0)}(t) > l\}$$

and

$$\begin{aligned}\omega_l &= [(\alpha^c \cup \theta_l) \times (\alpha^c \cup \theta_l)] \setminus (\alpha^c \times \alpha^c) \\ &= (\alpha^c \times \theta_l) \cup (\theta_l \times \alpha^c) \cup (\theta_l \times \theta_l).\end{aligned}$$

We note that

$$\mu(\omega_l) = 2\mu(\alpha_c)\mu(\theta_l) + \mu^2(\theta_l)$$

and also

$$\mu(\theta_l)l^2 \leq \int_{\theta_l} \left(u^{(0)}(t)\right)^2 dt \leq \epsilon^2,$$

since $u^{(0)} \in \overline{B_2(u^*, \epsilon)}$ and $u^*(t) = 0$ almost everywhere in $\alpha \supset \theta_l$. Consequently,

$$\mu(\omega_l) \leq \left(2\mu(\alpha^c) + \left(\frac{\epsilon}{l}\right)^2\right)\left(\frac{\epsilon}{l}\right)^2.$$

Given $\rho > 0$ in (33), let m be the unique positive root l of

$$\left(2\mu(\alpha^c) + \left(\frac{\epsilon}{l}\right)^2\right)\left(\frac{\epsilon}{l}\right)^2 = \rho.$$

Put $\theta = \theta_m$, $\omega = \omega_m$, $\nu = \max\{0, 1 - \frac{1}{2}\underline{a}c_3\} < 1$, and

$$n = \begin{cases} -\dfrac{\log\left(1 + \frac{2m(1-\nu)}{\underline{a}c_1}\right)}{\log \nu}, & \nu \geq 0 \\ 0, & \nu = 0 \end{cases}.$$

By constuction, and conditions (5), (27), (28), and (30)–(33), it follows that

$$\mu(\omega) \leq \rho$$

and for all k,

$$\begin{aligned} k > n &\Rightarrow u^{(k)}(t) = 0 \text{ a.e. in } \alpha \setminus \theta \\ &\Rightarrow u^{(k)} \in \overline{B_2(u^*, \epsilon)} \cap \Omega \cap T_{\alpha\setminus\theta} \end{aligned}$$

and for all u in $\overline{B_2(u^*, \epsilon)}$ and v in $T_{\alpha\setminus\theta}$,

$$\begin{aligned} \langle v, \nabla^2 J(u) v\rangle &= \int_{(\alpha\setminus\theta)^c} S(u)(t)v^2(t)dt + \int\int_{(\alpha\setminus\theta)^c\times(\alpha\setminus\theta)^c} K(u)(t,s)v(t)v(s)dtds \\ &= \int_{\alpha^c} S(u)(t)v^2(t)dt + \int\int_{\alpha^c\times\alpha^c} K(u)(t,s)v(t)v(s)dtds \\ &\quad + \int_{\theta} S(u)(t)v^2(t)dt + \int\int_{\omega} K(u)(t,s)v(t)v(s)ds \\ &\geq \frac{1}{2}d\int_{\alpha^c} v^2(t)dt + \frac{1}{2}d\int_{\theta} v^2(t)dt \\ &\quad - \left(\left(\int\int_{\omega} K^2(u^*)(t,s)dtds\right)^{\frac{1}{2}} + \|K(u) - K(u^*)\|_2\right)\int_{\alpha^c\cup\theta} v^2(t)dt \\ &\geq \frac{1}{4}d\int_{(\alpha\setminus\theta)^c} v^2(t)dt \\ &= \frac{1}{4}d\|v\|_2^2. \end{aligned}$$

We have shown that if $\{u^{(k)}\}$ has range in $\overline{B_2(u^*, \epsilon)} \cap \Omega$, then for $k > n$, $u^{(k)}$ is confined to the closed convex set $\overline{B_2(u^*, \epsilon)} \cap \Omega \cap T_{\alpha\setminus\theta}$, where $J|_{T_{\alpha\setminus\theta}}$ is convex and satisfies the growth condition,

$$J|_{T_{\alpha\setminus\theta}}(u) - J|_{T_{\alpha\setminus\theta}}(u^*) \geq c\|u - u^*\|_2^2.$$

As in the proof of Theorem 3, we can also see that for all $k > n$,

$$J|_{T_{\alpha\setminus\theta}}(u^{(k)}) - J|_{T_{\alpha\setminus\theta}}(u^{(k+1)}) \geq \sigma\langle \nabla\left(J|_{T_{\alpha\setminus\theta}}\right)(u^{(k)}), u^{(k)} - u^{(k+1)}\rangle$$

and

$$u^{(k+1)} = P_{\overline{B_2(u^*,\epsilon)}\cap\Omega\cap T_{\alpha\setminus\theta}}\left[u^{(k)} - a(u^{(k)})\nabla\left(J|_{T_{\alpha\setminus\theta}}\right)(u^{(k)})\right].$$

Our convergence assertions then follow from the stability of u^*, the uniform local step length bound (28), and [2, Thm. 4.3]. □

Note 1. The closed subspace $T_{\alpha\setminus\theta}$ in the proof of Theorem 4 varies with the starting iterate $u^{(0)}$ in $\overline{B_2(u^*, \epsilon)} \cap \Omega$; however, the number n and the convergence ratio λ in Theorem 4 do *not* depend on $u^{(0)}$.

5. Active constraint identification. Again, it is appropriate to begin our development with some remarks on the finite-dimensional analysis in [1]. If the strict complementarity condition holds at a stationary point u^* in the nonnegative orthant $\mathbb{R}^n_+$, and if the GP method in [1] generates a sequence $\{u^{(k)}\}$ that converges to u^*, then for k sufficiently large, and for $i = 1, \ldots, n$,

$$u_i^{(k)} = 0 \Leftrightarrow u_i^* = 0,$$

i.e., the iterates eventually "identify" the active constraints at u^*. Formal counterparts of this result are generally false in convex sets defined by infinitely many inequality constraints; however, there are circumstances under which infinite active constraint index sets are identified *asymptotically* by GP iterates that converge in some sense. A theorem of this kind is formulated in [20] for constrained compact fixed-point problems, and related mesh independence results are established in [21] for finite-dimensional approximations to continuous-time optimal control problems. Our theorem for (1) and (5) is stated in terms of the Lebesgue measure of symmetric set differences,

$$\alpha \Delta \beta \stackrel{def}{=} (\alpha^c \cap \beta) \cup (\alpha \cap \beta^c).$$

THEOREM 5. *Let $J : \mathcal{L}^2(0,1) \to \mathbb{R}^1$ be a continuously Fréchet differentiable function satisfying* (3B), (3E), *and* (4). *Suppose that* (8C) *holds at u^* in the nonnegative $\mathcal{L}^2$ cone Ω, and that $\alpha(u^*)$ is closed. If $\{u^{(k)}\}$ is a GP iterate sequence generated by* (5), *and if $u^{(k)}$ converges to u^* in the $\mathcal{L}^2$ norm, then*

$$\lim_{k \to \infty} \mu\left(\alpha(u^*) \Delta \alpha(u^{(k)})\right) = 0.$$

Proof. Since $\mathcal{L}^2$ convergence insures convergence in measure, we have

$$\forall \eta > 0 \quad \lim_{k \to \infty} \mu\{t \in \alpha(u^{(k)}) : u^*(t) \geq \eta\} = 0, \tag{34A}$$

$$\forall \eta > 0 \quad \lim_{k \to \infty} \mu\{t \in \alpha(u^*) : u^{(k)}(t) \geq \eta\} = 0. \tag{34B}$$

We will prove the theorem by showing that (34A) implies

$$\lim_{k \to \infty} \mu\left(\alpha(u^*)^c \cap \alpha(u^{(k)})\right) = 0, \tag{35A}$$

and that (34B) and the postulated conditions on J and u^* imply

$$\lim_{k \to \infty} \mu\left(\alpha(u^*) \cap \alpha(u^{(k)})^c\right) = 0. \tag{35B}$$

The function u^* is measurable, hence for every $\epsilon > 0$, there is an $\eta > 0$ such that

$$\mu\{t \in \alpha(u^{(k)}) : \eta > u^*(t) > 0\} \leq \mu\{t \in [0,1] : \eta > u^*(t) > 0\} \leq \epsilon,$$

$$\{t \in \alpha(u^{(k)}) : u^*(t) > 0\} = \{t \in \alpha(u^{(k)}) : u^*(t) \geq \eta\} \cup \{t \in \alpha(u^{(k)}) : \eta > u^*(t) > 0\}.$$

and therefore

$$\mu\{t \in \alpha(u^{(k)}) : u^*(t) > 0\} \leq \mu\{t \in \alpha(u^{(k)}) : u^*(t) \geq \eta\} + \epsilon.$$

According to (34A), we then have

$$0 \leq \overline{\lim_{k\to\infty}} \mu\{t \in \alpha(u^{(k)}) : u^*(t) > 0\} \leq \epsilon,$$

and since ϵ can be arbitrarily small, this proves (35A).

To prove (35B), we first note that $\alpha(u^*)^c$ is open in $[0,1]$ and $\overline{\alpha(u^*)^c}$ is measurable. Consequently,

$$\mu\left(\overline{\alpha(u^*)^c}\right) = \mu\left(\alpha(u^*)^c\right),$$

and for each $\epsilon > 0$ there is an open set $\mathcal{O}$ such that

$$\overline{\alpha(u^*)^c} \subset \mathcal{O},$$

and

$$\mu\left(\overline{\alpha(u^*)^c}\right) \leq \mu(\mathcal{O}) \leq \mu\left(\overline{\alpha(u^*)^c}\right) + \epsilon,$$

and therefore

$$\mu\left(\alpha(u^*)^c\right) \leq \mu(\mathcal{O}) \leq \mu\left(\alpha(u^*)^c\right) + \epsilon.$$

Fix ϵ and put $\gamma = \mathcal{O}^c$. Then γ is closed and

$$\gamma \subset \left(\overline{\alpha(u^*)^c}\right)^c = INT\ \alpha(u^*)$$

with

$$\begin{aligned}\mu\left(\alpha(u^*) \setminus \gamma\right) &= \mu\left(\alpha(u^*) \cap \mathcal{O}\right) \\ &= \mu(\mathcal{O}) - \mu\left(\alpha(u^*)^c \cap \mathcal{O}\right) \\ &\leq \epsilon.\end{aligned}$$

Now note that

$$\begin{aligned}\alpha(u^*) \cap \alpha(u^{(k+1)})^c = &\{t \in \alpha(u^*) \setminus \gamma : u^{(k+1)}(t) > 0\} \\ &\cup \{t \in \gamma : u^{(k+1)}(t) \geq \eta\} \cup \{t \in \gamma : \eta > u^{(k+1)}(t) > 0\}\end{aligned}$$

and therefore

(36)

$$\mu\left(\alpha(u^*) \cap \alpha(u^{(k+1)})^c\right) \leq \epsilon + \mu\{t \in \gamma : u^{(k+1)}(t) \geq \eta\} + \mu\{t \in \gamma : \eta > u^{(k+1)}(t) > 0\}.$$

We will complete our proof by showing that the rightmost term in this estimate is eventually bounded above by the middle term *at the previous iteration*, i.e.,

$$\mu\{t \in \gamma : \eta > u^{(k+1)}(t) > 0\} \leq \mu\{t \in \gamma : u^{(k)}(t) \geq \eta\}. \tag{37}$$

To see this, note that by (8C), there are positive numbers c_1 and η such that

$$\nabla J(u^*)(t) \geq c_1 \text{ a.e. in } \gamma$$

and

$$\frac{1}{2}c_1 - 2\eta\|S(u^*)\|_\infty \geq \frac{\eta}{\underline{a}}.$$

Therefore, by (3B), (3E), (4), and (5), there is a $\rho > 0$ such that for all u,

$$u \in B_2(u^*, \rho) \cap \Omega \Rightarrow \nabla J(u)(t) \geq \frac{\eta}{a} \text{ a.e. in } \{t \in \gamma : \eta > u(t) \geq 0\}$$
$$\Rightarrow G(u)(t) = 0 \text{ a.e. in } \{t \in \gamma : \eta > u(t) \geq 0\}.$$

Hence there is an N such that for all k,

$$k \geq N \Rightarrow u^{(k)} \in B_2(u^*, \rho) \cap \Omega$$
$$\Rightarrow u^{(k+1)}(t) = 0 \text{ a.e. in } \{t \in \gamma : \eta > u^{(k)}(t) \geq 0\}$$
$$\Rightarrow \mu\{t \in \gamma : \eta > u^{(k+1)}(t) > 0\} \leq \mu\{t \in \gamma : u^{(k)}(t) \geq \eta\}.$$

Thus (37) is true for sufficiently large k, and it follows from (34B), (36), and (37) that

$$0 \leq \overline{\lim_{k \to \infty}} \mu\left(\alpha(u^*) \cap \alpha(u^{(k)})^c\right) \leq \epsilon.$$

Since ϵ can be arbitrarily small, this establishes (35B) and completes the proof. □

6. Optimal control problems. The sufficient conditions in §2 and the GP convergence analysis of §§3–5 rest on the structure and smoothness properties (3)–(4). These properties are exhibited by a nontrivial class of Bolza objective functions for continuous-time optimal control problems with Hamiltonians that are quadratic in u.

Bolza objective functions are defined by rules of the form,

$$J(u) = P(x(u)(1)) + \int_0^1 f_0(t, x(u)(t), u(t))\, dt, \tag{38A}$$

where $x(u)$ is the solution of an initial value problem,

$$\frac{dx}{dt}(t) \overset{a.e.}{=} f(t, x(t), u(t)) \tag{38B}$$

$$x(0) = x^0 \tag{38C}$$

with u in a domain of $\mathcal{L}^p$ functions, and x^0 fixed in $\mathbb{R}^n$. In the present context, u is in $\mathcal{L}^2(0,1)$ and the functions P, f, and f_0 map $\mathbb{R}^n$ to $\mathbb{R}^1$, $([0,1] \times \mathbb{R}^n \times \mathbb{R}^1)$ to $\mathbb{R}^n$ and $([0,1] \times \mathbb{R}^n \times \mathbb{R}^1)$ to $\mathbb{R}^1$, respectively. To insure that J is twice continuously Fréchet differentiable and meets conditions (3) and (4) on $\mathcal{L}^2(0,1)$, we will assume that:

A1. P is twice continuously differentiable.

A2. For $i = 0, \cdots, n$,

$$f_i(t, x, u) = q_i(t, x) + r_i(t, x)u + s_i(t, x)u^2,$$

where the real valued functions q_i, r_i, and s_i and their first and second partial derivatives with respect to x are continuous on $[0,1] \times \mathbb{R}^n$.

A3. For $i = 1, \ldots, n$, the first partial derivatives of q_i, r_i, and s_i with respect to x are bounded on $[0,1] \times \mathbb{R}^n$.

Assumptions A2 and A3 imply the following growth properties for f and f_0:

G1. For every compact set $X \subset \mathbb{R}^n$, there are nonnegative numbers a_X, b_X, and c_X such that for (t,x,u) in $[0,1] \times X \times \mathbb{R}^1$, $0 \le i \le n$, and $1 \le j,k \le n$,

$$\left|\frac{\partial f_0}{\partial x_j}(t,x,u)\right| \le a_X + b_X|u| + c_X|u|^2,$$
$$\left|\frac{\partial f_i}{\partial u}(t,x,u)\right| \le b_X + c_X|u|,$$
$$\left|\frac{\partial^2 f_i}{\partial x_j \partial x_k}(t,x,u)\right| \le a_X + b_X|u| + c_X|u|^2,$$
$$\left|\frac{\partial^2 f_i}{\partial x_j \partial u}(t,x,u)\right| \le b_X + c_X|u|,$$
$$\left|\frac{\partial^2 f_i}{\partial u^2}(t,x,u)\right| \le c_X.$$

G2. There are nonnegative numbers a, b, and c such that for (t,x,u) in $[0,1] \times \mathbb{R}^n \times \mathbb{R}^1$ and $1 \le i, j \le n$

$$\left|\frac{\partial f_i}{\partial x_j}(t,x,u)\right| \le a + b|u| + c|u|^2.$$

With A1, A2, G1, G2, Gronwall's lemma and standard existence-uniqueness and dependence-on-parameters arguments from the theory of ordinary differential equations [22], [23], it can be shown that:

(i) For all u in $\mathcal{L}^2(0,1)$, the initial value problem (38B)–(38C) has a unique absolutely continuous solution $x(u)$ on $[0,1]$, and the integral in (38A) exists in Lebesgue's sense.

(ii) The mapping $x(\cdot) : \mathcal{L}^2(0,1) \to C^0([0,1],\mathbb{R}^n)$ is continuous.

(iii) For all u and v in $\mathcal{L}^2(0,1)$ the mapping $x\,(u + (\cdot)v) : \mathbb{R}^1 \to C^0([0,1],\mathbb{R}^n)$ is continuously Fréchet differentiable near 0 in $\mathbb{R}^1$, with

$$d^1x(u;v) \stackrel{def}{=} \lim_{s \to 0} \frac{x(u+sv) - x(u)}{s}, \quad = y(u;v) \tag{39A}$$

where $y(u;v)$ is the unique solution of the affine *equations of variation*,

$$\frac{dy}{dt}(t) \stackrel{a.e.}{=} A(u)(t)y(t) + B(u)(t)v(t), \tag{39B}$$

$$y(0) = 0, \tag{39C}$$

on $[0,1]$, with

$$A(u)(t) = \frac{\partial f}{\partial x}\,(t, x(u)(t), u(t)) \tag{39D}$$

and

$$B(u)(t) = \frac{\partial f}{\partial u}\,(t, x(u)(t), u(t))\,. \tag{39E}$$

(iv) For all u in $\mathcal{L}^2(0,1)$ the affine *adjoint* final value problem

$$\frac{d\psi}{dt}(t) \overset{a.e.}{=} -A(u)(t)^T\psi(t) - \nabla_x f_0\,(t, x(u)(t), u(t)) \tag{40A}$$

$$\psi(1) = \nabla P\,(x(u)(1)) \tag{40B}$$

has a unique absolutely continuous solution $\psi(u)$.

(v) The mapping $\psi(\cdot) : \mathcal{L}^2(0,1) \to C^0([0,1], \mathbb{R}^n)$ is continuous.

(vi) For all u and v in $\mathcal{L}^2(0,1)$ the mapping $\psi\,(u + (\cdot)v) : \mathbb{R}^1 \to C^0([0,1], \mathbb{R}^n)$ is continuously Fréchet differentiable near 0 in $\mathbb{R}^1$, with

$$\begin{aligned} d^1\psi(u;v) &\overset{def}{=} \lim_{s\to 0} \frac{\psi(u+sv) - \psi(u)}{s} \\ &= \zeta(u;v), \end{aligned} \tag{41A}$$

where $\zeta(u;v)$ is the unique absolutely continuous solution of the *adjoint* equations of variation,

$$\frac{d\zeta}{dt}(t) \overset{a.e}{=} -A(u)(t)^T\zeta(t) - Q(u)(t)y(u;v)(t) - R(u)(t)v(t), \tag{41B}$$

$$\zeta(1) = \nabla^2 P\,(x(u)(1))\,y(u;v)(1), \tag{41C}$$

with

$$Q(u)(t) = \nabla^2_{xx} H\,(t, \psi(u)(t), x(u)(t), u(t))\,, \tag{41D}$$

$$R(u)(t) = \nabla^2_{xu} H\,(t, \psi(u)(t), x(u)(t), u(t))\,, \tag{41E}$$

and

$$H(t, \psi, x, u) = \psi^T f(t, x, u) + f_0(t, x, u) \tag{41F}$$

for (t, ψ, x, u) in $[0,1] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^1$.

In the circumstances outlined above, the Bolza objective function J has first and second Gâteaux differentials with Riesz–Fréchet gradient and Hessian representors of the form (4A) and (3A). To see this, we note that

$$\begin{aligned} d^1J(u;v) = {} & \frac{\partial P}{\partial x}\,(x(u)(1))\,y(u;v)(1) \\ & + \int_0^1 \Big[\frac{\partial f_0}{\partial x}\,(t, x(u)(t), u(t))\,y(u;v)(t) \\ & \qquad + \frac{\partial f_0}{\partial u}\,(t, x(u)(t), u(t))\,v(t)\Big]\,dt \end{aligned} \tag{42A}$$

and

$$\frac{d}{dt}\left[\psi(t)^T y(t)\right] \overset{a.e.}{=} -\frac{\partial f_0}{\partial x}\,(t, x(u)(t), u(t))\,y(t) + \psi(t)^T B(u)(t)v(t) \tag{42B}$$

for any two solutions $y(\cdot)$ and $\psi(\cdot)$ of (39B) and (40A). By integrating (42B) from 0 to 1 and applying the boundary conditions in (39) and (40), we find that

$$d^1J(u;v) = \langle \nabla J(u), v\rangle, \tag{43A}$$

with

$$\begin{aligned}\nabla J(u)(t) &\overset{a.e.}{=} \nabla_u H\left(t, \psi(u)(t), x(u)(t), u(t)\right)\\ &= \phi(u)(t) + S(u)(t)u(t),\end{aligned} \tag{43B}$$

where

$$\phi(u)(t) = \psi(u)(t)^T r\left(t, x(u)(t)\right) + r_0\left(t, x(u)(t)\right), \tag{43C}$$

and

$$\begin{aligned}S(u)(t) &= 2\left[\psi(u)(t)^T s\left(t, x(u)(t)\right) + s_0\left(t, x(u)(t)\right)\right]\\ &= \nabla^2_{uu} H\left(t, \psi(u)(t), x(u)(t), u(t)\right).\end{aligned} \tag{43D}$$

Similarly, we have

$$\begin{aligned}d^2 J(u; v, w) &\overset{def}{=} \lim_{s\to 0} \frac{d^1 J(u+sv; w) - d^1 J(u; w)}{s}\\ &= \lim_{s\to 0} \frac{\langle \nabla J(u+sv) - \nabla J(u), w\rangle}{s}\\ &= \int_0^1 S(u)(t)v(t)w(t)dt\\ &\quad + \int_0^1 \left[\zeta(u; v)(t)^T B(u)(t) + y(u; v)(t)^T R(u)(t)\right] w(t)dt\end{aligned} \tag{44A}$$

and

$$\begin{aligned}\frac{d}{dt}[\zeta(u; v)(t)^T y(u; w)(t)] \overset{a.e.}{=}\ & -y(u; v)(t)^T Q(u)(t) y(u; w)(t) - v(t)R(u)(t)^T y(u; w)(t)\\ & + \zeta(u; v)(t)^T B(u)(t)w(t).\end{aligned} \tag{44B}$$

By integrating (44B) and applying (39C) and (41C), we obtain

$$\begin{aligned}d^2 J(u; v, w) =\ & y(u; v)(1)^T Q_1(u) y(u; w)(1)\\ & + \int_0^1 [y(u; v)(t)^T Q(u)(t) y(u; w)(t) + y(u; v)(t)^T R(u)(t)w(t)\\ & \quad + v(t)R(u)(t)^T y(u; w)(t) + S(u)(t)v(t)w(t)]dt,\end{aligned} \tag{45A}$$

with

$$Q_1(u) = \nabla^2 P\left(x(u)(1)\right). \tag{45B}$$

As in [15], (45) can be carried further by observing that

$$y(u; v)(t) = \int_0^1 \Phi(u)(t, \tau) B(u)(\tau) v(\tau) d\tau, \tag{46A}$$

where $\Phi(u)(t, \tau)$ are fundamental solution matrices uniquely prescribed by the initial value problems

$$\frac{\partial}{\partial t}\Phi(t, \tau) \overset{a.e.}{=} A(u)(t)\Phi(t, \tau), \tag{46B}$$

$$\Phi(\tau, \tau) = I \tag{46C}$$

for τ in $[0,1]$. Hence

$$d^2J(u;v) = \int_0^1 [\nabla^2 J(u)v](t)w(t)dt, \tag{47A}$$

with

$$[\nabla^2 J(u)v](t) \overset{a.e.}{=} S(u)(t)v(t) + \int_0^1 K(u)(t,s)v(s)ds, \tag{47B}$$

where

(47C)

$$\begin{aligned} K(u)(t,s) = \; & B(u)(t)^T\hat{\Phi}(u)(s,t)^T R(u)(s) + B(u)(s)^T\hat{\Phi}(u)(t,s)^T R(u)(t) \\ & + B(u)(t)^T \left[\int_{\max(t,s)}^1 \Phi(u)(\tau,t)^T Q(u)(\tau)\Phi(u)(\tau,s)d\tau\right] B(u)(s) \\ & + B(u)(t)^T\Phi(u)(1,t)^T Q_1(u)\Phi(u)(1,s)B(u)(s), \end{aligned}$$

and

$$\hat{\Phi}(u)(s,t) = \begin{cases} \Phi(u)(s,t), & t \le s \\ 0\,, & s < t \end{cases}. \tag{47D}$$

Thus far, we have shown that J has first and second Gâteaux differentials with representations (43) and (47) that are formally like (4A) and (3A). The remaining symmetry, integrability, and continuity conditions in (3) and (4) are now readily inferred from Assumptions A1–A3, and continuity of the maps $x(\cdot) : \mathcal{L}^2(0,1) \to C^0([0,1],\mathbb{R}^n)$ and $\psi(\cdot) : \mathcal{L}^2(0,1) \to C^0([0,1],\mathbb{R}^n)$. In particular, with reference to (43D), we see that $S(u)$ is in $C^0(0,1)$ and the associated map $S(\cdot) : \mathcal{L}^2(0,1) \to C^0(0,1)$ is continuous. Similarly, in view of (39D)–(39E), (41D)–(41F), (45B), (46B)–(46C), and properties G1–G2, it follows that for u in $\mathcal{L}^2(0,1)$, $A(u)$ and $Q(u)$ are in $\mathcal{L}^1([0,1],\mathbb{R}^{n\times n})$; $B(u)$ and $R(u)$ are in $\mathcal{L}^2([0,1],\mathbb{R}^{n\times 1})$; $\Phi(u)$ is in $C^0([0,1]\times[0,1],\mathbb{R}^{n\times n})$; $K(u)$ is in $\mathcal{L}^2([0,1]\times[0,1])$; and the corresponding maps $A(\cdot)$, $B(\cdot)$, $\Phi(\cdot)$, $Q(\cdot)$, $R(\cdot)$, and $K(\cdot)$ are continuous. Thus, Bolza objective functions (38) have the desired structure/continuity properties (3)–(4) when Assumptions A1–A3 hold.

Note 2. The foregoing conclusions are still valid if the smoothness restrictions in Assumption A2 are replaced by weaker conditions of the Carathéodory type [22]. In this case, the nonnegative constants in Properties G1–G2 become *functions* a and a_X in $\mathcal{L}^1(0,1)$, b and b_X in $\mathcal{L}^2(0,1)$, and c and c_X in $\mathcal{L}^\infty(0,1)$; and the resulting enlarged class of Bolza objective functions includes the general linear-quadratic regulator objectives treated in [16].

Note 3. For K in (47), condition (3C)′ will hold at u^* if u^* is in $\mathcal{L}^\infty(0,1)$, or if $s(t,x) = 0$ and $s_0(t,x)$ depends only on t (condition (3C)′ is invoked in Theorems 1 and 4).

REFERENCES

[1] D. P. BERTSEKAS, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Auto. Control, AC-10 (1976), pp. 174–184.

[2] J. C. DUNN, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368–400.

[3] J. C. DUNN, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203–215.

[4] J. V. BURKE AND J. J. MORÉ, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197–1211.

[5] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Appl. Math. Optim., 17 (1988), pp. 103–119.

[6] J. C. DUNN, *A subspace decomposition principle for scaled gradient projection methods: Global theory*, SIAM J. Control Optim., 29 (1991), pp. 1160–1175.

[7] ———, *A subspace decomposition principle for scaled gradient projection methods: Local theory*, SIAM J. Control Optim., 31 (1992), pp. 219–246.

[8] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.

[9] J. C. DUNN, *Gradient projection methods for systems optimization problems*, in Control and Dynamic Systems, 29, C. T. Leondes, ed., Academic Press, Orlando, FL, 1988.

[10] J. C. DUNN, *Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals*, SIAM J. Control Optim., 17 (1979), pp. 187–211

[11] G. C. HUGHES AND J. C. DUNN, *Newton–Goldstein convergence rates for convex constrained minimization problems with singular solutions*, Appl. Math. Optim., 12 (1984), pp. 203–230.

[12] J. C. DUNN, *Extremal types for certain $\mathcal{L}^p$-minimization problems and associated large scale nonlinear programs*, Appl. Math., Optim., 10 (1983), pp. 303–335.

[13] J. C. DUNN AND E. W. SACHS, *The effects of perturbations on the convergence rates of optimization algorithms*, Appl. Math., Optim., 10 (1983), pp. 143–157.

[14] H. MAURER AND J. ZOWE, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[15] J. C. DUNN AND T. TIAN, *Variants of the Kuhn–Tucker sufficient conditions in cones of non-negative functions*, SIAM J. Control Optim., 30 (1991), pp. 1361–1384.

[16] T. TIAN, *Convergence Analysis of a Projected Gradient Method for a Class of Optimal Control Problems*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 1992.

[17] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[18] H. MAURER, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[19] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, NY, 1982.

[20] C. T. KELLEY AND E. W. SACHS, *Multi-level Algorithms for Constrained Compact Fixed Point Problems*, preprint, 1992.

[21] ———, *Mesh independence of the gradient projection method for optimal control problems*, SIAM J. Control Optim., 30 (1992), pp. 477-493.

[22] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw Hill, New York, 1974.

[23] V. M. ALEKSEEV, V. M. TIKHOMIROV, AND S. V. FOMIN, *Optimal control*, Plenum Publishing Co., New York, 1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 538–552, March 1994

012

CONTINUOUS-TIME SHORTEST PATH PROBLEMS AND LINEAR PROGRAMMING*

A. B. PHILPOTT†

Abstract. Shortest path problems are considered for a graph in which edge distances can vary with time, each edge has a transit time, and parking (with a corresponding penalty) is allowed at the vertices. The problem is formulated as a continuous-time linear program, and a dual problem is derived for which the absence of a duality gap is proved. The existence of an extreme-point solution to the continuous-time linear program is also demonstrated, and a correspondence is derived between extreme points and continuous-time shortest paths. Strong duality is then derived in the case where the edge distances satisfy a Lipschitz condition.

Key words. shortest path, continuous linear programming, extreme point optimal solution, duality

AMS subject classifications. 05C38, 49A27, 49B36, 90C35, 90C48

1. Introduction.

Many problems in operations research require the solution of a shortest path problem: that of computing the shortest path between two vertices of a graph. This problem arises in a natural way in routing applications, and often as a subproblem that must be solved as part of some algorithm to solve a more complicated problem. Suppose c_{jk} is the distance along the edge joining j and k in a graph with n vertices. (We set $c_{jk} = \infty$ if this edge does not exist.) Then the classical shortest path problem can be formulated as the following linear program.

$$\text{SP: minimize} \sum_j \sum_k c_{jk} x_{jk}$$

$$\text{subject to} \sum_k x_{kj} - x_{jk} = \begin{cases} -1, & j = 1, \\ 0, & j \neq 1, n, \\ 1, & j = n, \end{cases}$$

$$x_{jk} \geq 0.$$

It is well known that as a consequence of the unimodularity of the constraint matrix the basic feasible solutions of SP have components which are zero or one. Furthermore it is easy to see that the nonzero components of any basic feasible solution determine a path from vertex 1 to vertex n, and every such path corresponds to some basic feasible solution. Thus an optimal basic feasible solution to SP gives variables x_{jk}, which indicate whether the edge joining j and k is on a shortest path from vertex 1 to vertex n.

The problem SP has the following dual problem:

$$\text{SP*: maximize } \pi_n - \pi_1$$

$$\text{subject to } \pi_k - \pi_j \leq c_{jk}.$$

The difference in dual variables $\pi_k - \pi_j$ can be interpreted as the shortest distance from vertex j to vertex k. Most solution techniques for SP use some labelling procedure to construct vertex by vertex a feasible solution to SP*, along with a complementary slack solution to SP. (See, e.g., [1] for a survey of efficient procedures for solving SP.)

* Received by editors March 11, 1991; accepted for publication (in revised form) December 29, 1992.

† Department of Engineering Science, University of Auckland, Auckland, New Zealand.

In this paper we consider a generalization of the problem SP to the case where there is a transit time on each edge, edge distances can vary as functions of time, and parking with some penalty is allowed at the vertices of the graph. We draw a distinction between edge distances and transit times, and assume that the transit times are constant with time. The motivation for our model comes from an application due to Mees [8] involving the scheduling of trains in a railway network. Here the transit times, which depend only on the speed that the trains traverse the arcs of the network, are independent of time but we discourage travel on given edges at certain times by imposing a time-varying penalty for traversing these edges.

A number of authors have considered generalizing SP to the time-varying case. Problems that seek a path with least total transit time have been studied in the absence of parking by Cooke and Halsey [3], who consider integer-valued transit times, and by Dreyfus [4] for real-valued transit times. A similar model described by Halpern [6] admits parking, with a unit penalty, only in certain specified time intervals, giving an effective penalty on parking that equals 1 or ∞. More recently, Orda and Rom have developed algorithms for the computation of minimum delay paths (see [9]), and minimum distance paths (see [10]), both for networks with time-varying transit times and a number of different parking models.

Given a graph G with n vertices and a time interval $[0,T]$, we define a *vertex-time pair* (VTP) to be a member of $\{1,2,\ldots,n\}\times[0,T]$. A *continuous-time path* from $(1,0)$ to (j,t) is a sequence of VTPs

$$(1,0)=(j_0,t_0),(j_1,t_1),\ldots,(j_p,t_p)=(j,t)$$

in which either $j_i = j_{i+1}$, in which case parking occurs at vertex j_i for the interval $[t_i,t_{i+1})$, or $j_i \neq j_{i+1}$, in which case traffic leaves vertex j_i for vertex j_{i+1} at time t_i and arrives at t_{i+1}. If the transit time between vertices j and k is denoted by τ_{jk} then $t_i + \tau_{j_i,j_{i+1}} = t_{i+1}$. (We assume for convenience that τ_{jk} and c_{jk} are defined for all j and k and are equal to zero when $j=k$.) The *length* of a continuous-time path is defined by

$$C=\sum_{j_i\neq j_{i+1}} c_{j_ij_{i+1}}(t_i)+\sum_{j_i=j_{i+1}}\int_{t_i}^{t_{i+1}} s_{j_i}(t)dt$$

where $c_{jk}(t)$ is the edge distance from j to k at time t, and $s_j(t)$ is the penalty on parking at node j at time t. The continuous-time shortest path problem seeks a continuous-time path that minimizes C for paths from $(1,0)$ to (n,T).

The presence of transit times in our model means that there are some VTPs (j,t) that will not appear in any path from $(1,0)$ to (n,T) because we cannot reach vertex j from 1 in time t, or we cannot reach vertex n from j in time $T-t$. The remaining VTPs are called *admissible* VTPs. We may characterize the set of admissible VTPs by computing for each vertex,

$$\rho_j=\min_{\alpha}\sum_{(p,q)\in P_\alpha}\tau_{pq},\qquad \sigma_j=T-\min_{\alpha}\sum_{(p,q)\in Q_\alpha}\tau_{pq},$$

where $\{P_\alpha\}$ is the set of paths in G from node 1 to node j, and $\{Q_\alpha\}$ is the set of paths in G from node j to node n. Thus ρ_j is the earliest time we may arrive at vertex j if we leave vertex 1 at time zero, and σ_j is the latest time that we may leave j so as to arrive at vertex n by time T. (Clearly $\rho_1=0$ and $\sigma_n=T$, and for every j and

k we have $\sigma_k \leq \sigma_j + \tau_{jk}$ and $\rho_k \leq \rho_j + \tau_{jk}$.) Admissible VTPs can now be defined as those (j,t) with $t \in [\rho_j, \sigma_j]$.

We note in passing that if the departure times in this model must be chosen from a finite set S, say, then the model is equivalent to a classical shortest path problem formulated in the time-expanded graph with vertices in $\{1,2,\ldots,n\} \times S$, and edge distances between vertices equal to infinity, except for the edges from (j,t) to (k,u) for $u \geq t + \tau_{jk}$, which have distance $c_{jk}(t) + \int_{t+\tau_{jk}}^{u} s_k(\tau)d\tau$.

In this paper we wish to address the following question. To what extent does the continuous-time shortest path problem described above have an equivalent formulation as a linear program? We show that if the problem is suitably posed as a linear program in an (infinite-dimensional) space of measures then the results outlined above for SP and SP* remain true. The analysis is based on the theory of linear programming in infinite-dimensional vector spaces as expounded in Anderson and Nash [2]. In the next section we introduce a linear programming formulation (CSP) and its dual (CSP*) and demonstrate a duality theorem that states that CSP and CSP* have the same value. We then show that the extreme-point solutions of CSP correspond to continuous-time paths. It follows that the optimal solution to CSP can be taken as a vector of measures with finite support. We conclude with a proof of strong duality in the case where the edge distances are Lipschitz functions and give a brief sketch of a computational technique for solving instances of CSP, the details of which may be found in [13] and [14].

2. Continuous-time linear programming. The central purpose of this paper is to show that the continuous-time shortest-path problem introduced in the previous section has an equivalent formulation as the following continuous-time linear programming problem:

$$\text{CSP: minimize } \sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} c_{jk}(t)dx_{jk}(t) + \sum_j \int_{\rho_j}^{\sigma_j} y_j(t)s_j(t)dt$$

subject to

$$(1) \quad \sum_k \int_{\rho_k}^{\sigma_j - \tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{\sigma_k - \tau_{jk}} dx_{jk}(\tau) = \begin{cases} -1, & j = 1, \\ 0, & j \neq 1, n, \\ 1, & j = n, \end{cases}$$

$$(2) \quad y_1(t) = 1 + \sum_k \int_{\rho_k}^{t-\tau_{k1}} dx_{k1}(\tau) - \int_{\rho_1}^{t} dx_{1k}(\tau), \quad t \in [\rho_1, \sigma_1],$$

$$(3) \quad y_j(t) = \sum_k \int_{\rho_k}^{t-\tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau), \quad t \in [\rho_j, \sigma_j], \quad j = 2,\ldots,n,$$

$$y_j(t) \geq 0, \quad t \in [\rho_j, \sigma_j], \quad x_{jk}(t) \geq 0, \quad t \in [\rho_j, \sigma_k - \tau_{jk}] \quad j,k = 1,\ldots,n.$$

Here the edge distances c_{jk} are taken to be continuous functions on $[\rho_j, \sigma_k - \tau_{jk}]$, each parking cost s_j is a bounded measurable function, and x_{jk} is a regular Borel measure on $[\rho_j, \sigma_k - \tau_{jk}]$, which indicates in an optimal solution whether a shortest path involves leaving vertex j for vertex k at time t. The variables y_j are intended to indicate whether parking occurs in vertex j. To see this observe that if for some vertex j we have $x_{kj}(\{t - \tau_{kj}\}) = x_{jk}(\{t\}) = 0$ for every vertex k and time t, except for two

vertices l and m, and two time instants $t_1 < t_2$ for which $x_{lj}(\{t_1 - \tau_{lj}\}) = x_{jm}(\{t_2\}) = 1$, then $y_j(t) = 1$, $t \in [t_1, t_2)$, and zero elsewhere. Such a solution corresponds to a path leaving vertex l for j at $t_1 - \tau_{lj}$, parking in j until t_2, and then leaving vertex j for m. By summing (2) and (3) and using the nonnegativity conditions on x and y, it is easy to show that for all j, $0 \leq y_j(t) \leq 1$. This ensures that the number of arrivals at each vertex (apart from vertex 1) up to time t is always greater by at most 1 than the number of departures up to t. In what follows we show that an optimal solution to CSP exists for which $y_j(t) \in \{0, 1\}$.

We will find occasion to characterize each measure x_{jk} by a Lebesgue–Stieltjes distribution function X_{jk}, which is of bounded variation on $[\rho_j, \sigma_k - \tau_{jk}]$ and continuous from the right on $(\rho_j, \sigma_k - \tau_{jk})$. We set $X_{jk}(\rho_j) = 0$, implying that $\int_{\rho_j}^{t} dx_{jk}(\tau) = X_{jk}(t)$. (Note that this is the same as $x_{jk}([\rho_j, t])$ and $\int_{[\rho_j, t]} dx_{jk}(\tau)$, except when $t = \rho_j$, when these expressions are not necessarily equal to $X_{jk}(\rho_j) = 0$, but are both equal to $x_{jk}(\{\rho_j\}) = \lim_{\epsilon \to 0} X_{jk}(\rho_j + \mid \epsilon \mid)$). To avoid introducing complicated notation, we often interpret the values of x and y to be zero at any point where they are not defined. In particular, we assume that for every j and k

$$x_{jk}([-\tau_{jk}, \rho_j)) = x_{jk}((\sigma_k - \tau_{jk}, T]) = 0. \tag{4}$$

We say that any nonnegative $\{x, y\}$ satisfying the constraints (1)–(3) is *feasible* for CSP, and define the *value* $V(\text{CSP})$ to be the infimum of the objective function over all feasible $\{x, y\}$. For CSP to have a feasible solution, T must be chosen sufficiently large, as shown by the following lemma.

LEMMA 2.1. CSP *has a feasible solution if and only if* $\rho_n \leq T$.

Proof. If $\rho_n \leq T$, then for some path P from 1 to n in the underlying graph G, $\sum_{(j,k) \in P} \tau_{jk} \leq T$. Let $\{1 = j_0, j_1, \ldots, j_m = n\}$ be the vertices in P, which we may assume to be distinct. Define $x_{jk} = 0$ for every j and k except where $j = j_u, k = j_{u+1}$ for some u, in which case

$$x_{jk}\left(\left\{\sum_{p=1}^{p=u} \tau_{j_{p-1} j_p}\right\}\right) = 1.$$

(Here the sum is void if $u = 0$.) This x and the y it generates are easily shown to be feasible for CSP.

Conversely, suppose (x,y) is feasible for CSP. We show $\rho_n \leq T$ by constructing a path P in G from 1 to n with $\sum_{(j,k) \in P} \tau_{jk} \leq T$. Setting $j_0 = n$, we may find a vertex $j_1 \neq n$ with $x_{j_1 j_0}([0, T - \tau_{j_1 j_0}]) > 0$, otherwise (1) fails to hold for $j = j_0$. Since $T - \tau_{j_1 j_0} \leq \sigma_{j_1}$, (3) and the condition $y_{j_1}(T - \tau_{j_1 j_0}) \geq 0$ yields

$$\sum_k \int_{\rho_k}^{T - \tau_{j_1 j_0} - \tau_{k j_1}} dx_{k j_1}(t) \geq \int_{\rho_{j_1}}^{T - \tau_{j_1 j_0}} \sum_k dx_{j_1 k}(t) \geq x_{j_1 j_0}([\rho_{j_1}, T - \tau_{j_1 j_0}]) > 0.$$

Thus there is some vertex $j_2 \neq j_1$ with $x_{j_2 j_1}([\rho_{j_2}, T - \tau_{j_1 j_0} - \tau_{j_2 j_1}]) > 0$, implying that $T - \tau_{j_1 j_0} - \tau_{j_2 j_1} \geq \rho_{j_2}$.

Since there is a minimum nonzero traversal time, we may repeat this process until either some member j_u, say, of the sequence $j_0, j_1, \ldots$ equals 1, in which case the result follows because $P = \{j_u, j_{u-1}, \ldots, j_0\}$ is a path from 1 to n with $\sum_{(j,k) \in P} \tau_{jk} \leq T$, or on the other hand some subsequence $C = \{j_q, j_{q+1}, \ldots, j_{q+v} = j_q\}$ occurs with $\tau_{jk} = 0$

for j and k consecutive members of C, $j_{q-1} \notin C$, and $x_{j_q j_{q-1}}([\rho_q, T - \sum_{p=1}^{p=q} \tau_{j_p j_{p-1}}]) > 0$. In this case if we let $S = [\rho_q, T - \sum_{p=1}^{p=q} \tau_{j_p j_{p-1}}]$ then we have

$$\sum_{j \in C} \sum_{k \notin C} x_{jk}(S) > 0.$$

By virtue of the positivity of x it follows using the convention of (4) that

$$\sum_{j \in C} \sum_{k \in C} x_{jk}(S) \geq \sum_{j \in C} \sum_{k \in C} x_{kj}(S - \tau_{kj}),$$

whence

$$\sum_{j \in C} \sum_{k} x_{jk}(S) > \sum_{j \in C} \sum_{k \in C} x_{kj}(S - \tau_{kj}),$$

from which we obtain

$$\sum_{j \in C} \sum_{k \notin C} x_{kj}(S - \tau_{kj}) > \sum_{j \in C} \sum_{k} x_{kj}(S - \tau_{kj}) - x_{jk}(S).$$

The right-hand side of this inequality is nonnegative by virtue of (3), and so there must be some vertex $l \notin C$, and some vertex $j \in C$ with $x_{lj}(S - \tau_{lj}) > 0$. If $l \neq 1$ and $\tau_{lj} = 0$ then we add l to C and repeat the above argument until either $l = 1$ or $\tau_{lj} > 0$. Since there is a minimum nonzero traversal time, and a finite number of vertices, this procedure must terminate at vertex 1 at some time $t \in [0, \sigma_1]$. From the ordered set of vertices visited, it is straightforward to construct a path P from 1 to n with $\sum_{(j,k) \in P} \tau_{jk} \leq T$. □

We assume henceforth that T satisfies the conditions of Lemma 2.1. In what follows we require the following integration by parts formula, which is easily derived from a standard result.

LEMMA 2.2. *Let α and β be two functions of bounded variation on $[\rho, \sigma]$ with α continuous on $[\rho, \sigma]$, and β continuous from the right on (ρ, σ). Then*

$$\int_{\rho}^{\sigma} \alpha(t) d\beta(t) + \int_{\rho}^{\sigma} \beta(t) d\alpha(t) = \alpha(\sigma)\beta(\sigma) - \alpha(\rho)\beta(\rho)$$

Proof. For the proof see [5, p. 154]. □

LEMMA 2.3. *Suppose for each j that w_j is a Lebesgue integrable function on $[\rho_j, \sigma_j]$, and let $W_j(t) = \int_t^{\sigma_j} w_j(\tau) d\tau$. Then x is feasible for* CSP *implies that*

$$\sum_j \int_{\rho_j}^{\sigma_j} w_j(t) \sum_k [\int_{\rho_k}^{t - \tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau)] dt$$

$$= \sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} [W_k(t + \tau_{jk}) - W_j(t)] dx_{jk}(t).$$

Proof. Since each $W_j(t)$ is a continuous function of bounded variation on $[\rho_j, \sigma_j]$ we may let $\alpha = W$ and $\beta = X$ in Lemma 2.2 to give

$$\int_{\rho_j}^{\sigma_j} w_j(t) \sum_k [\int_{\rho_k}^{t - \tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau)] dt$$

$$= \left[-W_j(t) \sum_k X_{kj}(t - \tau_{kj}) - X_{jk}(t) \right]_{\rho_j}^{\sigma_j}$$

$$+ \int_{\rho_j}^{\sigma_j} W_j(t) \sum_k dx_{kj}(t - \tau_{kj}) - dx_{jk}(t)$$

$$= \int_{\rho_j}^{\sigma_j} W_j(t) \sum_k dx_{kj}(t - \tau_{kj}) - dx_{jk}(t),$$

using the fact that $W_j(\sigma_j) = 0$ and $X_{kj}(\rho_j - \tau_{kj}) = 0$ which follows from $\rho_j \leq \rho_k + \tau_{kj}$.

Now summing over j gives

$$\sum_j \int_{\rho_j}^{\sigma_j} w_j(t) \sum_k [\int_{\rho_k}^{t-\tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau)]dt$$

$$= \sum_j \int_{\rho_j}^{\sigma_j} W_j(t) \sum_k dx_{kj}(t - \tau_{kj}) - dx_{jk}(t)$$

$$= \sum_j \sum_k \int_{\rho_j - \tau_{kj}}^{\sigma_j - \tau_{kj}} W_j(t + \tau_{kj}) dx_{kj}(t) - \int_{\rho_j}^{\sigma_j} W_j(t) dx_{jk}(t)$$

$$= \sum_j \sum_k \int_{\rho_k - \tau_{jk}}^{\sigma_k - \tau_{jk}} W_k(t + \tau_{jk}) dx_{jk}(t) - \int_{\rho_j}^{\sigma_j} W_j(t) dx_{jk}(t)$$

$$= \sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} W_k(t + \tau_{jk}) dx_{jk}(t) - \int_{\rho_j}^{\sigma_k - \tau_{jk}} W_j(t) dx_{jk}(t),$$

using (4) and the inequalities $\rho_j \geq \rho_k - \tau_{jk}$, and $\sigma_j \geq \sigma_k - \tau_{jk}$. The result now follows by combining the integrals. □

If w is chosen to be s and $S_j(t) = \int_t^{\sigma_j} s_j(\tau) d\tau$ then Lemma 2.3 may be invoked to remove the dependence on y of the objective function of CSP, giving the following problem expressed entirely in terms of x.

CSP1: minimize $\sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} \bar{c}_{jk}(t) dx_{jk}(t)$

(5) subject to $\sum_k \int_{\rho_k}^{\sigma_j - \tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{\sigma_k - \tau_{jk}} dx_{jk}(\tau) = \begin{cases} -1, & j = 1, \\ 0, & j \neq 1, n, \\ 1, & j = n, \end{cases}$

(6) $$\sum_k \int_{\rho_k}^{t - \tau_{k1}} dx_{k1}(\tau) - \int_{\rho_1}^{t} dx_{1k}(\tau) \geq -1, \quad t \in [\rho_1, \sigma_1],$$

(7) $$\sum_k \int_{\rho_k}^{t - \tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau) \geq 0, \quad t \in [\rho_j, \sigma_j], \quad j = 2, \ldots, n,$$

$$x_{jk}(t) \geq 0, \quad t \in [\rho_j, \sigma_k - \tau_{jk}], \quad j, k = 1, \ldots, n,$$

where $\bar{c}_{jk}(t) = c_{jk}(t) + [S_k(t + \tau_{jk}) - S_j(t)]$ and $V(\text{CSP1}) = V(\text{CSP}) - \int_{\rho_1}^{\sigma_1} s_1(\tau) d\tau$.

The advantage of this formulation is that we can readily develop a duality theory for CSP1 based on the paired-space methodology introduced by Kretschmer [7] and adopted by Anderson and Nash [2]. Following [2] we specify two dual pairs of topological vector spaces, (X, Y) and (Z, W), each endowed with a bilinear form $\langle , \rangle$, and each having the weak topologies $\sigma(X, Y)$ and $\sigma(Z, W)$ given by the respective dual pairings. For CSP1, we choose

- X to be $\prod_{j,k} M[\rho_j, \sigma_k - \tau_{jk}]$, where $M[a, b]$ is the space of regular signed Borel measures on $[a, b]$,
- Y to be $\prod_{j,k} C[\rho_j, \sigma_k - \tau_{jk}]$, where $C[a, b]$ is the space of continuous functions on $[a, b]$,
- Z to be $R^n \times \prod_j L_1[\rho_j, \sigma_j]$,
- W to be $R^n \times \prod_j L_\infty[\rho_j, \sigma_j]$.

Observe that each of these spaces may also be endowed with their standard norm topologies, and as such are Banach spaces. The problem CSP1 is an example of the abstract linear programming problem

$$\begin{aligned} \text{IP: minimize } & \langle c, x \rangle \\ \text{subject to } & Ax - b \in Q \\ & x \in P \end{aligned}$$

presented in [2]. Here $\langle c, x \rangle$ is a linear objective functional, b is a fixed element of Z, P and Q are the positive cones in X and Z, respectively, and $A : X \to Z$ is the constraint operator. Thus for CSP1

$$\langle c, x \rangle = \sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} \bar{c}_{jk}(t) dx_{jk}(t),$$

$$P = \{x : x_{jk} \in M[\rho_j, \sigma_k - \tau_{jk}], x_{jk}(B) \geq 0 \text{ for every Borel set } B \subseteq [\rho_j, \sigma_k - \tau_{jk}]\},$$

$$Q = \{(0, z) : z \in \prod_j L_1[\rho_j, \sigma_j], z(t) \geq 0 \text{ a.e.}\},$$

$$b = (-1, 0, \ldots, 0, 1) \times (-1(t), 0(t), \ldots, 0(t)),$$

and the constraint operator A maps an n^2-tple of measures into a pair consisting of a real vector defined by the left-hand side of (5), and a Lebesgue-integrable vector function defined by the left-hand sides of (6) and (7).

It is straightforward to show that A is linear and continuous with respect to the weak topologies on X and Z. In order to specify a dual problem for CSP1 we must define dual cones

$$P^* = \{y \in Y : \langle x, y \rangle \quad \geq 0, \quad \forall x \in P\}, \quad Q^* = \{w \in W : \langle z, w \rangle \quad \geq 0, \quad \forall z \in Q\},$$

and an adjoint operator $A^* : W \to Y$, which satisfies $\langle Ax, w \rangle = \langle x, A^* w \rangle$. Now

$$\langle Ax, w \rangle = \left\langle Ax, \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \right\rangle,$$

where $\lambda \in R^n$ and $\mu \in \prod_j L_\infty[\rho_j, \sigma_j]$, whence $\langle Ax, w\rangle$ may be written as

$$\sum_j \lambda_j \sum_k \left(\int_{\rho_k}^{\sigma_j - \tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{\sigma_k - \tau_{jk}} dx_{jk}(\tau) \right)$$
$$+ \sum_j \int_{\rho_j}^{\sigma_j} \mu_j(t) \left[\sum_k \int_{\rho_k}^{t-\tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau) \right] dt.$$

Applying Lemma 2.3 yields

$$\langle Ax, w\rangle = \sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} \left[(\lambda_k - \lambda_j) + \int_{t+\tau_{jk}}^{\sigma_k} \mu_k(\tau) d\tau - \int_t^{\sigma_j} \mu_j(\tau) d\tau \right] dx_{jk}(t),$$

whereby the jkth element of $A^* \begin{pmatrix} \lambda \\ \mu \end{pmatrix}$ is given by

$$A^* \begin{pmatrix} \lambda \\ \mu \end{pmatrix}_{jk} = (\lambda_k - \lambda_j) + \int_{t+\tau_{jk}}^{\sigma_k} \mu_k(\tau) d\tau - \int_t^{\sigma_j} \mu_j(\tau) d\tau.$$

From the dual (see [2]) to IP, which is

$$\begin{aligned} \text{IP}^*\text{: maximize } & \langle b, w\rangle \\ \text{subject to } & c - A^* w \in P^* \\ & w \in Q^*, \end{aligned}$$

we may write down the following dual problem for CSP1:

$$\begin{aligned} \text{CSP1}^*\text{: maximize } & \lambda_n - \lambda_1 - \int_{\rho_1}^{\sigma_1} \mu_1(t) dt \\ \text{subject to } & (\lambda_k - \lambda_j) + \int_{t+\tau_{jk}}^{\sigma_k} \mu_k(\tau) d\tau - \int_t^{\sigma_j} \mu_j(\tau) d\tau \le \bar{c}_{jk}(t), \\ & \qquad t \in [\rho_j, \sigma_k - \tau_{jk}], \quad j, k = 1, \ldots, n, \\ & \mu_j(t) \ge 0, \quad t \in [\rho_j, \sigma_j], \quad j = 1, \ldots, n. \end{aligned}$$

Any λ and μ that satisfies the constraints for CSP1* is said to be (dual) *feasible*, and the *value* of CSP1* is the supremum of its objective function over all feasible solutions. We observe in passing that the change of variables $\pi_j(t) = \lambda_j + \int_t^{\sigma_j} (\mu_j(\tau) - s_j(\tau)) d\tau$ gives the following dual problem for CSP:

$$\begin{aligned} \text{CSP}^*\text{: maximize } & \pi_n(T) - \pi_1(0) \\ \text{subject to } & \frac{d\pi_j}{dt} \le s_j(t), \quad t \in [\rho_j, \sigma_j], \quad j = 1, \ldots, n, \\ & \pi_k(t + \tau_{jk}) - \pi_j(t) \le c_{jk}(t), \quad t \in [\rho_j, \sigma_k - \tau_{jk}], \quad j, k = 1, \ldots, n. \end{aligned}$$

The following weak duality result for CSP1 and CSP1* derives directly from the definition of the adjoint operator.

THEOREM 2.4. *For any feasible solution x to* CSP1 *and any feasible solution (λ, μ) to* CSP1*

$$\sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} \bar{c}_{jk}(t) dx_{jk}(t) \ge \lambda_n - \lambda_1 - \int_{\rho_1}^{\sigma_1} \mu_1(t) dt.$$

We now turn our attention to the values of CSP1 and its dual. If these differ then we say that there is a *duality gap*. We may show under the assumption stated below that CSP1 and CSP1* have no duality gap by appealing to the following theorem.

THEOREM 2.5. *Let X, Y, W and Z be normed spaces with X the normed dual of Y and W the normed dual of Z. For any $z \in Z$ let $F(z) = \{x \mid Ax - z \in Q\}$. Suppose that P and Q are closed in the appropriate weak topologies. Then if* IP *has a finite value and there exist constants K and L such that for every $z \in Z$, $x \in F(z) \Rightarrow \| x \| \leq K \| z \| + L$, then* IP *and* IP* *have no duality gap.*

Proof. See the proof of Theorem 3.15 in [2, p. 56]. □

If we are to invoke Theorem 2.5 then we must show that CSP1 has a finite value. CSP1 might be unbounded if in the underlying graph for CSP1 there exists a cycle containing edges with zero traversal times and negative total length. This situation can be avoided by making all edge distances positive, or by imposing a rank ordering on the vertices of the graph which will preclude such a cycle occurring. We adopt instead a third device, and make the following assumption.

ASSUMPTION 1. *The traversal times $\tau_{jk} > 0$ for every $j \neq k$.*

This assumption implies the existence of strictly positive $\hat{\tau} \leq \tau_{jk}$, for every j and k. The existence of this bound allows us to prove the following lemma.

LEMMA 2.6. *Suppose x is feasible for the problem* CSP1(b) *having constraints* (5) *with right-hand sides $b_j \in R$, $j = 1, 2, \ldots, n$, and constraints* (6) *and* (7) *with right-hand sides $f_j \in L_1[\rho_j, \sigma_j]$, $j = 1, 2, \ldots, n$. Then $\| x \| \leq (1/\hat{\tau}) \| f \| + (T/\hat{\tau}) \| b \|$.*

Proof. Suppose that x is a feasible solution for CSP1(b). For each $t \in [0, T]$ let $J(t) = \{j : \sigma_j < t\}$, and for notational convenience extend f_j to $[0, T]$ by setting $f_j = 0$, $t \notin [\rho_j, \sigma_j]$. For any j and t, if $t \leq \sigma_j$ then

$$f_j(t) \leq \sum_k \int_{\rho_k}^{t-\tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau), \tag{8}$$

where the right-hand side becomes zero by (4) if $t < \rho_j$. On the other hand if $t > \sigma_j$ then the right-hand side of (8) becomes b_j. It follows that for every $t \in [0, T]$

$$\sum_j f_j(t) + \sum_{j \in J(t)} b_j \leq \sum_j \sum_k \int_{\rho_k}^{t-\tau_{kj}} dx_{kj}(\tau) - \int_{\rho_j}^{t} dx_{jk}(\tau) = -\sum_j \sum_k \int_{t-\tau_{jk}}^{t} dx_{jk}(\tau).$$

It follows that

$$0 \leq \sum_j \sum_k \int_{t-\tau_{jk}}^{t} dx_{jk}(\tau) \leq -\sum_j f_j(t) - \sum_{j \in J(t)} b_j$$

whence

$$\sum_j \sum_k \int_{t-\tau_{jk}}^{t} dx_{jk}(\tau) \leq \sum_j | f_j(t) | + \sum_j | b_j | .$$

It is easy to show using integration by parts and (4) that

$$\int_0^T \int_{t-\tau_{jk}}^{t} dx_{jk}(\tau) dt = \tau_{jk} \int_0^{T-\tau_{jk}} dx_{jk}(\tau),$$

yielding

$$\sum_j \sum_k \tau_{jk} \int_0^{T-\tau_{jk}} dx_{jk}(\tau) \leq \| f \| + T \| b \|$$

whence the lower bound on τ_{jk} gives

$$\| x \| = \sum_j \sum_k \int_{\rho_j}^{\sigma_k - \tau_{jk}} dx_{jk}(\tau) \leq (1/\hat{\tau}) \| f \| + (T/\hat{\tau}) \| b \|,$$

which is the desired result. □

The following simple corollaries to Lemma 2.6 are immediate.

COROLLARY 2.7. *The feasible region of* CSP1 *is bounded in the norm of* x.

COROLLARY 2.8. CSP1 *has a finite value.*

It remains to verify that the positive cones P and Q are closed in their respective weak topologies. This can be seen by observing that they are convex sets in respective locally convex topological vector spaces. Such sets are (weakly) closed by virtue of being closed in their norm topologies. We thus have the following lemmas.

LEMMA 2.9. *The set* P *is closed in the weak topology on* X.

LEMMA 2.10. *The set* $Q = \{(0, z) : z \in \prod_j L_1[\rho_j, \sigma_j], z(t) \geq 0 \text{ a.e.}\}$ *is closed in the weak topology on* Z.

We may now invoke Theorem 2.5 to give the following result.

THEOREM 2.11. $V(\text{CSP1}) = V(\text{CSP1}^*)$.

3. Extreme points. In the previous section we established a duality relationship between CSP1 and its dual. In fact, Theorem 2.5 may be used with Theorem 3.22 in [2] to show that CSP1 is solvable. Thus there exists a primal feasible x that attains the optimal value of the primal objective function. In this section we show that x may be taken to be an extreme point of the feasible region of CSP1. The approach taken is that of [2, pp. 60–61], which invokes Alaoglu's theorem (see [5, p. 424]). This states that the unit ball in the topological dual Y of a normed space X is compact in the weak topology $\sigma(X, Y)$. Thus any norm-bounded, $\sigma(X, Y)$-closed subset F of X is $\sigma(X, Y)$-compact, and by the Krein–Milman theorem is therefore the $\sigma(X, Y)$-closed convex hull of its extreme points. Any linear $\sigma(X, Y)$-continuous functional (such as the objective function of CSP1) will attain a minimum over F at such an extreme point.

Since we have shown in Corollary 2.7 that the feasible region F of CSP1 is bounded in the norm of X, it suffices to show that F is $\sigma(X, Y)$-closed. This is done by the following lemma.

LEMMA 3.1. *The feasible region of* CSP1 *is* $\sigma(X, Y)$*-closed.*

Proof. In the notation of IP, $F = \{x : Ax \in Q + b\} \cap P$. Since P is $\sigma(X, Y)$-closed by Lemma 2.9, it suffices to show that $\{x : Ax \in Q + b\}$ is $\sigma(X, Y)$-closed. This follows immediately from Lemma 2.10 and the fact that A is $\sigma(X, Y) - \sigma(W, Z)$ continuous. □

We have now established the following theorem.

THEOREM 3.2. *There exists a solution to* CSP1 *that is an extreme point of the set defined by the constraints of* CSP1.

In the classical shortest path problem, extreme points of the feasible region correspond to shortest paths. This has an analogue in the continuous-time case. The key result we require is that the extreme points of the feasible region of CSP1 are measures with finite support. We prove this result by first establishing a lemma, the full proof of which is rather long and technical, so we choose to give only a sketch proof here. (The reader is referred to [12] for a rigorous proof.)

LEMMA 3.3. *Suppose that x is feasible for* CSP1 *and that for some j and k there are disjoint Borel sets B and C and some t with*

$$x_{jk}([t, t+\hat{\tau}) \cap B) > 0, \qquad x_{jk}([t, t+\hat{\tau}) \cap C) > 0.$$

Then x is not an extreme point.

Sketch of proof. The proof proceeds by constructing for B and C, respectively, two distinct sequences of edges, each containing (j,k), and each starting at vertex 1 and ending at vertex n. For each member (p,q) in the first sequence we may add (or subtract) a nonzero measure z^B_{pq} to each x_{pq}, and for each member (p,q) in the second sequence we may respectively subtract (or add) a non-zero measure z^C_{pq} from each x_{pq} in such a way that $z^B - z^C \neq 0$, and $x \pm (z^B - z^C)$ is feasible for CSP1.

Each sequence of edges is constructed using the same labelling argument, which essentially states that if $x_{jk}(E) > 0$ for some set E then either $x_{kl}(E) > 0$ for some l, or there is parking in vertex k over some interval which intersects E. Proceeding in this fashion we can label new time intervals and new vertices until after a finite number of steps (by virtue of Assumption 1) both vertex 1 and vertex n are labelled.

Since the argument for z^C is identical, we restrict our attention to the construction of z^B from the sequence containing B. Let this consist of intervals I_{j_r}, Borel sets B_{j_r} and edges (j_{r-1}, j_r), giving

$$\{I_{j_0}, B_{j_0}, (j_0, j_1), I_{j_1}, B_{j_1}, (j_1, j_2), \ldots, I_{j_{q-1}}, B_{j_{q-1}}, (j_{q-1}, j_q), I_{j_q}\},$$

where $j_0 = 1$, $j_q = n$. We define the measure z^B by scaling down the measure x (by possibly different amounts) on the Borel sets in the sequence. This scaling gives a nonzero measure that has the property that total flow into a vertex equals total flow out, and the scale factors are chosen sufficiently small so that z^B can be subtracted from x without violating the constraints on x and y.

The nonnegativity constraints (6) and (7) require particular care. To avoid cases where a y variable equals zero at an endpoint of an interval I_{j_r} in the sequence, we can assume without loss of generality that y_{j_r} is greater than some strictly positive ϵ on such an interval. It then becomes possible to ensure by scaling z^B that any changes in y_{j_r} are less than ϵ on I_{j_r}, and zero elsewhere. The details are given in [12]. □

The previous lemma can be used to show that if x is an extreme point of CSP1 then each x_{jk} is concentrated on a finite set. We make use of the following simple result.

LEMMA 3.4. *Suppose that D is any Borel set in $[0,T]$ and ψ a regular Borel measure with $\psi(D) > 0$. Suppose further that for every B and C with $B \cup C = D$ and $B \cap C = \emptyset$, either $\psi(B) = 0$ or $\psi(C) = 0$. Then there is some $\bar{t} \in D$ with $\psi(\{\bar{t}\}) > 0$, and $\psi(D \setminus \{\bar{t}\}) = 0$.*

Proof. If $\bar{v} = \inf\{v : \psi((v,T] \cap D) = 0\}$, and $\bar{u} = \sup\{u : \psi([0,u] \cap D) = 0\}$, then since $\psi(D) > 0$, $\bar{u} \leq \bar{v}$. Suppose $\bar{u} < \bar{v}$, and define $w = (\bar{u} + \bar{v})/2$. If we then let $B = [0,w] \cap D$ and $C = (w,T] \cap D$ then either $\psi(B) = 0$ or $\psi(C) = 0$ which contradicts the definition of $\bar{u}$ and $\bar{v}$. Thus $\bar{u} = \bar{v}$, whence setting $\bar{t} = \bar{u}$ gives the result. □

COROLLARY 3.5. *If x is an extreme point of* CSP1 *then each x_{jk} is concentrated on a finite set.*

Proof. Suppose x is extreme. Since $\hat{\tau} > 0$, we may cover $[0,T]$ with a finite number of intervals of the form $[t, t+\hat{\tau})$. Then for any one of these intervals, D, say, and for each j and k, either $x_{jk}(D) = 0$, or since x is extreme by Lemma 3.3 there

exists B and C which satisfy the conditions of the previous lemma. Hence there is some $\bar{t} \in D$ with $x_{jk}(\{\bar{t}\}) > 0$, and $x_{jk}(D \setminus \{\bar{t}\}) = 0$, which gives the result. □

The importance of Corollary 3.5 is that it may be used to give the following characterization of the extreme points of the feasible region of CSP1.

THEOREM 3.6. *Every extreme point of* CSP1 *corresponds to a continuous-time path from* $(1,0)$ *to* (n,T).

Proof. Recall that the extreme points of SP correspond to paths from 1 to n. If x is an extreme point of CSP1 then each x_{jk} is concentrated on a finite set S_{jk}, say. Let $S = \bigcup_{j,k} S_{jk}$ and consider the graph having as nodes the finite number of VTPs in $\{1,2,\ldots,n\} \times S$, and as edges pairs of VTPs (j,t), (k,u) if $j \neq k$ and $u \geq t + \tau_{jk}$. Then x represents a feasible solution to SP posed in this graph, where a path is sought from $(1,0)$ to (n,T). Since x is extreme for CSP1, it must also be extreme for this version of SP, whence its nonzero components indicate a path from $(1,0)$ to (n,T). □

4. Conclusion. Theorem 3.6 implies that any extreme point of CSP1 can be characterized by a corresponding continuous-time path. It is also clear that any continuous-time path determines a feasible solution to CSP1 with finite support. Moreover this is an extreme point for CSP1, for if it were not then we could demonstrate that it was not extreme for SP formulated in the time-expanded graph

$$\{1,2,\ldots,n\} \times S,$$

where S is the set of departure times in the path. By virtue of Theorem 3.2 it follows that

1. there exists an optimal solution to CSP1 defining a shortest continuous-time path,
2. a shortest continuous-time path will define an optimal solution to CSP1.

It is tempting to consider extending the results above to models in which both the transit times and the edge distances vary, since in many situations this will be the case. As shown in [10] it is possible to construct algorithms for such problems when the paths contain a finite number of links. However, it seems impossible to incorporate varying transit times into a linear programming framework, at least without making the presentation considerably more complicated than that given above.

The assumption that all transit times are strictly positive is central to the arguments presented in this paper. This assumption, which gives a bound on the norm of x, also guarantees that the number of links traversed in any continuous-time path is finite. A similar finiteness assumption arises in [2] where the relationship between a continuous-time max flow-min cut theorem and duality theory is explored. In this context the minimum cut problem is formulated as an infinite-dimensional linear program over a space of measures, and the absence of a duality gap is demonstrated. However, in general there is no guarantee that the optimal value to the dual problem is attained, and a correspondence between solutions to the dual problem and continuous-time cuts can only be established when the cuts have a finite number of switches. (A continuous-time version of the max flow-min cut theorem without such an assumption is established in [11] using a labelling algorithm.)

It is interesting to conjecture whether a strong duality result can be derived whereby the value of CSP1* is attained and equals V(CSP1). This depends on being able to construct dual feasible μ in $\prod_j L_\infty[\rho_j, \sigma_j]$ having value V(CSP1*), which is straightforward if we assume that c satisfies a Lipschitz condition.

ASSUMPTION 2. *There exists K such that for every $j,k = 1,\ldots,n$,*

$$| c_{jk}(t_1) - c_{jk}(t_2) | < K | t_1 - t_2 |, \quad t_1, t_2 \in [\rho_j, \sigma_k - \tau_{jk}].$$

Observe that since $s \in \prod_j L_\infty[\rho_j, \sigma_j]$, this assumption implies that $\bar{c}_{jk}(t) = c_{jk}(t) + \int_{t+\tau_{jk}}^{\sigma_k} s_k(t)dt - \int_t^{\sigma_j} s_j(t)dt$ also satisfies a Lipschitz condition. This allows us to prove the following result.

THEOREM 4.1. *Suppose that c satisfies Assumption 2. Then there exists $\lambda \in R^n$, and $\mu \in \prod_j L_\infty[\rho_j, \sigma_j]$ with λ and μ feasible for* CSP1* *and*

$$V(\text{CSP1}^*) = \lambda_n - \lambda_1 - \int_{\rho_1}^{\sigma_1} \mu_1(t)dt.$$

Proof. The results of the previous section show that for every vertex j and time $t \in [\rho_j, \sigma_j]$ there exists a continuous-time shortest path from $(1,0)$ to (j,t). When $j \neq 1$ this path may be obtained by solving CSP1 with (n,T) replaced by (j,t), but because of the form of the constraints we must modify CSP1 slightly to compute a continuous-time shortest path from $(1,0)$ to $(1,t)$. Formally, we add a dummy vertex $n+1$ with $s_{n+1}(t) = \| s_1 \|_\infty$, $c_{jn+1}(t) = c_{n+1j}(t) = \infty$ for all j except $c_{1n+1}(t) = 0$, and $\tau_{1n+1} = \hat{\tau}$; a continuous-time shortest path from $(1,0)$ to $(n+1, t+\hat{\tau})$ is a solution to CSP1 with n replaced by $n+1$. This path must travel via $(1,t)$ and hence defines a shortest path from $(1,0)$ to $(1,t)$.

Now for each vertex j and $t \in [\rho_j, \sigma_j]$ let $\psi_j(t)$ be the length of a continuous-time shortest path from $(1,0)$ to (j,t), computed by solving the appropriate version of CSP1 described above. Each of these problems has edge distances $\bar{c}_{jk}(t)$ and zero parking penalties implying that ψ_j is a monotonic decreasing function of t. Furthermore, by virtue of Assumption 1, each ψ_j is a finite sum of edge distances, and thus satisfies a Lipschitz condition by Assumption 2. The functions ψ_j therefore have derivatives almost everywhere, which, being bounded except on a set of measure zero, may be taken to lie in $L_\infty[\rho_j, \sigma_j]$.

If for each j we define

$$\mu_j(t) = -\frac{d\psi_j(t)}{dt}, \qquad \lambda_j = \psi_j(\sigma_j),$$

then λ and μ are easily seen to be feasible for CSP1* since $\mu_j \geq 0$ and

$$(\lambda_k - \lambda_j) + \int_{t+\tau_{jk}}^{\sigma_k} \mu_k(\tau)d\tau - \int_t^{\sigma_j} \mu_j(\tau)d\tau = \psi_k(t+\tau_{jk}) - \psi_j(t).$$

The right-hand side of this equation cannot exceed $\bar{c}_{jk}(t)$, otherwise the length of a continuous-time shortest path from $(1,0)$ to $(k, t+\tau_{jk})$ may be improved by travelling from j to k at time t.

Moreover, λ and μ have an objective function value of

$$\lambda_n - \lambda_1 - \int_{\rho_1}^{\sigma_1} \mu_1(t)dt = \psi_n(\sigma_n) - \psi_1(\sigma_1) + \int_{\rho_1}^{\sigma_1} \frac{d\psi_1(t)}{dt}dt = \psi_n(T) - \psi_1(0),$$

which equals the length of a continuous-time shortest path from $(1,0)$ to (n,T) evaluated with edge distances $\bar{c}_{jk}(t)$ and zero parking penalties. Since this equals $V(\text{CSP1})$, the result follows immediately from Theorem 2.4. □

One might hope that the above results would lead to the development of algorithms to solve instances of continuous-time shortest path problems. Such algorithms can be derived directly using a dynamic programming approach. Orda and Rom give a conceptual labelling method in [10] for solving instances of CSP. In our notation this carries out the following steps.

Initialise: Set

$$\pi_1(t) = \int_0^t s_1(\tau)d\tau), \quad \pi_j(t) = \infty, \ j \neq 1.$$

Iterate: Replace $\pi_j(t)$ by

$$\min\{\pi_j(t), \min_{k \neq j} \min_{\tau_{kj} \leq t_0 \leq t} \left\{ \pi_k(t_0 - \tau_{kj}) + c_{kj}(t_0 - \tau_{kj}) + \int_{t_0}^t s_j(\tau)d\tau \right\}\}$$

until for every j, $\pi_j(t)$ does not change from one iteration to the next.

Upon termination this algorithm gives a solution π to CSP*, as well as a continuous-time path corresponding to a solution x which is feasible for CSP and complementary slack with π. For each j and t the value of $\pi_j(t)$ gives the length of a shortest path from $(1,0)$ to (j,t). The minimum length path (and corresponding solution x) is easy to recover from $\pi_j(t)$ by tracing back through the sequence of VTPs yielding the minimum values. The above algorithm can be shown to terminate in a finite time when the edge distances are piecewise linear functions and the parking penalties are piecewise constant functions. Details of these results are given in [14].

Acknowledgments. We are grateful to the referees of this paper for their helpful comments which have considerably improved the exposition.

REFERENCES

[1] R. AHUJA, K. MEHLHORN, J. ORLIN, AND R. TARJAN, *Faster algorithms for the shortest path problem*, J. Assoc. Comput. Mach., 37 (1990), pp. 213–223.

[2] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces Theory and Applications*, John Wiley, New York, 1987.

[3] K. COOKE AND E. HALSEY, *The shortest route through a network with time dependent internodal transit times*, J. Math. Anal. Appl., 14 (1966), pp. 493–498.

[4] S. DREYFUS, *An appraisal of some shortest path algorithms*, Oper. Res., 17 (1969), pp. 395–412.

[5] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part I: General Theory*, Wiley–Interscience, New York, 1958.

[6] J. HALPERN, *The shortest route with time dependent length of edges and limited delay possibilities in nodes*, Z. Oper. Res., 21 (1977), pp. 117–124.

[7] K. KRETSCHMER, *Programmes in paired spaces*, Canad. J. Math., 13 (1961), pp. 221–238.

[8] A. I. MEES, *Railway scheduling by network optimization*, Math. Comput. Modelling, 15 (1991), pp. 32–42.

[9] A. ORDA AND R. ROM, *Shortest-path algorithms for time-dependent networks*, in IEEE Infocom 88, IEEE Computer and Communications Soc., 1988, pp. 282–287.

[10] ———, *Minimum-weight paths in time-dependent networks*, Networks, 21 (1991), pp. 295–319.

[11] A. B. PHILPOTT, *Continuous-time flows in networks*, Math. Oper. Res., 15 (1990), pp. 640–661.

[12] ———, *Continuous-time shortest path problems and linear programming*, Tech. Report 509, School of Engineering, University of Auckland, 1992.

[13] A. B. PHILPOTT AND A. I. MEES, *Continuous-time shortest path problems with stopping and starting costs*, Appl. Math. Lett., 5 (1992), pp. 63–66.

[14] ———, *A finite-time algorithm for shortest path problems with time-varying costs*, Appl. Math. Lett., 6 (1993), pp. 91–94.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 553-571, March 1994

013

SENSITIVITY ANALYSIS OF PARAMETRIZED PROGRAMS VIA GENERALIZED EQUATIONS*

ALEXANDER SHAPIRO†

Abstract. This paper investigates local behavior of optimal solutions of parametrized optimization problems with cone constraints in Banach spaces. The corresponding first-order optimality conditions are formulated in a form of generalized equations (variational inequalities) and solutions of these generalized equations are studied. It is shown that under certain second-order sufficient optimality conditions and a regularity assumption related to the associated Lagrange multipliers, the considered optimal solutions are Lipschitzian stable. This is compared with a similar result in Shapiro and Bonnans [*SIAM J. Control Optim.*, 30 (1992), pp. 1409–1422]. Under the additional assumption of uniqueness of the Lagrange multipliers, first-order expansions of the optimal solutions are given in terms of solutions of auxiliary optimization problems. Finally, as an example, semi-infinite programming problems are discussed.

Key words. nonlinear optimization, parametric programming, sensitivity analysis, generalized equations, Lipschitzian stability, directional differentiability, semi-infinite programming

AMS subject classifications. 49K40, 90C31

1. Introduction.

In this paper we study optimization programs of the form

$$(P_t) \qquad \min_{x \in X} f(x,t) \quad \text{subject to } x \in \Phi(t),$$

depending on the parameter $t \in \Re_+$. Here

$$\Phi(t) = \{x \in S : g(x,t) \in K\},$$

S is a closed convex subset of a Banach space X, K is a closed convex cone in a Banach space Y and $g : X \times \Re_+ \to Y$. We investigate continuity and differentiability properties of an optimal solution of the program (P_t) , considered as a function of t, by writing the corresponding first-order optimality conditions in a form of *generalized equations.*

The finite-dimentional case, when the feasible set $\Phi(t)$ is defined by a finite number of constraints, has been studied extensively (cf. Fiacco [9]). In recent years its theory was developed and the local behavior of the optimal solutions in that case is now well understood [4], [5], [10], [12], [29], [31]. We refer to Bonnans, Ioffe, and Shapiro [6] for a discussion of these results. After the pioneering work of Robinson [25]–[27], sensitivity analysis of generalized equations has been discussed by several authors (e.g., [7], [14], [16], [21], [28]). In these papers the considered space is finite-dimensional and the corresponding cone (convex set) is polyhedral. The case of infinite-dimensional spaces and general cone constraints is discussed in recent publications [1], [8], [18], [32], [34]. In this paper we further develop and complement this theory. In the next section we show that under certain second-order sufficient conditions and a regularity assumption related to the corresponding Lagrange multipliers, the considered optimal solutions are Lipschitzian stable. In §3 we discuss a relation of this result to a similar result in Shapiro and Bonnans [34]. In §4 we derive, under the assumption of uniqueness of the Lagrange multipliers, various expansions of the optimal solutions of (P_t) in terms

* Received by the editors May 6, 1992; accepted for publication (in revised form) November 24, 1992.

† School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205.

of optimal solutions of auxiliary optimization problems. Finally, as an example, we briefly discuss in §5 semi-infinite programming.

We assume subsequently that $f(x,t)$ and $g(x,t)$ are *twice continuously differentiable* jointly in x and t. For $x \in X$ and $\xi \in X^*$ we use the notation $\langle \xi, x \rangle$ or $\langle x, \xi \rangle$ for the value $\xi(x)$ of the linear functional ξ at x. By $\overline{S}$ we denote the topological closure of a set S. The scalar product of two vectors x and y in the finite-dimensional space $\Re^n$ is denoted by $x \cdot y$. For a convex subset C of the dual space X^* and $\xi \in C$ we denote by $N_C(\xi)$ or $N(\xi, C)$ the normal cone to C at ξ,

$$N_C(\xi) = \{x \in X : \langle x, \zeta - \xi \rangle \leq 0 \text{ for all } \zeta \in C\}.$$

For a convex subset of X the normal cone is defined similarly. By $T_S(x)$ or $T(x,S)$ we denote the tangent cone to the convex set S at a point $x \in S$. Note that $T_S(x)$ is polar (negative dual) of the normal cone $N_S(x)$ and that if S is a convex cone, then $T_S(x) = \overline{S + [x]}$, where $[x]$ is the linear space generated by x. The linear space generated by a convex set S will be denoted by $Sp(S)$, that is

$$Sp(S) = \{z = t(x-y) : x, y \in S,\ t \in \Re\}.$$

In particular, if S is a convex cone, then $Sp(S) = S + (-S)$.

Consider the Lagrangian function

$$L(x, \lambda, t) = f(x,t) + \langle \lambda, g(x,t) \rangle,$$

$\lambda \in Y^*$, associated with the program (P_t). Then, under a constraint qualification, to an optimal solution $\bar{x}(t)$ of (P_t) corresponds a vector $\bar{\lambda}(t)$ of Lagrange multipliers satisfying the first-order necessary conditions (cf. [19], [22])

$$\begin{aligned} &0 \in D_x L(x, \lambda, t) + N_S(x), \\ &\lambda \in K^-,\ \langle \lambda, g(x,t) \rangle = 0. \end{aligned} \tag{1.1}$$

Here

$$K^- = \{\lambda \in Y^* : \langle \lambda, v \rangle \leq 0 \text{ for all } v \in K\}$$

is the polar (negative dual) cone of K. The optimality conditions (1.1), together with the feasibility conditions $g(x,t) \in K$, $x \in S$, can be written in the form of generalized equations (variational inequality) as follows (cf. [26]):

$$0 \in F(z,t) + \Omega(z), \tag{P_t^*}$$

where $z = (x, \lambda) \in X \times Y^*$,

$$F(z,t) = (D_x L(x, \lambda, t), -g(x,t)) : X \times Y^* \times \Re_+ \to X^* \times Y$$

and

$$\Omega(z) = N_{S \times K^-}(x, \lambda) = N_S(x) \times N_{K^-}(\lambda) \subset X^* \times Y.$$

In the subsequent analysis we study solutions of the generalized equations (P_t^*) rather than solutions of the optimization program (P_t). We say that $\bar{x} = \bar{x}(t) \in S$ is a *stationary solution* of the program (P_t) if $g(\bar{x}, t) \in K$ and there exists a vector $\bar{\lambda} = \bar{\lambda}(t)$ satisfying the first-order optimality conditions (1.1). The corresponding point $\bar{z} = \bar{z}(t) = (\bar{x}(t), \bar{\lambda}(t))$ is then a solution of the generalized equations (P_t^*).

Of course, generalized equations (P_t^*) have a particular structure that we exploit here. In this respect let us make the following observations.

(i) The multifunction $\Omega(z)$ is monotone, i.e., for any z_1, z_2 and $\zeta_1 \in \Omega(z_1), \zeta_2 \in \Omega(z_2)$,

$$\langle \zeta_1 - \zeta_2, z_1 - z_2 \rangle \geq 0.$$

Note that $\Omega(z) = \emptyset$ if $z \notin S \times K^-$.

(ii) Consider the mapping $F(z) = F(z, 0)$. We have that

$$DF(z) = \begin{bmatrix} D^2_{xx} L(x, \lambda, 0) & D_x g(x, 0) \\ -D_x g(x, 0)^* & 0 \end{bmatrix}$$

and hence for $z_1 = (x_1, \lambda_1)$ and $z_2 = (x_2, \lambda_2)$,

$$\begin{aligned} &\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \\ &= \int_0^1 \langle z_1 - z_2, DF(z_2 + \tau(z_1 - z_2))(z_1 - z_2) \rangle d\tau \\ &= \int_0^1 \langle x_1 - x_2, D^2_{xx} L(x_2 + \tau(x_1 - x_2), \lambda_2 + \tau(\lambda_1 - \lambda_2), 0)(x_1 - x_2) \rangle d\tau. \end{aligned} \tag{1.2}$$

Equation (1.2) indicates a relation between monotonicity properties of the mapping $F(z)$ and second-order conditions for the program (P_t). That is, if for all z in a convex region W and some $\beta > 0$ and $y \in X$,

$$\langle y, D^2_{xx} L(z, 0) y \rangle \geq \beta \|y\|^2, \tag{1.3}$$

then for all $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2) \in W$ such that $x_1 - x_2 = y$ we have

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \beta \|x_1 - x_2\|^2. \tag{1.4}$$

Note that the lower bound given in the right-hand side of (1.4) involves only components x_1 and x_2 of the respective vectors z_1 and z_2 and does not depend on λ_1 and λ_2. We will discuss this later.

2. Lipschitz stability of the solutions. In this section we study continuity properties of solutions $\bar{z}(t) = (\bar{x}(t), \bar{\lambda}(t))$ of the generalized equations (P^*_t). Before proceeding further we introduce the regularity conditions that will be responsible for a relation between continuity properties of $\bar{x}(t)$ and $\bar{\lambda}(t)$.

Let x_0 be an optimal solution of the unperturbed program (P_0). We assume that x_0 is a *regular* point of $g(x) = g(x, 0)$, with respect to S and K, in the sense of Robinson [23]. That is

$$0 \in \operatorname{int}\{g(x_0) + Dg(x_0)(S - x_0) - K\}.$$

It follows then that the set Λ_0 of Lagrange multipliers satisfying the first-order optimality conditions (1.1), with $x = x_0$ and $t = 0$, is *nonempty* and *bounded*. These optimality conditions can be written in the form

$$-\lambda \circ \bar{G} - \bar{a} \in N_S(x_0), \ \langle \lambda, \bar{c} \rangle = 0, \ \lambda \in K^-, \tag{2.1}$$

where $\bar{G} = D_x g(x_0, 0)$, $\bar{a} = D_x f(x_0, 0)$ and $\bar{c} = g(x_0, 0)$.

Consider

$$\Lambda(a, \gamma, x) = \{\lambda \in K^- : \ -\lambda \circ \bar{G} - a \in N_S(x), \ \langle \lambda, \bar{c} \rangle = \gamma \}. \tag{2.2}$$

It is clear that $\Lambda_0 = \Lambda(\bar{a}, 0, x_0)$. By $\mathrm{dom}(\Lambda)$ we denote the domain of the multifunction $\Lambda(a, \gamma, x)$, i.e., $\mathrm{dom}(\Lambda) = \{(a, \gamma, x) : \Lambda(a, \gamma, x) \neq \emptyset\}$. We assume that the multifunction $\Lambda(a, \gamma, x)$ is *upper Lipschitzian* at $(\bar{a}, 0, x_0)$ in the following sense.

ASSUMPTION A1. *There exists a positive constant k such that*

$$\sup_{\lambda \in \Lambda(a,\gamma,x)} \mathrm{dist}(\lambda, \Lambda_0) \leq k(\|a - \bar{a}\| + |\gamma| + \|x - x_0\|) \tag{2.3}$$

for all $(a, \gamma, x) \in \mathrm{dom}(\Lambda)$ in a neighborhood of $(\bar{a}, 0, x_0)$.

Since Λ_0 is bounded it follows from (2.3) that $\Lambda(a, \gamma, x)$ is uniformly bounded for all (a, γ, x) in a neighborhood of $(\bar{a}, 0, x_0)$. It also follows from (2.3) that if λ satisfies conditions (2.1) but with $\bar{a}, \bar{G}, \bar{c}$, and x_0 replaced by a, G, c, and x, respectively, then

$$\mathrm{dist}(\lambda, \Lambda_0) \leq k(\|a - \bar{a}\| + \|\lambda\|\|G - \bar{G}\| + \|\lambda\|\|c - \bar{c}\| + \|x - x_0\|).$$

Consequently, we obtain that if $\bar{\lambda}(t)$ satisfies the necessary conditions (1.1), then

$$\begin{aligned}\mathrm{dist}(\bar{\lambda}(t), \Lambda_0) = O(&\|D_x f(\bar{x}(t), t) - D_x f(x_0, 0)\| + \|D_x g(\bar{x}(t), t) - D_x g(x_0, 0)\| \\ &+ \|g(\bar{x}(t), t) - g(x_0, 0)\| + \|\bar{x}(t) - x_0\|).\end{aligned}$$

If we assume further that $\bar{x}(t) \to x_0$ as $t \to 0$, we obtain

$$\mathrm{dist}(\bar{\lambda}(t), \Lambda_0) = O(\|\bar{x}(t) - x_0\| + t). \tag{2.4}$$

This shows that Assumption A1 guarantees that perturbations of the Lagrange multipliers are of the same order as perturbations of the corresponding optimal solutions. We will discuss Assumption A1 further in the next section.

We now introduce second-order optimality conditions for the progaram (P_t) (cf. [19]). For $\eta > 0$ consider the cone

$$C_\eta = \{y \in T_S(x_0) : \ Dg(x_0)y \in T_K(g(x_0)), \ \langle Df(x_0), y\rangle \leq \eta\|y\| \ \}. \tag{2.5}$$

ASSUMPTION A2 (General second-order sufficient conditions). *There exist $\alpha > 0$ and $\eta > 0$ such that*

$$\langle y, D^2_{xx} L(x_0, \lambda, 0)y\rangle \geq \alpha\|y\|^2, \tag{2.6}$$

for all $y \in C_\eta$ and all $\lambda \in \Lambda_0$.

Note that since $D^2_{xx} L(x, \lambda, 0)$ is linear in λ, it follows from (2.6) that for a positive constant β less than α,

$$\langle y, D^2_{xx} L(x, \lambda, 0)y\rangle \geq \beta\|y\|^2$$

for all (x, λ) in a neighborhood W of $\{x_0\} \times \Lambda_0$ and all $y \in C_\eta$. Consequently condition (1.4) holds for all $z_1, z_2 \in W$ and such that $x_1 - x_2 = y$.

THEOREM 2.1. *Suppose that the point x_0 is regular, that Assumptions* A1 *and* A2 *hold and let $\bar{x}(t)$ be a stationary solution of (P_t) such that $\bar{x}(t)$ converges to x_0 as $t \to 0^+$. Then there exists a positive constant c such that*

$$\|\bar{x}(t) - x_0\| \leq ct \tag{2.7}$$

for all $t \geq 0$ sufficiently small.

Proof. Let $\bar{z}(t) = (\bar{x}(t), \bar{\lambda}(t))$ be a solution of the generalized equations (P_t^*) corresponding to the stationary solution $\bar{x}(t)$. We argue by a contradiction. Suppose that (2.7) is false. Then there are $t_n \to 0^+$, $x_n = \bar{x}(t_n)$ and $\kappa_n = \|x_n - x_0\|$ such that

$$\lim_{n\to\infty} \kappa_n^{-1} t_n = 0. \tag{2.8}$$

Consider $y_n = \kappa_n^{-1}(x_n - x_0)$, $\lambda_n = \bar{\lambda}(t_n)$ and $z_n = (x_n, \lambda_n)$. Note that it follows from the regularity of the point x_0 that Λ_0 and $\{\lambda_n\}$ are bounded.

Since $x_n \in S$ we have that $y_n \in T_S(x_0)$. Furthermore, because of (2.8),

$$g(x_n, t_n) = g(x_0) + \kappa_n Dg(x_0) y_n + o(\kappa_n).$$

Since $g(x_n, t_n) \in K$ it follows that

$$\operatorname{dist}(Dg(x_0)y_n,\ K + [g(x_0)]) = \operatorname{dist}(Dg(x_0)y_n,\ T_K(g(x_0)) \to 0.$$

By the Robinson [24]–Ursescu [35] stability theorem this implies that $\operatorname{dist}(y_n, \Sigma) \to 0$, where

$$\Sigma = \{y \in T_S(x_0) :\ Dg(x_0)y \in T_K(g(x_0))\}. \tag{2.9}$$

Therefore there exists $\bar{y}_n \in \Sigma$ such that $\bar{y}_n - y_n \to 0$.

Moreover, since

$$-D_x L(x_n, \lambda_n, t_n) \in N_S(x_n)$$

we have that

$$\langle D_x L(x_n, \lambda_n, t_n), x_n - x_0 \rangle \le 0$$

and hence

$$\langle D_x f(x_n, t_n), x_n - x_0 \rangle + \langle \lambda_n, D_x g(x_n, t_n)(x_n - x_0) \rangle \le 0. \tag{2.10}$$

We also have

$$\langle \lambda_n, g(x_n, t_n) - g(x_0, 0) \rangle = \langle \lambda_n, -g(x_0, 0) \rangle \ge 0$$

and

$$g(x_n, t_n) - g(x_0, 0) = D_x g(x_n, t_n)(x_n - x_0) + o(\kappa_n)$$

and hence

$$-\langle \lambda_n, D_x g(x_n, t_n)(x_n - x_0) \rangle \le o(\kappa_n). \tag{2.11}$$

Consequently we obtain from (2.10) and (2.11),

$$\langle D_x f(x_n, t_n), x_n - x_0 \rangle \le o(\kappa_n).$$

This implies that

$$\langle y_n, Df(x_0) \rangle \le \eta$$

and hence

$$\langle \bar{y}_n, Df(x_0) \rangle \leq \eta$$

for all n large enough. We obtain then that $\bar{y}_n \in C_\eta$.

Now let $\hat{z}_n = (x_0, \hat{\lambda}_n)$ be an element of $\{x_0\} \times \Lambda_0$ such that

$$\|\lambda_n - \hat{\lambda}_n\| = O(\|x_n - x_0\| + t_n). \tag{2.12}$$

Existence of such $\hat{\lambda}_n$ is ensured by Assumption A1. Since $-F(\hat{z}_n) \in \Omega(\hat{z}_n)$ and $-F(z_n, t_n) \in \Omega(z_n)$ and $\Omega(z)$ is monotone, we obtain that

$$\langle F(\hat{z}_n) - F(z_n, t_n), \hat{z}_n - z_n \rangle \leq 0. \tag{2.13}$$

By (1.2) we have

$$\langle F(\hat{z}_n) - F(z_n), \hat{z}_n - z_n \rangle = \kappa_n^2 \int_0^1 \langle y_n, D^2_{xx} L(z_n + \tau(\hat{z}_n - z_n), 0) y_n \rangle d\tau$$

and hence, since $\bar{y}_n - y_n \to 0$, it follows from Assumption A2 that for all n large enough and $0 < \beta < \alpha$,

$$\langle F(\hat{z}_n) - F(z_n), \hat{z}_n - z_n \rangle \geq \beta \|x_0 - x_n\|^2. \tag{2.14}$$

From (2.13) we obtain

$$\langle F(\hat{z}_n) - F(z_n), \hat{z}_n - z_n \rangle \leq \langle F(z_n, t_n) - F(z_n), \hat{z}_n - z_n \rangle$$

which, together with (2.14), implies

$$\beta \|x_0 - x_n\|^2 \leq \langle F(z_n, t_n) - F(z_n), \hat{z}_n - z_n \rangle.$$

It follows that

$$\beta \|x_0 - x_n\|^2 \leq \|F(z_n, t_n) - F(z_n)\| \|\hat{z}_n - z_n\| = O(t_n \|\hat{z}_n - z_n\|)$$

and hence, because of (2.12),

$$\|x_0 - x_n\|^2 = O(t_n \|x_n - x_0\| + t_n^2). \tag{2.15}$$

Finally we observe that (2.15) contradicts (2.8) and hence the proof is complete. □

Remarks. Denote by $\Lambda(t)$ the set of Lagrange multipliers corresponding to the optimal solution $\bar{x}(t)$. Assumption A1 was used in Theorem 2.1 only to ensure the existence of a Lagrange multiplier $\bar{\lambda}(t) \in \Lambda(t)$ satisfying condition (2.4). A similar assumption was used in Alt [1], [2] but in a considerably stronger form. It was required there that

$$\|\bar{\lambda}(t) - \lambda_0\| = O(\|\bar{x}(t) - x_0\| + t) \tag{2.16}$$

for a particular (independent of t) element λ_0 of the set Λ_0. It seems that condition (2.16) will be difficult to verify in situations where the set Λ_0 is not a singleton.

Compared with a similar result in Shapiro and Bonnans [34], the second-order conditions of Assumption A2 appear to be unnecessarily strong in cases where Λ_0 is

not a singleton. We will discuss in the next section the relation between Assumption A1 and a corresponding regularity condition used in [34].

We did not assume in Theorem 2.1 that the optimal (stationary) solution $\bar{x}(t)$ of (P_t) does exist for all t. It was only stated that if such a solution exists for some $t > 0$ in a neighborhood of zero, then condition (2.7) holds. We can consider ϵ-optimal solutions with $\epsilon = \epsilon(t) > 0$ representing a possible error in solution of the optimization problem (P_t). That is, $\bar{x}(t) \in \Phi(t)$ and

$$f(\bar{x}(t), t) \leq \inf_{x \in \Phi(t)} f(x, t) + \epsilon.$$

It follows then by Ekeland's variational principle [3], that there is an ϵ-optimal solution $\hat{x}(t)$ of (P_t) such that $\|\hat{x}(t) - \bar{x}(t)\|$ is less than $\epsilon^{1/2}$ and that $\hat{x}(t)$ is the minimizer of the function

$$f_\epsilon(x, t) = f(x, t) + \epsilon^{1/2}\|x - \hat{x}(t)\|.$$

This implies existence of $\hat{\lambda}(t)$ and $\Upsilon(t) = (\Upsilon_1(t), 0) \in X^* \times Y$ such that $\|\Upsilon_1(t)\| \leq \epsilon^{1/2}$ and $\hat{z}(t) = (\hat{x}(t), \hat{\lambda}(t))$ is a solution of the generalized equation

$$0 \in F(z, t) + \Upsilon(t) + \Omega(z). \tag{2.17}$$

It is possible then to show, in a way similar to the proof of Theorem 2.1, that if the assumptions of Theorem 2.1 are satisfied with $\bar{x}(t)$ being an $\epsilon(t)$-optimal solution of (P_t) and $\epsilon(t) = O(t^2)$, then the Lipschitz stability result (2.7) still holds.

Let us finally remark that under the assumptions of Theorem 2.1, x_0 is an *isolated* stationary solution (isolated locally optimal solution) of the program (P_0). This can be shown by arguments similar to those used in the proof of Theorem 2.1 and by taking $t_n = 0$.

3. Discussion of regularity assumptions. In this section we discuss the regularity Assumption A1. This assumption and its consequence (2.4) were crucial in derivation of the Lipschitzian stability result of Theorem 2.1.

There are at least two situations where Assumption A1 holds. That is, when $S = X$, the space Y is finite-dimensional and the cone K is *polyhedral.* This result is due to Walkup and Wets [36]. The other case where Assumption A1 holds is when $S = X$ and the point x_0 is regular with respect to the cone $K_0 = K(\lambda_0)$, $\lambda_0 \in \Lambda_0$, [32, Lem. 4.4], where

$$K(\lambda) = \{y \in K : \langle \lambda, y \rangle = 0\}.$$

Note that in this case the set $\Lambda_0 = \{\lambda_0\}$ is a singleton [32, Lem. 4.3]. Since this result will be important in the next section, we state it in the following proposition.

PROPOSITION 3.1. *Let $S = X$, x_0 be an optimal solution of (P_0), λ_0 be a Lagrange multiplier satisfying the first-order necessary conditions and suppose that x_0 is a regular point of $g(x)$ with respect to $K_0 = K(\lambda_0)$. Then $\Lambda_0 = \{\lambda_0\}$ is a singleton and Assumption* A1 *holds.*

In the remainder of this section we suppose that $S = X$ and compare assumption A1 with the following regularity condition used in Shapiro and Bonnans [34].

ASSUMPTION SB. *For some $\lambda_0 \in \Lambda_1$ the tangent cone $T(\lambda_0, \Lambda_0)$ is representable in the form*

$$T(\lambda_0, \Lambda_0) = \{\lambda \in T(\lambda_0, K^-) : \lambda \circ Dg(x_0) = 0, \langle \lambda, g(x_0) \rangle = 0\}. \tag{3.1}$$

Here Λ_1 *is a subset of* Λ_0 *defined by*

$$\Lambda_1 := \operatorname{argmax}\{D_t L(x_0, \lambda, 0) : \lambda \in \Lambda_0\}.$$

In the present case when $S = X$ the set $\Lambda(a, \gamma, x)$ depends only on a and γ and can be written as

$$\Lambda(a, \gamma) = \{\lambda \in K^- : \lambda \circ Dg(x_0) + a = 0, \ \langle \lambda, g(x_0) \rangle = \gamma\}.$$

PROPOSITION 3.2. *Assumption* A1 *implies Assumption* SB.

Proof. Let us first observe that $T(\lambda_0, \Lambda_0)$ is always contained in the set given in the right-hand side of (3.1). We argue now by a contradiction. Suppose that Assumption A1 holds and that (3.1) is false. This implies existence of a nonzero vector μ in the set given by the right-hand side of (3.1) and such that $\mu \notin T(\lambda_0, \Lambda_0)$. Since $\mu \in T(\lambda_0, K^-)$ we have that there exists a sequence $\{\lambda_n\} \subset K^-$ converging to λ_0 and $\tau_n \to 0^+$ such that $\tau_n^{-1}(\lambda_n - \lambda_0) \to \mu$. Consider $a_n = -\lambda_n \circ Dg(x_0)$ and $\gamma_n = \langle \lambda_n, g(x_0) \rangle$. We have that $\lambda_n \in \Lambda(a_n, \gamma_n)$. Also, since $\mu \notin T(\lambda_0, \Lambda_0)$, it follows that

$$\liminf_{n \to \infty} \tau_n^{-1} \operatorname{dist}(\lambda_0 + \tau_n \mu, \ \Lambda_0) \ > 0$$

and hence, since $\operatorname{dist}(\cdot, \Lambda_0)$ is Lipschitz continuous,

$$\liminf_{n \to \infty} \tau_n^{-1} \operatorname{dist}(\lambda_n, \ \Lambda_0) \ > 0.$$

Moreover,

$$\tau_n^{-1}(\bar{a} - a_n) = \tau_n^{-1}(\lambda_n - \lambda_0) \circ Dg(x_0) \to \mu \circ Dg(x_0) = 0$$

and hence

$$\|a_n - \bar{a}\| = o(\tau_n).$$

Similarly

$$|\gamma_n| = o(\tau_n).$$

Consequently we obtain that

$$(\|a_n - \bar{a}\| + |\gamma_n|)^{-1} \operatorname{dist}(\lambda_n, \Lambda_0) \to \infty,$$

which contradicts condition (2.3) of Assumption A1. □

This shows that in general Assumption A1 is stronger than Assumption SB (see further discussion in §5). It is not difficult to show that if the space Y is finite-dimensional and the set Λ_0 is a singleton, then Assumption SB implies Assumption A1 and hence in this case both assumptions are equivalent. Note that apart from Assumption SB it was also assumed in Shapiro and Bonnans [34] that the cone $Dg(x_0)X - K_0 + [g(x_0)]$ is closed.

4. First-order expansions of the solutions. In this section we discuss various expansions of solutions $\bar{z}(t) = (\bar{x}(t), \bar{\lambda}(t))$ of the generalized equations (P_t^*) and the corresponding optimal solution $\bar{x}(t)$ of the optimization problem (P_t). We assume throughout this section that a *unique* Lagrange multiplier λ_0, satisfying the first-order necessary conditions, corresponds to the optimal solution x_0. This assumption of uniqueness of λ_0 is quite restrictive. In this respect we would like to mention that some results derived in the case of finite-dimensional spaces and finite number of constraints suggest a close relation between the obtained formulas for directional derivatives of the optimal solutions and the *optimization* structure of the considered problems [4], [6], [31]. In particular, when Λ_0 is not a singleton, those formulas are related to duality properties of the optimization problems. These duality properties are not apparent in the equations representing the corresponding first-order optimality conditions. Therefore, in the author's opinion, the generalized equations approach is not very well suited for studying differentiability properties of the optimal solutions when Λ_0 is not a singleton.

Let us consider the generalized equations

$$(Q_t^*) \qquad 0 \in F(z) + H(t) + \Omega(z),$$

where $H(t) = F(z_0, t) - F(z_0, 0)$ and $z_0 = (x_0, \lambda_0)$.

ASSUMPTION A3. *For all positive t in a neighborhood of zero the generalized equations (Q_t^*) have a solution $z^*(t) = (x^*(t), \lambda^*(t))$ converging to z_0 as $t \to 0^+$.*

We employ in this section the following strong form of second-order sufficient conditions.

ASSUMPTION A4. *There exists a positive constant α such that*

$$\langle y, D_{xx}^2 L(x_0, \lambda_0, 0) y \rangle \geq \alpha \|y\|^2 \tag{4.1}$$

for all $y \in Z$, where

$$Z = \{y \in \overline{Sp(S)} : Dg(x_0)y \in \overline{Sp(K)}\}.$$

Note that $T_K(g(x_0)) \subset \overline{Sp(K)}$ and therefore Assumption A2 follows from Assumption A4. Note also that Assumption A4 implies existence of a quadratic form on the linear space Z which induces on Z a norm equivalent to the original norm of X restricted to Z. Endowed with this new norm, Z becomes a Hilbert space.

THEOREM 4.1. *Suppose that the point x_0 is regular, that the corresponding Lagrange multipliers set $\Lambda_0 = \{\lambda_0\}$ is a singleton, that Assumptions* A1, A3, *and* A4 *hold and let $\bar{z}(t) = (\bar{x}(t), \bar{\lambda}(t))$ be a solution of (P_t^*) converging to z_0. Then*

$$\|\bar{x}(t) - x^*(t)\| = o(t). \tag{4.2}$$

Proof. Let us first observe that by Theorem 2.1 it follows from Assumptions A1 and A4 that $\|\bar{z}(t) - z_0\|$, $\|z^*(t) - z_0\|$ and hence $\|\bar{z}(t) - z^*(t)\|$ are of order $O(t)$.

We argue now by a contradiction. Suppose that (4.2) is false. Then there are $t_n \to 0^+$, $\bar{z}_n = (\bar{x}_n, \bar{\lambda}_n) = \bar{z}(t_n)$ and $z_n^* = (x_n^*, \lambda_n^*) = z^*(t_n)$ such that

$$t_n \|\bar{x}_n - x_n^*\|^{-1} = O(t_n). \tag{4.3}$$

Since $-F(\bar{z}_n, t_n) \in \Omega(\bar{z}_n)$ and $-F(z_n^*) - H(t_n) \in \Omega(z_n^*)$ and $\Omega(z)$ is monotone, we have that

$$\langle F(\bar{z}_n, t_n) - F(z_n^*) - H(t_n),\ \bar{z}_n - z_n^* \rangle \leq 0. \tag{4.4}$$

We also have that

$$g(\bar{x}_n, t_n) - [g(x_n^*, 0) + g(x_0, t_n) - g(x_0, 0)] \in K + (-K). \tag{4.5}$$

By the mean value theorem, it follows from (4.5) that

$$Dg(x_0, 0)(\bar{x}_n - x_n^*) + o(t_n) \in Sp(K),$$

which, together with (4.3), implies

$$Dg(x_0, 0)(\bar{x}_n - x_n^*) + o(\|\bar{x}_n - x_n^*\|) \in Sp(K). \tag{4.6}$$

We also have that $\bar{x}_n, x_n^* \in S$ and hence $\bar{x}_n - x_n^* \in Sp(S)$. Moreover, since x_0 is regular, zero is a regular point of the linear mapping $Dg(x_0, 0)$ with respect to the set $\overline{Sp(S)}$ and the cone $\overline{Sp(K)}$. It follows therefore from (4.6) that there is $y_n \in Z$ such that $\|y_n - (\bar{x}_n - x_n^*)\| = o(\|\bar{x}_n - x_n^*\|)$.

By (1.2) and continuity arguments it follows then from Assumption A4 that for all n large enough and some $\beta > 0$,

$$\langle F(\bar{z}_n) - F(z_n^*),\ \bar{z}_n - z_n^* \rangle \geq \beta \|\bar{x}_n - x_n^*\|^2. \tag{4.7}$$

Inequalities (4.4) and (4.7) imply

$$\beta \|\bar{x}_n - x_n^*\|^2 \leq \langle F(\bar{z}_n) - F(\bar{z}_n, t_n) + H(t_n), \bar{z}_n - z_n^* \rangle$$

and hence

$$\beta \|\bar{x}_n - x_n^*\|^2 \leq \|F(\bar{z}_n, t_n) - F(\bar{z}_n) - H(t_n)\| \|\bar{z}_n - z_n^*\|. \tag{4.8}$$

Note that

$$\|F(\bar{z}_n, t_n) - F(\bar{z}_n) - H(t_n)\| = o(t_n). \tag{4.9}$$

Indeed, we have that

$$F(z_0, t) - F(z_0, 0) = t D_t F(z_0, 0) + o(t)$$

and by continuity of $D_t F(z, t)$

$$F(\bar{z}(t), t) - F(\bar{z}(t), 0) = t D_t F(z_0, 0) + o(t)$$

and hence (4.9) follows. Now since $\|\bar{z}_n - z_n^*\| = O(t_n)$ it follows then from (4.8) that

$$\|\bar{x}_n - x_n^*\|^2 = o(t_n^2)$$

which contradicts (4.3). □

Remarks. Uniqueness of the Lagrange multiplier λ_0 was used in the above theorem only to ensure that $\|\bar{\lambda}(t) - \lambda^*(t)\|$ is of order $O(t)$. If the set Λ_0 is not a singleton we can try to replace the function $H(t)$ in (Q_t^*) by

$$H(t) = F(\hat{z}(t), t) - F(\hat{z}(t), 0)$$

with $\hat{z}(t) = (x_0, \hat{\lambda}(t))$ and $\hat{\lambda}(t) \in \Lambda_0$ is such that $\|\bar{\lambda}(t) - \hat{\lambda}(t)\|$ and $\|\lambda^*(t) - \hat{\lambda}(t)\|$ are of order $O(t)$. It is not clear, however, what regularity conditions will be required to guarantee existence of such $\hat{\lambda}(t)$.

Similar results were obtained by Malanowski [18] under stronger regularity conditions and by using different techniques. Note that Assumption A4 and regularity of x_0 imply that the point x_0 is an isolated stationary solution of the program (P_0). Since the second-order conditions of Assumption A4 are retained under small perturbations, $\bar{x}(t)$ is also locally unique for all t sufficiently small. Note, however, that the assumptions of Theorem 4.1 do not guarantee local uniqueness of $\bar{\lambda}(t)$ or $\lambda^*(t)$. Therefore, to ensure that $\|\bar{z}(t) - z^*(t)\| = o(t)$, we must impose some additional conditions.

In the generalized equations (Q_t^*) we can linearize $F(z)$ and $H(t)$ as well. This leads to the following generalized equations:

$$(\mathcal{L}_t^*) \qquad 0 \in F(z_0) + DF(z_0)(z - z_0) + tD_tF(z_0, 0) + \Omega(z).$$

Suppose that for all sufficiently small $t \geq 0$ the generalized equations $(\mathcal{L}_t^*)$ have a solution $z'(t) = (x'(t), \lambda'(t))$ converging to z_0 as $t \to 0^+$. It is possible then to show that, under the assumptions of Theorem 4.1 (except Assumption A3, which is not required here),

$$\|\bar{x}(t) - x'(t)\| = o(t). \tag{4.10}$$

It is also possible to add an error term $\Upsilon(t)$ of order $o(t)$ to the right-hand sides of the generalized equations (Q_t^*) and $(\mathcal{L}_t^*)$.

Let us return to the optimization problem (P_t). The generalized equations $(\mathcal{L}_t^*)$ can be written in the form

$$\begin{aligned} 0 \in\ & D_xL(x_0, \lambda_0, 0) + D_{xx}^2L(x_0, \lambda_0, 0)(x - x_0) \\ & +(\lambda - \lambda_0) \circ Dg(x_0) + tD_{xt}^2L(x_0, \lambda_0, 0) + N_S(x), \\ & g(x_0) + Dg(x_0)(x - x_0) + tD_tg(x_0, 0) \in N_{K^-}(\lambda). \end{aligned} \tag{4.11}$$

Note that

$$D_xL(x_0, \lambda_0, 0) - \lambda_0 \circ Dg(x_0) = Df(x_0)$$

and therefore the generalized equations $(\mathcal{L}_t^*)$ correspond to the optimization problem

$$(\mathcal{L}_t) \qquad \begin{aligned} \min_{x \in S} \quad & \langle x - x_0, Df(x_0) + tD_{xt}^2L(x_0, \lambda_0, 0)\rangle \\ & + \tfrac{1}{2}\langle x - x_0, D_{xx}^2L(x_0, \lambda_0, 0)(x - x_0)\rangle \\ \text{subject to} \quad & g(x_0) + Dg(x_0)(x - x_0) + tD_tg(x_0, 0) \in K. \end{aligned}$$

For every t the feasible set of the program $(\mathcal{L}_t)$ is convex and, because of the second-order condition (4.1) of Assumption A4, the objective function of $(\mathcal{L}_t)$ is convex on the linear space generated by this feasible set. By (4.11) it follows then that x_0 is the optimal solution of $(\mathcal{L}_0)$. Moreover, by Assumption A4 and regularity of x_0 we have here that for all t the program $(\mathcal{L}_t)$ has a unique optimal solution $x'(t)$ and $x'(t) \to x_0$ as $t \to 0^+$ (see the proof of Theorem 2 in [34]). Together with Theorem 4.1 this implies the following result.

THEOREM 4.2. *Consider the optimal solution $x'(t)$ of $(\mathcal{L}_t)$. Suppose that the point x_0 is regular, that $\Lambda_0 = \{\lambda_0\}$ is a singleton, that Assumptions* A1 *and* A4 *hold and let $\bar{x}(t)$ be an optimal solution of (P_t) converging to x_0 as $t \to 0^+$. Then*

$$\|\bar{x}(t) - x'(t)\| = o(t). \tag{4.12}$$

Remarks. Let $S = X$ and suppose that the point x_0 is regular with respect to the cone $K_0 = K(\lambda_0)$. By Proposition 3.1 we have then that λ_0 is unique and Assumption A1 holds. It follows that in this case second-order conditions of Assumption A4 imply the first-order equivalence (4.12) between optimal solutions of the programs (P_t) and $(\mathcal{L}_t)$. Note also that $\bar{x}(t)$ in Theorem 4.2 can be an $\epsilon(t)$-optimal solution of (P_t) with $\epsilon(t) = o(t^2)$.

DEFINITION. *We say that a closed convex set C is conical at a point $y \in C$ if $C - y$ locally coincides with the tangent cone $T_C(y)$, i.e., there is a neighborhood W of y such that $C \cap W = (y + T_C(y)) \cap W$.*

For example, a convex polyhedron in a finite-dimensional space is conical at every point and any closed convex cone is conical at $y = 0$.

Let us consider again generalized equations (4.11). Denote $y = x - x_0$, $\mu = \lambda - \lambda_0$ and substitute $N(\mu, K^- - \lambda_0)$ in place of the cone $N(\lambda, K^-)$. Suppose now that K^- is *conical* at λ_0. Then for all t sufficiently close to zero we can further replace $K^- - \lambda_0$ by $T(\lambda_0, K^-)$. Note that

$$T(\lambda_0, K^-) = \overline{K^- + [\lambda_0]} = K(\lambda_0)^- = K_0^-.$$

Therefore the obtained generalized equations will correspond to the optimization problem

$$(\mathcal{M}_t) \qquad \begin{array}{ll} \min_{y \in S - x_0} & \langle y, D_x L(x_0, \lambda_0, 0) + t D^2_{xt} L(x_0, \lambda_0, 0) \rangle \\ & + \frac{1}{2} \langle y, D^2_{xx} L(x_0, \lambda_0, 0) y \rangle \\ \text{subject to} & g(x_0) + Dg(x_0) y + t D_t g(x_0, 0) \in K_0. \end{array}$$

We obtain the following result (cf. [32, Thm. 4.4]).

COROLLARY 4.3. *Suppose that the assumptions of Theorem 4.2 hold, that K^- is conical at λ_0 and let $y'(t)$ be the optimal solution of $(\mathcal{M}_t)$. Then*

$$\|\bar{x}(t) - x_0 - y'(t)\| = o(t). \tag{4.13}$$

Suppose further that the set S is also conical at x_0. Then for all $z = (x, \lambda)$ sufficiently close to $z_0 = (x_0, \lambda_0)$, the normal cone $N(z, S \times K^-)$ coincides with the normal cone $N(z, z_0 + \Sigma)$, where

$$\Sigma = T_{S \times K^-}(z_0) = T_S(x_0) \times T_{K^-}(\lambda_0) = T_S(x_0) \times K_0^-.$$

Therefore for t small enough, solutions of the generalized equations $(\mathcal{L}_t^*)$ will coincide with solutions of the generalized equations

$$(\mathcal{L}'_t) \qquad 0 \in F(z_0) + DF(z_0)(z - z_0) + t D_t F(z_0, 0) + N_\Sigma(z - z_0).$$

Let us consider the generalized equations

$$(\mathcal{N}_t^*) \qquad 0 \in DF(z_0) v + t D_t F(z_0, 0) + N(v, \Sigma_0),$$

where

$$\Sigma_0 = \{v \in \Sigma : \langle F(z_0), v \rangle = 0\}.$$

For $t = 1$ we write $(\mathcal{N}^*)$ to denote the generalized equations $(\mathcal{N}_1^*)$. Note that solutions of $(\mathcal{N}_t^*)$ are linear in t. That is, if $\bar{v}$ is a solution of $(\mathcal{N}^*)$, then $t\bar{v}$ is a solution of $(\mathcal{N}_t^*)$.

LEMMA 4.4. *Consider the optimal solution $x'(t)$ of $(\mathcal{L}_t)$. Suppose that the assumptions of Theorem 4.2 hold, that the sets S and K^- are conical at the points x_0 and λ_0, respectively, and let $\bar{v} = (\bar{y}, \bar{\mu})$ be a solution of $(\mathcal{N}^*)$. Then*

$$\|x'(t) - x_0 - t\bar{y}\| = o(t). \tag{4.14}$$

Proof. Let $\lambda'(t)$ be a Lagrange multiplier corresponding to the optimal solution $x'(t)$ of $(\mathcal{L}_t)$. Then $z'(t) = (x'(t), \lambda'(t))$ is a solution of $(\mathcal{L}_t^*)$. Denote $v'(t) = z'(t) - z_0$. By Theorem 2.1 we have that

$$\|v'(t)\| = O(t), \tag{4.15}$$

and hence for all t small enough $z'(t)$ is a solution of $(\mathcal{L}_t')$. Since $z'(t)$ is solution of $(\mathcal{L}_t')$ we have that

$$\langle F(z_0) + DF(z_0)v'(t) + tD_tF(z_0, 0), v'(t)\rangle = 0,$$

and since $\bar{v} \in \Sigma$

$$\langle F(z_0) + DF(z_0)v'(t) + tD_tF(z_0, 0), t\bar{v}\rangle \geq 0.$$

It follows that

$$\langle F(z_0) + DF(z_0)v'(t) + tD_tF(z_0, 0), v'(t) - t\bar{v}\rangle \leq 0. \tag{4.16}$$

Also since $\bar{v}$ is a solution of $(\mathcal{N}^*)$ we have

$$\langle DF(z_0)\bar{v} + D_tF(z_0, 0), \bar{v}\rangle = 0.$$

Moreover, by the definition of Σ_0,

$$\langle F(z_0), \bar{v}\rangle = 0$$

and hence

$$\langle F(z_0) + tDF(z_0)\bar{v} + tD_tF(z_0, 0), \bar{v}\rangle = 0. \tag{4.17}$$

Note that z_0 is a solution of $(\mathcal{L}_0')$ and that $-F(z_0) \in \Sigma^-$. We also have that

$$-DF(z_0)\bar{v} - D_tF(z_0, 0) \in \Sigma_0^- = T(-F(z_0), \Sigma^-).$$

Therefore

$$\operatorname{dist}(-F(z_0) - tDF(z_0)\bar{v} - tD_tF(z_0, 0), \Sigma^-) = o(t)$$

and hence there exists $F'(t) \in \Sigma^-$ such that

$$\| - F(z_0) - tDF(z_0)\bar{v} - tD_tF(z_0, 0) - F'(t)\| = o(t).$$

Since

$$\langle F'(t), v'(t)\rangle \leq 0$$

and because of (4.15) we obtain then

$$-\langle F(z_0) + tDF(z_0)\bar{v} + tD_tF(z_0,0), v'(t)\rangle \leq o(t^2).$$

Together with (4.17) this implies

$$-\langle F(z_0) + tDF(z_0)\bar{v} + tD_tF(z_0,0), v'(t) - t\bar{v}\rangle \leq o(t^2). \tag{4.18}$$

It follows from (4.16) and (4.18) that

$$\langle DF(z_0)(v'(t) - t\bar{v}), v'(t) - t\bar{v}\rangle \leq o(t^2). \tag{4.19}$$

On the other hand Assumption A4 implies

$$\langle DF(z_0)(v'(t) - t\bar{v}), v'(t) - t\bar{v}\rangle \geq \alpha\|x'(t) - x_0 - t\bar{y}\|^2. \tag{4.20}$$

Inequalities (4.18) and (4.20) imply (4.14) and hence the proof is complete. □

Generalized equations $(\mathcal{N}^*)$ correspond to the optimization problem

$$(\mathcal{N}) \qquad \begin{array}{ll} \min_{y\in T_0} & \langle y, D^2_{xt}L(x_0,\lambda_0,0)\rangle + \frac{1}{2}\langle y, D^2_{xx}L(x_0,\lambda_0,0)y\rangle \\ \text{subject to} & Dg(x_0)y + D_tg(x_0,0) \in T(g(x_0), K_0), \end{array}$$

where

$$T_0 = \{y \in T_S(x_0) : \ \langle D_xL(x_0,\lambda_0,0), y\rangle = 0\}.$$

Theorem 4.2 and Lemma 4.1 imply the following result.

THEOREM 4.5. *Suppose that the assumptions of Theorem* 4.2 *hold, that the sets* S *and* K^- *are conical at* x_0 *and* λ_0*, respectively, and let* $\bar{y}$ *be an optimal solution of the optimization problem* $(\mathcal{N})$*. Then* $\bar{x}(t)$ *is right side differentiable at* $t = 0$ *and* $d^+\bar{x}(0)/dt$ *is equal to* $\bar{y}$*.*

Note that under Assumption A4 the optimization problem $(\mathcal{N})$ is convex and it has a unique optimal solution provided its feasible set is nonempty.

5. Concluding remarks and semi-infinite programming example. Our basic regularity assumptions are related to the *unperturbed* optimization problems and the corresponding generalized equations when the parameter value $t = 0$. Therefore, although for $t = 0$ the generalized equations are explicitly related to the associated optimization problems, the considered generalized equations need not be connected to a particular optimization problem for other values of $t > 0$. Also we can formulate the required regularity assumptions directly in terms of generalized equations without referring to the respective optimization problems. This will make the presented theory more abstract although possibly with a wider range of applications. As another remark we mention that, for the sake of simplicity, the parameter t was considered to be a scalar varying in the parameter set $\Re_+$. It is possible to extend our results to other, normed or metric, parameter spaces as well.

In the remainder of this section we discuss, as an example, semi-infinite programming problems. This will provide us with some interesting counterexamples and will show what can go wrong. We assume now that the space X is finite-dimensional, say $X = \Re^n$, and that the feasible set $\Phi(t)$ of the program (P_t) is given in the form

$$\Phi(t) = \{x \in X : \ h(x,t,\tau) \leq 0, \ \ \tau \in T\}, \tag{5.1}$$

where $h(x,t,\tau)$ is a real-valued function and T is a compact subset of $\Re^m$ (equipped with the Euclidean norm $\|\cdot\|$). It will be assumed that for every $\tau \in T$, $h(x,t,\tau)$ is twice differentiable jointly in x and t and that $h(x,t,\tau)$ together with the first- and second-order derivativs in (x,t) are continuous on $X \times \Re_+ \times T$.

Semi-infinite programming problems can be analysed by the so-called reduction method. That is, the feasible set $\Phi(t)$ can be defined by one constraint $m(x,t) \leq 0$, where $m(x,t) = \max_{\tau \in T} h(x,t,\tau)$. In certain situations the max-function $m(x,t)$ can be represented as maximum of a finite number of smooth functions and consequently the problem can be reduced to a nonlinear programming problem with a finite number of constraints (cf. [11], [13], [15], [30]). Our approach here will be different. We propose a direct analysis of the semi-infinite programs by employing results of §§2 and 4.

Denote by $\Delta(x,t)$ the set

$$\Delta(x,t) = \{\tau \in T : h(x,t,\tau) = 0\}$$

of active at (x,t) constraints. Note that if x is a feasible point of the program (P_t) and the set $\Delta(x,t)$ is nonempty, then $\Delta(x,t)$ represents the set of maximizers of $h(x,t,\cdot)$ over T. Unless stated otherwise all gradients will be written with respect to x.

The following first-order necessary conditions for the considered semi-infinite programming problem (P_0) are well known (e.g., [20]). If x_0 is an optimal solution of (P_0), then there exist $\tau_i \in \Delta(x_0)$, $i = 1,\ldots,n$, and nonnegative multipliers $\lambda_0, \lambda_1, \ldots, \lambda_n$, not all of them zero, such that

$$\lambda_0 \nabla f(x_0) + \sum_{i=1}^{n} \lambda_i \nabla h(x_0, 0, \tau_i) = 0. \tag{5.2}$$

The feasible set $\Phi(t)$, given in (5.1), can be written in the form of cone constraints as follows. Consider the normed space $Y = C(T)$ of continuous functions $y : T \to \Re$, equipped with the sup-norm

$$\|y\|_\infty = sup\{|y(\tau)| : \tau \in T\},$$

and the cone $K \subset C(T)$ formed by nonpositive valued continuous functions $y(\tau)$. Consider also the mapping $g : X \times \Re_+ \to Y$ taking a point (x,t) into the function $y = g(x,t)$, $y(\cdot) = h(x,t,\cdot)$. Then the feasible set $\Phi(t)$ is formed by points $x \in X$ such that $g(x,t) \in K$. It follows from the above differentiability assumptions for the function $h(x,t,\tau)$ that the mapping g is twice continuously differentiable and, for example,

$$[D_x g(x,t)v](\cdot) = v \cdot \nabla h(x,t,\cdot).$$

The dual space Y^* of $Y = C(T)$ is the space of finite signed measures on $(T, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra of T, with the norm given by the total variation of the corresponding measure. The polar cone K^- of the cone K is formed by the set of nonnegative Borel measures on T.

Assume further that there exist a vector $v \in X$ such that

$$v \cdot \nabla h(x_0, 0, \tau) < 0, \quad \text{for all } \tau \in \Delta(x_0). \tag{5.3}$$

(For the unperturbed program we sometimes omit $t = 0$ and write $g(x)$, $\Delta(x)$, etc.) In case the set T is finite this is the Mangasarian–Fromovitz constraint qualification.

Under condition (5.3) the multiplier λ_0 in (5.2) is nonzero and can be taken $\lambda_0 = 1$. It is also not difficult to show that (5.3) is *equivalent* to the condition that the point x_0 is a regular point (in the sense of Robinson [23]) of the mapping $g(x)$ with respect to the cone K (see [33]).

Consider the discrete measure

$$\mu = \sum_{i=1}^{n} \lambda_i \delta(\tau_i), \tag{5.4}$$

where $\delta(\tau)$ denotes the measure of mass one at the point τ and λ_i are the multipliers in (5.2) corresponding to the points $\tau_i \in \Delta(x_0)$. This measure μ represents a Lagrange multiplier, corresponding to the optimal solution x_0, in the dual space Y^*. Let us observe now that the cone K^- is *conical* at μ. Consequently Assumption SB (discussed in §3) holds at μ. (Assumption SB is implied by the property $T(\mu, K^-) = R(\mu, K^-)$, where $R(\mu, K^-)$ denotes the radial cone to K^- at μ, see [34].) On the other hand, in general, Assumption A1 is not satisfied here. The total variation norm distance between two atomic measures $\mu_1 = \alpha_1\delta(\tau_1)$ and $\mu_2 = \alpha_2\delta(\tau_2)$ equals $|\alpha_1| + |\alpha_2|$ provided $\tau_1 \neq \tau_2$. Therefore even small changes in values of τ_i in (5.4) can result in a finite jump in the distance between the corresponding measures. This shows that the results of §§2 and 4 cannot be applied in a straightforward manner in the considered framework of the space $C(T)$. The corresponding Lipschitzian stability result of Shapiro and Bonnans [34] cannot be applied here in a straightforward way either and this is becuase typically in the present case the cone $Dg(x_0) - K_0 + [g(x_0]$ is not closed. A slightly modified approach of [34], however, leads to a description of the Lipschitzian stability of optimal solutions of semi-infinite programs (see [33] for details).

It is also interesting to note that if the set $\Delta(x_0)$ is finite and the gradients $\nabla h(x_0, 0, \tau)$, $\tau \in \Delta(x_0)$, are linearly independent, then the corresponding measure μ, given in (5.4), represents the *unique* Lagrange multiplier. However, the point x_0 here is not a regular point of $g(x)$ with respect to the cone $K_0 = K(\mu)$. This can be compared with a result in [32, Lemma 4.3]. It was shown there that regularity of x_0 with respect to K_0 is ensured by (i) uniqueness of μ, (ii) the conical property, and (iii) closedness of $Dg(x_0) - K_0 + [g(x_0]$. Again the last condition (iii) is violated here.

Let us finish this section by suggesting an alternative approach to sensitivity analysis of semi-infinite programs. The above choice of the Banach space $C(T)$ was quite arbitrary. It failed because even small perturbations in the support of the considered measures could result in large changes in the corresponding distances with respect to the considered dual norm. Let us consider now the space $Y = \mathrm{Lip}(T)$ of Lipschitz continuous functions $y(\tau)$ equipped with the norm

$$\|y\|_L = \|y\|_\infty + \sigma(y),$$

where

$$\sigma(y) = \sup\left\{\frac{|y(\tau_1) - y(\tau_2)|}{\|\tau_1 - \tau_2\|} : \tau_1, \tau_2 \in T,\ \tau_1 \neq \tau_2\right\}$$

is the Lipschitz constant of the function $y(\tau)$. To every finite signed Borel measure on T will still correspond a linear bounded functional on $\mathrm{Lip}(T)$ (although such measures will not form *all* possible bounded linear functionals on $\mathrm{Lip}(T)$). If we consider now two atomic measures $\mu_1 = \alpha_1\delta(\tau_1)$ and $\mu_2 = \alpha_2\delta(\tau_2)$, then

$$\|\mu_1 - \mu_2\|_L \leq |\alpha_1|\|\tau_1 - \tau_2\| + |\alpha_1 - \alpha_2|.$$

Suppose now that for all x and t the functions $h(x,t,\cdot)$ together with the first and second-order deravatives with respect to x and t, are Lipschitz continuous on T and that these functions are continuous in (x,t) with respect to the norm $\|\cdot\|_L$. Then the mapping $g(x,t)$, defined as above, becomes a twice continuously differentiable mapping from $X \times \Re_+$ into $\mathrm{Lip}(T)$. Suppose further that

(i) The set $\Delta(x_0) = \{\tau_1, ..., \tau_k\}$ is finite.

(ii) Condition (5.3) holds.

(iii) The multifunction $\Delta(x,t)$ is upper Lipschitzian at $(x_0,0)$, i.e., there exists a positive constant c such that for all $(x,t) \in \Phi(t)$ in a neighborhood of $(x_0,0)$ and all $\tau \in \Delta(x,t)$

$$\mathrm{dist}(\tau, \Delta(x_0)) \leq c(\|x - x_0\| + |t|). \tag{5.5}$$

(iv) For all $(x,t) \in \Phi(t)$ sufficiently close to $(x_0,0)$ and all $i = 1,\ldots,k$, the set $\Delta(x,t)$ cannot have two points in a neighborhood of the point τ_i.

Consider the set Λ_0 of vectors $\lambda = (\lambda_1,\ldots,\lambda_k)$ with nonnegative components λ_i, $i = 1,\ldots,k$, satisfying the condition

$$\nabla f(x_0) + \sum_{i=1}^{k} \lambda_i \nabla h(x_0, 0, \tau_i) = 0. \tag{5.6}$$

By the above condition (ii), the set Λ_0 is nonempty and bounded. Moreover, condition (ii) guarantees that if $\bar{x}(t)$ is an optimal solution of (P_t) sufficiently close to x_0 and t is small enough, then there exist points $\bar{\tau}_i = \bar{\tau}_i(t) \in \Delta(\bar{x}(t),t)$ and nonnegative multipliers $\bar{\lambda}_i = \bar{\lambda}_i(t)$, $i = 1,\ldots,n$, such that

$$\nabla f(\bar{x}(t),t) + \sum_{i=1}^{n} \bar{\lambda}_i \nabla h(\bar{x}(t), t, \bar{\tau}_i) = 0. \tag{5.7}$$

Also by condition (iii), for every $\bar{\tau}_i$ there is $\tau_i \in \Delta(x_0)$ such that

$$\|\bar{\tau}_i - \tau_i\| = O(\|\bar{x}(t) - x_0\| + t). \tag{5.8}$$

Actually because of the condition (iv), by adding zero multipliers if necessary, we can always assume that the number of multipliers in (5.7) is the same as in (5.6) and that condition (5.8) holds for all $i = 1,\ldots,k$. Consider vector $\bar{\lambda}(t) = (\bar{\lambda}_1(t),\ldots,\bar{\lambda}_k(t))$. It follows then from (5.8) that

$$\mathrm{dist}(\bar{\lambda}(t), \Lambda_0) = O(\|\bar{x}(t) - x_0\| + t).$$

If we consider now the set M_0 of discrete mesures $\mu = \sum_{i=1}^{k} \lambda_i \delta(\tau_i)$ corresponding to the multipliers of the set Λ_0 and the discrete measure $\bar{\mu}(t)$ corresponding to $\bar{\lambda}(t)$, we obtain that the set M_0 is bounded and

$$\mathrm{dist}(\bar{\mu}(t), M_0) = O(\|\bar{x}(t) - x_0\| + t),$$

where the distance is taken with respect to the dual norm of the norm $\|\cdot\|_L$. It follows then by the results of §2 that, under suitable second-order sufficient conditions, the optimal solutions $\bar{x}(t)$ of (P_t) are *Lipschitz stable* at $t = 0$.

Finally let us remark that if in addition to the above conditions we assume that the set $\Lambda_0 = \{\lambda_0\}$ is a singleton, then the result of theorem 4.2 will apply. Consequently we obtain that, under suitable second-order conditions, an optimal solution

$\bar{x}(t)$ of (P_t) can be first-order approximated by the optimal solution of the semi-infinite programming problem

$$(\mathcal{S}_t) \qquad \begin{array}{ll} \min_{x \in X} & (x - x_0) \cdot [\nabla f(x_0) + t \nabla^2_{xt} L(x_0, \lambda_0, 0)] \\ & + \frac{1}{2}(x - x_0) \cdot \nabla^2_{xx} L(x_0, \lambda_0, 0)(x - x_0) \\ \text{subject to} & (x - x_0) \cdot a(\tau) + b(t, \tau) \leq 0, \quad \tau \in T, \end{array}$$

where $L(x, \lambda, t) = f(x, t) + \sum_{i=1}^{k} \lambda_i h(x, t, \tau_i)$ with τ_i, $i = 1, \ldots, k$, being the points forming the set $\Delta(x_0)$, $a(\tau) = \nabla h(x_0, \tau)$ *and* $b(t, \tau) = h(x_0, \tau) + t \partial h(x_0, 0, \tau)/\partial t$.

REFERENCES

[1] W. Alt, *Stability of solutions for a class of nonlinear cone constrainted optimization problems, Part* 1: *basic theory*, Numer. Funct. Anal. Optim., 10 (1989), pp. 1053–1064.

[2] ———, *Local stability of solutions to differentiable optimization problems in Banach spaces*, J. Optim. Theory Appl., 70 (1991), pp. 443–466.

[3] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.

[4] A. Auslender and R. Cominetti, *First and second order sensitivity analysis of nonlinear programs under directional constraint qualification conditions*, Optimization, 21 (1990), pp. 351–363.

[5] J. F. Bonnans, *Directional derivatives of optimal solutions in smooth nonlinear programming*, J. Optim. Theory Appl., 73 (1992), pp. 27–45.

[6] J. F. Bonnans, A. D. Ioffe and A. Shapiro, *Expansion of exact and approximate solutions in nonlinear programming*, Proc. French-German Conference on Optimization, D. Pallaschke, ed., Lecture Notes in Economics and Mathematical Systems, Vol. 382 Springer-Verlag, Berlin, 1992, pp. 103–117.

[7] S. Dafermos, *Sensitivity analysis in variational inequalities*, Math. Oper. Res., 13 (1988), pp. 421–434.

[8] A. L. Dontchev and W. W. Hager, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.

[9] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New york, 1983.

[10] J. Gauvin and R. Janin, *Directional behaviour of optimal solutions in nonlinear mathematical programming*, Math. Oper. Res., 13 (1988), pp. 629–649.

[11] R. P. Hettich and H. Th. Jongen, *Semi-infinite programming: conditions of optimality and applications*, In Optimization Techniques, Proc. 8th IFIP Conf. on Optimization Techniques, Würzburg. Part 2, J. Stoer, ed., Springer–Verlag, New York, 1977.

[12] A. D. Ioffe, *On sensitivity analysis of nonlinear programs in Banach spaces: the approach via composite unconstrained optimization*, SIAM J. Optim., 4 (1994), pp. 000–000.

[13] H. Th. Jongen, W. Wetterling and G. Zwier, *On sufficient conditions for local optimality in semi-infinite programming*, Optimization, 18 (1987), pp. 165-178.

[14] A. J. King and R. T. Rockafellar, *Sensitivity analysis for nonsmooth generalized equations*, Math. Prog., 55 (1992), pp. 193–212.

[15] D. Klatte, *Stability of stationary solutions in semi-infinite optimization via the reduction approach*, Proc. French-German Conference on Optimization, D. Pallaschke ed., Lecture Notes in Economics and Mathematical Systems, Vol. 382, Springer-Verlag, Berlin, 1992, pp. 155–170.

[16] J. Kyparisis, *Sensitivity analysis framework for variational inequalities*, Math. Prog., 38 (1987), pp. 190–203.

[17] F. Lempio and H. Maurer, *Differential stability in infinite-dimensional nonlinear programming*, Appl. Math. Optim., 6 (1980), pp. 139–152.

[18] K. Malanowski, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.

[19] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Prog., 16 (1979), pp. 98–110.

[20] B. N. Pshenichnyi, *Necessary Conditions for an Extremum*, Marcel Dekker, New York, 1971.

[21] Y. QIU AND T. L. MAGNANTI, *Sensitivity analysis for variational inequalities defined on polyhedral sets*, Math. Oper. Res., 14 (1989), pp. 410–432.
[22] S. M. ROBINSON, *First order conditions for general nonlinear optimization*, SIAM J. Appl. Math., 30 (1976), pp. 597–607.
[23] ———, *Stability theorems for systems of inequalities, Part* II: *differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[24] ———, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[25] ———, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[26] ———, *Generalized equations and their solutions, Part* II: *applications to nonlinear programming*, Math. Prog. Stud., 19 (1982), pp. 200–221.
[27] ———, *Implicit B-differentiability in Generalized Equations*, Technical Report 2854, Mathematics Research Center, University of Wisconsin-Madison, 1985.
[28] ———, *An implicit-function theorem for a class of nonsmooth functions*, Math. Oper. Res., 16 (1991), pp. 292–309.
[29] A. SHAPIRO, *Second-order sensitivity analysis and asymptotic theory of parametrized, nonlinear programs*, Math. Prog., 33 (1985), pp. 280–299.
[30] ———, *Second-order derivatives of extremal-value functions and optimality conditions for semi-infinite programs*, Math. Oper. Res., 10 (1985), pp. 207–219.
[31] ———, *Sensitivity analysis of nonlinear programs and differentiability properties of metric projections*, SIAM J. Control Optim., 26 (1988), pp. 628–645.
[32] ———, *Perturbation analysis of optimization problems in Banach spaces*, Numer. Funct. Anal. Optim., 13 (1992), pp. 97–116.
[33] ———, *On Lipschitzian stability of optimal solutions of parametrized semi-infinite programs*, Math. Oper. Res., to appear.
[34] A. SHAPIRO AND J. F. BONNANS, *Sensitivity analysis of parametrized programs under cone constraints*, SIAM J. Control Optim., 30 (1992), pp. 1409–1422.
[35] C. URSESCU, *Multifunctions with convex closed graph*, Czechoslovak Math. J., 25 (1975), pp. 438–441.
[36] D. W. WALKUP AND R. J. B. WETS, *A Lipschitzian characterization of convex polyhedra*, Proc. Amer. Math. Soc., 23 (1969), pp. 167–173.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 2, pp. 572–590, March 1994

014

SIMULTANEOUS STABILIZATION OF THREE OR MORE PLANTS: CONDITIONS ON THE POSITIVE REAL AXIS DO NOT SUFFICE*

V. BLONDEL†, M. GEVERS†, R. MORTINI‡, AND R. RUPP‡

Abstract. The problem of the simultaneous stabilizability of a finite family of single-input, single-output time-invariant systems by a time-invariant controller is studied. The link between stabilization and avoidance is shown and is used to derive necessary conditions for the simultaneous stabilization of k plants. These necessary conditions are proved to be, in general, not sufficient. This result also disproves a long-standing conjecture on the stabilizability condition of a single plant with a stable minimum phase controller. The main result is to show that, unlike the case of two plants, the existence of a simultaneous stabilizing controller for more than two plants is not guaranteed by the existence of a controller such that the closed loops have no *real* unstable poles.

Key words. stabilization, simultaneous stabilization, strong stabilization, interpolation, avoidance

AMS subject classifications. 93D, 30C

1. Introduction.

Do you believe that simple questions always have simple answers? If you do, you may consider with interest the following problem: let

$$p_1(s) = 0, \quad p_2(s) = \frac{2}{17}\frac{(s-1)}{(s+1)}, \quad \text{and } p_3(s) = \frac{(s-1)^2}{(9s-8)(s+1)}$$

be three continuous-time rational transfer functions. Is it possible to find a single rational controller $c(s)$ that simultaneously stabilizes $p_i(s)$, $i = 1, 2, 3$ (i.e., such that the closed-loop transfer functions $p_i(s)c(s)(1+p_i(s)c(s))^{-1}$ have no poles in the complex right half plane for $i = 1, 2, 3$)? The question may not look too hard: it merely asks whether or not three plants are *simultaneously stabilizable* by a single controller. At present nobody is capable of answering such a question, and this paper is devoted to it.

Let us first state the problem clearly. We restrict our attention to single-input, single-output systems that are described by linear, time-invariant, rational but not necessarily proper transfer functions. Each one of these systems is thus represented by an arbitrary real rational function $p_i(s) \in \mathbb{R}(s)$, $i = 1, \ldots, k$. To control our systems, we allow ourselves to use a dynamic but time-invariant, rational and not necessarily proper controller $c(s) \in \mathbb{R}(s)$. Finally, our goal is to achieve continuous-time closed-loop internal stability with the controller. That is, we want that, with the chosen controller $c(s)$, the four transfer functions $p_ic(1+p_ic)^{-1}, p_i(1+p_ic)^{-1}, c(1+p_ic)^{-1}$, and $(1+p_ic)^{-1}$ have no poles in the extended right half plane. Our question is now: Under what conditions on the plants $p_i(s)$, $i = 1, \ldots, k$ is it possible to find such a simultaneous stabilizing controller? This problem has been formulated for some years now (see, e.g., [3], [5], [12], [14]–[18], [26], [27], [34]) and, despite many efforts, it has remained unsolved for $k \geq 3$. It is nowadays commonly referred to

* Received by the editors September 3, 1991; accepted for publication (in revised form) December 14, 1992. This work was supported by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State Prime Minister's Office, Science Policy Programming.

† Université Catholique de Louvain, Cesame, Place du Levant 2, B-1348 Louvain-La-Neuve, Belgium.

‡ Universität Karlsruhe, Mathematisches Institut, Englerstrasse 2, D-7500 Karlsruhe 1, Germany.

as the *simultaneous stabilization problem* and is recognized as one of the hard open problems in linear systems theory.

Although this paper does not solve the simultaneous stabilization problem, we provide some fresh angle of attack. By introducing the concept of avoidance, we produce a range of new necessary and sufficient conditions and, more importantly, we prove a negative result by showing that, unlike the case $k = 2$, the simultaneous stabilizability question of more than two plants cannot be answered by just checking whether a controller exists such that the closed-loop transfer functions have no *real* unstable poles.

We draw the reader's attention to the crucial point that we allow ourselves the use only of a *time-invariant* controller. It is not always possible to stabilize simultaneously two or more plants with such a controller. To overcome this limitation, alternative strategies have recently been developed with time-varying controllers, and we refer the reader to the existing literature for more details on this subject (see, for example, [18] and references therein). This paper deals only with the time-invariant case.

Historically, the first line of attack on the simultaneous stabilization question was given through the solution of a seemingly unrelated question: "When is it possible to stabilize a single plant with a *stable* controller?" This question, known as the *strong stabilization problem*, was fully solved by Youla, Bongiorno, and Lu [32] in a now classical paper. A plant is stabilizable by a stable controller if and only if it has an even number of real unstable poles between each pair of real unstable zeros. Such plants are said to have the *parity interlacing property*. A most remarkable feature of this condition is that it involves only the *real* unstable poles and zeros of the plant.

The link between strong stabilization and simultaneous stabilization of two plants was discovered, and used, by Saeks and Murray [23]. Roughly speaking, two plants p_1 and p_2 are simultaneously stabilizable if and only if the plant $p_1 - p_2$ is strongly stabilizable. Since a tractable condition for strong stabilization is known, this solves the problem of simultaneous stabilization of two plants. This result was further extended to a multi-input, multi-output setting by Vidyasagar and Viswanadham [26], where it was shown that from k plants p_i, $i = 1, \ldots, k$ it is possible to construct $k-1$ plants p'_i, $i = 1, \ldots, k-1$ in such a way that the plants p_i are simultaneously stabilizable if and only if the plants p'_i are simultaneously stabilizable by a *stable* controller. This equivalence, while theoretically interesting, does not provide a computable test for the simultaneous stabilization of three or more plants, since we have no criteria to decide if two or more plants are simultaneously stabilizable by a stable controller.

After these results were obtained, the main contributions to simultaneous stabilization this last decade have been in the form of necessary or sufficient conditions for simultaneous stabilization (but never necessary *and* sufficient conditions). *At present no tractable necessary and sufficient conditions exist for simultaneous stabilization except for the case of two plants.*

The simultaneous stabilization of two plants is equivalent, as stated above, to the stabilization of a single plant by a stable controller. This idea can be extended to the simultaneous stabilization of three plants. Modulo an avoidance condition, the simultaneous stabilization of three plants is equivalent to the stabilization of a single plant by a stable controller whose inverse is also stable. Such a controller is called a *unit* controller. The problem of finding a condition under which a plant p can be stabilized by a unit controller can thus be seen as an intermediate step towards the solution of the simultaneous stabilization of three plants. It is easy to see that a necessary condition is that both p and p^{-1} must have the parity interlacing

property, i.e., p must have an even number of real unstable poles between each pair of real unstable zeros and vice versa. Such plants are referred to as having the *even interlacing property*. It is shown in [29] that this even interlacing property condition is also sufficient to ensure that the plant p is stabilizable by a stable controller with no *real* unstable zeros. Note that such a controller may have complex unstable zeros, so that the result of [29] does not prove that the even interlacing property of a plant p is sufficient for stabilization by a unit controller. The even interlacing property also ensures that there exists a unit controller such that the closed-loop transfer function has no *real* unstable poles. In the same vein, [30] and [31] give a condition on three plants p_1, p_2, and p_3 under which it is possible to find a single controller such that none of the closed-loop transfer functions have *real* unstable poles.

In the first part of this paper, we pursue this line of thinking, and we give a thorough study of the question: "Given k plants p_i, $i = 1, \ldots, k$, under what condition is it possible to find a single controller such that none of the closed-loop transfer functions have *real* unstable poles?" Motivations to develop such results are threefold. First, the conditions obtained are tractable, which is seldom the case for simultaneous stabilization questions. Second, such conditions remain necessary when the closed-loop transfer functions are constrained not only to have no *real* unstable poles but no unstable poles at all. They are therefore necessary conditions for simultaneous stabilization. Third, it is known that these conditions are also sufficient for the strong stabilization of a single plant and for the simultaneous stabilization of two plants. A single plant is stabilizable by a stable controller if and only if there exists a stable controller such that the closed-loop transfer function has no *real* unstable poles. Two plants, p_1 and p_2, are simultaneously stabilizable if and only if there exists a single controller such that the closed-loop transfer functions associated to p_1 and to p_2 have no real unstable poles. By analogy, it was hoped (see, for example, the conclusion in [30]) that this property would extend to the simultaneous stabilization problem for three or more plants. As we shall see at the very end of the paper, this is unfortunately not the case.

The main contributions of this paper have been briefly described above. The layout is as follows. We introduce, in §3, the simultaneous stabilization problem as an avoidance problem in the complex plane. We show that k plants are simultaneously stabilizable if and only if there exists a controller that avoids, in a way that we will define, the k plants in the complex right half plane. This reinterpretation in terms of avoidance (i.e., nonintersection) of functions gives powerful new insights into stabilization and simultaneous stabilization problems. We use these insights in §4, where, after a quick review of some known results, we answer the question: "Given k plants p_i, $i = 1, \ldots, k$, under what condition is it possible to find a single controller such that the closed-loop transfer functions associated with each plant have no *real* unstable poles?" Strikingly, we show that under a weak assumption this can be achieved if and only if for each *pair of plants* there exist such stabilizing controllers. The fulfillment of these conditions can be checked by using the parity interlacing property so that we have a tractable test to answer the above question. Finally, in §5 we present negative results. We first show that the even interlacing property is not sufficient for stabilizability of a plant by a unit controller. It then follows that the condition given in [30] and presented in §4 is not sufficient either for simultaneous stabilization of three plants.

2. Notation. $\mathbb{R}[s]$ is the set of real polynomials. $\mathbb{R}(s)$ is the set of real rational functions. $\mathbb{C}_\infty$ is the extended complex plane, $\mathbb{C} \cup \{\infty\}$, adequately topologized, and

$\mathbb{R}_\infty$ is the extended real line, $\mathbb{R} \cup \{\infty\}$. D is the open unit disc $\{s \in \mathbb{C} :\mid s \mid < 1\}$. Ω is some chosen subset of $\mathbb{C}_\infty$. We shall assume throughout this paper that Ω is closed in the Riemann sphere topology, that it is symmetric with respect to the real axis, and that it contains at least one value of the extended real line $\mathbb{R}_\infty$ but not the whole extended real line $\mathbb{R}_\infty$. Ω is to be thought of as the complement in $\mathbb{C}_\infty$ of a region of stability. Classical examples of regions Ω are the closed unit disc $\overline{D} = \{s \in \mathbb{C} :\mid s \mid \leq 1\}$ and the extended closed right half plane $\mathbb{C}_{+\infty} = \{s \in \mathbb{C} : \Re(s) \geq 0\} \cup \{\infty\}$, which correspond, respectively, to the complement in $\mathbb{C}_\infty$ of the discrete and continuous-time stability regions. We define $I = \overline{D} \cap \mathbb{R}_\infty = [-1, 1]$ and $\mathbb{R}_{+\infty} = \mathbb{C}_{+\infty} \cap \mathbb{R}_\infty = [0, \infty]$. The subsets $\overline{D}, \mathbb{C}_{+\infty}, I$, and $\mathbb{R}_{+\infty}$ all satisfy the assumptions on Ω. A real rational function is Ω-*stable* if it has no poles in Ω. $S(\Omega)$ is the set of all Ω-stable rational functions. We use $U(\Omega)$ to denote the set of functions in $S(\Omega)$ whose inverses are in $S(\Omega)$, and we call such rational functions Ω-*units*. Finally, to shorten the notation, we define $U = U(\mathbb{C}_{+\infty})$ and $S = S(\mathbb{C}_{+\infty})$.

3. Stabilization as avoidance. The equivalence between the solvability of the simultaneous stabilization problem of two plants and interpolation conditions by real rational functions was pointed out by various authors (see [32], [25] [13], [15], [8], and [17]). By a few algebraic manipulations it is possible to show that the problem of stabilizing two plants simultaneously is equivalent to that of finding a stable rational function having a stable inverse that interpolates a set of values at a set of points in the right half plane. This interpretation of the problem has the advantage of giving a geometrical insight to the problem. In this section we develop a different view of the problem, which we call an *avoidance* approach. Roughly speaking, a controller stabilizes a set of k plants if and only if it avoids, in a sense that we shall define, the k plants in the extended right half plane. By the end of this section we hope that we will have convinced the reader that stabilization and avoidance are different names for the same mathematical question. We refer the reader to [6] or [7] for more details on avoidance concepts applied to simultaneous stabilization problems.

3.1. Internal stabilization. Throughout this paper we shall consider a controller to be within a unity feedback loop with the plant, and we shall adopt the following usual definition of stability for this closed-loop configuration.

DEFINITION 3.1. *A controller* $c(s) \in \mathbb{R}(s)$ *is an* internal stabilizer *of (or* internally stabilizes*) a plant* $p(s) \in \mathbb{R}(s)$ *if the four transfer functions* $p(s)c(s)(1 + p(s)c(s))^{-1}$, $c(s)(1 + p(s)c(s))^{-1}$, $p(s)(1 + p(s)c(s))^{-1}$, *and* $(1 + p(s)c(s))^{-1}$ *belong to* S *(i.e., they have no poles with nonnegative real part).*

These four unpractical conditions for internal stability can elegantly be condensed into a single one by using the so-called *factorization approach* described in [24] and [23]. Hereafter we give a short introduction to this approach and refer the interested reader to [24] for more details. In the sequel we will always mean *internal stability* when writing *stability.*

It is easy to check that the set S of stable rational functions is a commutative ring. The invertible elements (or units) in the ring S are the stable real rational functions whose inverses are stable, that is, the real rational functions with neither poles nor zeros in $\mathbb{C}_{+\infty}$. We have denoted this set by U. Two elements of S are called *coprime* if they have no common zeros in $\mathbb{C}_{+\infty}$. It can be proved (see [24, p. 10]) that S is an Euclidean ring and hence, if $a(s), b(s) \in S$ are coprime, then there exists $x(s), y(s) \in S$ such that $a(s)x(s) + b(s)y(s) = 1$. Such an identity is called a *Bezout identity.* Finally, the field of fractions of S is $\mathbb{R}(s)$. All this together shows that if $p(s) \in \mathbb{R}(s)$, then there exist $n_p(s), d_p(s) \in S$ and $x(s), y(s) \in S$ such that

$p(s) = n_p(s)/d_p(s)$ and $n_p(s)x(s) + d_p(s)y(s) = 1$ (such a fractional decomposition of $p(s)$ will be called a *coprime decomposition*). This is the only property of S that we will need in this paper. It provides us the following result (see [24, p. 45] for proof). For conciseness, we sometimes drop the reference to the complex variable s when writing rational functions.

THEOREM 3.2. *Let $p, c \in \mathbb{R}(s)$, and let $p = n_p/d_p$ and $c = n_c/d_c$ be any coprime decompositions of p and c. Then c stabilizes p if and only if $n_c n_p + d_c d_p \in U$.*

As a corollary of this Theorem 3.2, we may formulate the simultaneous stabilization problem as follows.

COROLLARY 3.3. *Let $p_i \in \mathbb{R}(s)$, $i = 1, \ldots, k$ and let $p_i = n_i/d_i$ be any coprime decomposition of p_i, $i = 1, \ldots, k$. Then p_i are simultaneously stabilizable if and only if there exist $n_c, d_c \in S$ such that $n_c n_i + d_c d_i \in U$, $i = 1, \ldots, k$.*

A controller $c \in \mathbb{R}(s)$ is stable if it has no poles in $\mathbb{C}_{+\infty}$, in other words, it is stable if for every coprime decomposition $c = n_c/d_c$ we have $d_c \in U$. The controller c is stable and inverse stable (we have called such functions units) if both n_c and d_c are in U. In the next section we will need the following natural definition.

DEFINITION 3.4. *Let $p \in \mathbb{R}(s)$ and let $p = n_p/d_p$ be any coprime decomposition of p in S. The plant p is* strongly stabilizable *(i.e., stabilizable by a stable controller) if and only if there exist $n_c \in S$ and $d_c \in U$ such that $n_c n_p + d_c d_p \in U$. The plant p is* unit stabilizable *(i.e., stabilizable by a stable controller whose inverse is stable) if and only if there exist $n_c \in U$ and $d_c \in U$ such that $n_c n_p + d_c d_p \in U$.*

Note that the definition above is independent of the choice of the coprime decompositions. Theorem 3.2 and Corollary 3.3 are proved for the case where the stability region is the extended closed right half plane $\mathbb{C}_{+\infty}$. It may, however, be useful to define the concept of stability in a more general framework. First, this allows us to treat continuous- and discrete-time stability questions in a general setting and, second, the use of stability regions different from the extended right half plane may be justified for practical purposes (see [25], for example). The generalization goes exactly along the same line. Let Ω be a closed subset of the extended complex plane $\mathbb{C}_\infty$ satisfying the assumptions given in §2. $S(\Omega)$ is the set of real rational functions with no poles in Ω, and $U(\Omega)$ is the set of invertible elements of $S(\Omega)$. Then the above results on the ring S carry over, namely, $S(\Omega)$ is an Euclidean commutative ring whose field of fractions is $\mathbb{R}(s)$. Since these are the only properties that are needed to prove Theorem 3.2 and Corollary 3.3, these results remain valid for a general stability region Ω. Let us state this clearly.

DEFINITION 3.5. *Let Ω be a subset of $\mathbb{C}_\infty$. A controller $c(s) \in \mathbb{R}(s)$ is an internal Ω-stabilizer of a plant $p(s) \in \mathbb{R}(s)$ if the four transfer functions $p(s)c(s)(1 + p(s)c(s))^{-1}, c(s)(1 + p(s)c(s))^{-1}, p(s)(1 + p(s)c(s))^{-1}$, and $(1 + p(s)c(s))^{-1}$ belong to $S(\Omega)$.*

We then have the following corollary.

COROLLARY 3.6. *Let Ω be a subset of $\mathbb{C}_\infty$ as described in §2. Let $p_i \in \mathbb{R}(s)$, $i = 1, \ldots, k$, and let $p_i = n_i/d_i$ be any coprime decompositions of p_i in $S(\Omega)$, $i = 1, \ldots, k$. Then p_i are simultaneously Ω-stabilizable if and only if there exist $n_c, d_c \in S(\Omega)$ such that $n_c n_i + d_c d_i \in U(\Omega)$, $i = 1, \ldots, k$.*

Proof. The proof follows exactly along the same lines as Theorem 3.2 and Corollary 3.3. See [24]. □

It is clear that if Ω' is a subset of Ω, then an Ω-stabilizing controller of a plant p is also an Ω'-stabilizing controller (indeed, if the transfer functions have no poles in Ω, they then have no poles in Ω'). In particular, if we define $\omega = \Omega \cap \mathbb{R}_\infty$, then an

Ω-stabilizing controller is also an ω-stabilizing controller. ω-stabilizability is thus a necessary condition for Ω-stabilizability. This necessary condition will play a crucial role in §4. First we show the link between stability and avoidance.

3.2. Avoidance. Functions in $\mathbb{R}(s)$ go from $\mathbb{C}_\infty$ to $\mathbb{C}_\infty$ and have the additional property that they take extended *real* values on $\mathbb{R}_\infty$. It is therefore easy to represent their behaviour on $\mathbb{R}_\infty$ with a two-dimensional graphic. On the other hand, we need four dimensions to represent their behaviour on the complex plane. With these representations in mind we may figure out where two plants $p_1(s), p_2(s) \in \mathbb{R}(s)$ possibly intersect on the $\mathbb{R}_\infty$ axis, that is, the set of points $s_0 \in \mathbb{R}_\infty$ for which $p_1(s_0) = p_2(s_0)$. It is still easy to define, but more difficult to represent geometrically, the points in $\mathbb{C}_\infty \backslash \mathbb{R}_\infty$ where two plants intersect. We give a formal definition for this.

DEFINITION 3.7. *Let $p_1(s), p_2(s) \in \mathbb{R}(s)$, let Ω be a subset of $\mathbb{C}_\infty$, and let $p_i(s) = n_i(s)/d_i(s)$ be any coprime decompositions of $p_i(s)$ in $S(\Omega)$, $i = 1, 2$. $s_0 \in \Omega$ is a point of* intersection *of multiplicity n between $p_1(s)$ and $p_2(s)$ if $n_1(s)d_2(s) - n_2(s)d_1(s) \in S(\Omega)$ has a zero of multiplicity n at s_0. $p_1(s)$ avoids $p_2(s)$ in Ω if $p_1(s)$ and $p_2(s)$ have no points of intersection in Ω.*

Note that the points of intersection in Ω between $p_1(s)$ and $p_2(s)$ do not depend on the particular choice of the coprime factorizations in $S(\Omega)$.

To illustrate Definition 3.7, consider, for example, $p_1(s) = (s+1)/s^2$ and $p_2(s) = (5s-1)/(s^3+s^2)$. The points of intersection between $p_1(s)$ and $p_2(s)$ in $\mathbb{C}_{+\infty}$ are at $s_0 = 1, s_0 = 2$, and $s_0 = \infty$ with multiplicity one and at $s_0 = 0$ with multiplicity two. The link between stabilization and avoidance is shown in the next theorem.

THEOREM 3.8. *Let $p(s), c(s) \in \mathbb{R}(s)$. Then the controller $c(s)$ Ω-stabilizes $p(s)$ if and only if $-c^{-1}(s)$ avoids $p(s)$ in Ω (or, equivalently, if and only if $-p^{-1}(s)$ avoids $c(s)$ in Ω).*

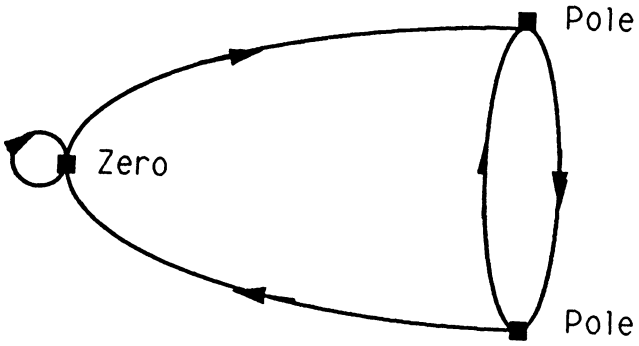
Proof. Let $p(s) = n_p(s)/d_p(s)$ and $c(s) = n_c(s)/d_c(s)$ be coprime decompositions of $p(s)$ and $c(s)$ in $S(\Omega)$. By Theorem 3.2, $c(s)$ Ω-stabilizes $p(s)$ if and only if $n_p(s)n_c(s) + d_p(s)d_c(s) \in U(\Omega)$. This last condition is satisfied if and only if $n_p(s)n_c(s) + d_p(s)d_c(s) \in S(\Omega)$ has no zeros in Ω or, alternatively, if and only if $-c(s)^{-1}$ avoids $p(s)$ in Ω. □

As a trivial consequence, notice that the plants that are Ω-stabilizable by a real constant feedback gain are precisely those that avoid a real value on Ω. With Theorem 3.8. we can formulate the general simultaneous stabilization problem of k plants in the form of an avoidance problem.

COROLLARY 3.9. *Let $p_i \in \mathbb{R}(s)$, $i = 1, \ldots, k$. The plants p_i are simultaneously Ω-stabilizable if and only if there exists a $q(s) \in \mathbb{R}(s)$ such that $q(s)$ avoids $p_i(s)$ in Ω, $i = 1, \ldots, k$, in which case $c(s) = -q^{-1}(s)$ is a Ω-stabilizing controller.*

The problem of the simultaneous Ω-stabilization of k plants thus has an easily understandable geometric interpretation. We are given a set of rational functions defined on a region Ω of the extended complex plane and we ask whether it is possible to find a rational function that avoids them all on Ω. If this is possible, then the plants are simultaneously Ω-stabilizable. Now, if it is possible to find a rational function that avoids k rational functions on Ω, then the same function avoids them all on $\omega = \mathbb{R}_\infty \cap \Omega$. *The existence of an ω-avoiding rational function is thus a necessary condition for simultaneous Ω-stabilization.* It is this necessary ω-stabilizability condition that we analyse in the next section.

When dealing with general stability regions Ω, the terminology and the notation get somewhat heavy. For our purposes a large class of such stability regions are equivalent: all the results contained in this paper are valid for closed, simply connected



GRAPH 1.1. *Parity interlacing property.*

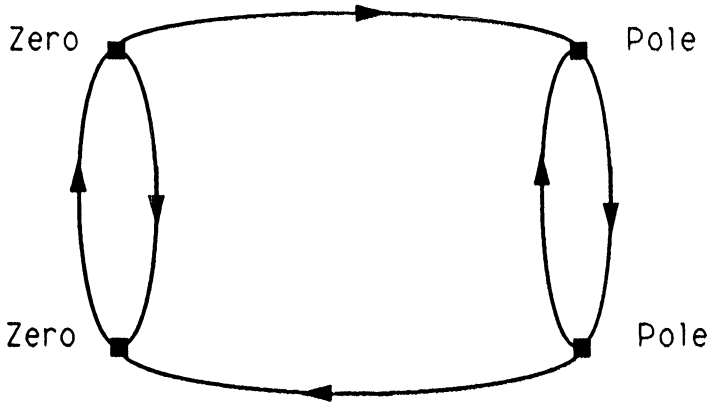
stability regions. In what follows we concentrate on *canonical* simply connected stability regions, in §4 we deal with $\mathbb{C}_{+\infty}$, and in §5 we analyse counterexamples in $\overline{D}$.

4. Stabilization on the real axis: The search for necessary conditions. The problem is simple. We examine k real rational functions on the interval $\mathbb{R}_{+\infty} = [0, \infty]$, where they are real-valued. They may have poles as well as zeros on $\mathbb{R}_{+\infty}$. Their behaviour can easily be represented on a two-dimensional graph as functions from $\mathbb{R}_{+\infty}$ to $\mathbb{R}_{\infty}$. Now we ask the question: "Is it possible to find a real rational function, with perhaps poles and zeros in $\mathbb{R}_{+\infty}$, which avoids this set of functions on the interval $\mathbb{R}_{+\infty}$?" In view of Corollary 3.9, this question is equivalent to the following: "Given a set of plants $p_i(s) \in \mathbb{R}(s)$, $i = 1, \ldots, k$, when is it possible to find a single controller $c(s) \in \mathbb{R}(s)$ such that $p_i c(1 + p_i c)^{-1}, p_i(1 + p_i c)^{-1}, c(1 + p_i c)^{-1}$, and $(1 + p_i c)^{-1}$ have no *real* unstable poles?" Obviously, this is a weaker requirement than simultaneous stabilization, where poles in the whole $\mathbb{C}_{+\infty}$ are to be avoided.

4.1. Stabilization of two plants and strong stabilization. In this section we will need the following well-known definitions.

DEFINITION 4.1. *Let $p(s) \in \mathbb{R}(s)$. $p(s)$ has the* parity interlacing property *if $p(s)$ has an even number (counting multiplicities) of poles between each pair of zeros in $\mathbb{R}_{+\infty}$. $p(s)$ has the* even interlacing property *if both $p(s)$ and $p^{-1}(s)$ have the parity interlacing property.*

An alternative way of defining this is by means of a graph. Let $z_i, p_j \in \mathbb{R}_{+\infty}$ $(i = 1, \ldots, l)$ $(j = 1, \ldots, m)$ be the l zeros and m poles of a plant $p(s) \in \mathbb{R}(s)$ in $\mathbb{R}_{+\infty}$. The plant $p(s)$ has the parity interlacing property if and only if the succession of its poles and zeros on $\mathbb{R}_{+\infty}$, as s increases from zero to infinity, corresponds to a possible path in Graph 1.1. In the same vein, $p(s)$ has the even interlacing property if and only if the succession of its poles and zeros on $\mathbb{R}_{+\infty}$ correspond to a possible path in Graph 1.2. For example, the succession of poles and zeros of $p(s) = (s-1)/((s-3)(s-2)s)$ gives the following pattern: PZPPZ and hence $p(s)$ has the parity interlacing property, but not the even interlacing property. The same kind of figure will be used in Theorem 4.13 to describe a $\mathbb{R}_{+\infty}$-stabilizability condition for three plants, but first we analyse the two-plant case.



GRAPH 1.2. *Even interlacing property.*

THEOREM 4.2. *Let $p(s) \in \mathbb{R}(s)$. If there exists a stable controller that $\mathbb{R}_{+\infty}$-stabilizes $p(s)$, then $p(s)$ has the parity interlacing property.*

Proof. Let $c(s)$ be a stable $\mathbb{R}_{+\infty}$-stabilizing controller of $p(s)$. Then by Theorem 3.8, $c(s)$ avoids $-p^{-1}(s)$ on $\mathbb{R}_{+\infty}$. Since $c(s)$ is stable, it also avoids ∞ on $\mathbb{R}_{+\infty}$. Suppose, to get a contradiction, that $p(s)$ has an odd number of poles between two zeros on $\mathbb{R}_{+\infty}$. Then $-p^{-1}(s)$ has an odd number of zeros between two poles on $\mathbb{R}_{+\infty}$. But then $c(s)$ has to avoid both a rational function $-p^{-1}(s)$, which has an odd number of zeros between two of its poles on $\mathbb{R}_{+\infty}$ and ∞ on $\mathbb{R}_{+\infty}$. This is impossible and hence $p(s)$ has an even number of poles between two zeros on $\mathbb{R}_{+\infty}$. □

A stronger version of Theorem 4.2 can be obtained. It can be shown that the parity interlacing property is in fact sufficient for $\mathbb{R}_{+\infty}$-stabilizability by a stable controller. This last result in turn is contained under a stronger form in the next theorem.

THEOREM 4.3. *Let $p(s) \in \mathbb{R}(s)$. There exists a stable controller that stabilizes $p(s)$ if and only if $p(s)$ has the parity interlacing property.*

The proof of this fundamental fact was first given in [32]. The reader may find both an elementary and an advanced proof in [24].

COROLLARY 4.4. *A plant is stabilizable by a stable controller if and only if it is $\mathbb{R}_{+\infty}$-stabilizable by a stable controller.*

Proof. Use the two previous theorems together with the fact that a stabilizing controller is also $\mathbb{R}_{+\infty}$-stabilizing. □

Using this last corollary, we stress in the next theorem a fundamental property of simultaneous stabilization of two plants: if there exists a controller such that the closed-loop transfer functions associated with each plant have no *real* unstable poles, then there exists a controller that simultaneously stabilizes the two plants.

THEOREM 4.5. *Two plants are simultaneously stabilizable if and only if they are simultaneously $\mathbb{R}_{+\infty}$-stabilizable.*

Proof. Let p_1 and $p_2 \in \mathbb{R}(s)$, and let $p_i = n_i/d_i$ be any coprime decompositions, $i = 1, 2$. By Theorem 3.2, p_i are simultaneously stabilizable if and only if there exist $n_c, d_c \in S$ such that $n_c n_i + d_c d_i \in U$, $i = 1, 2$. Since n_1, d_1 are coprime, there exists $x, y \in S$ such that $n_1 x + d_1 y = 1$. Any controller $c = (x + r d_1)/(y - r n_1)$, where $r \in S$ is a stabilizing controller of p_1. In fact, it can easily be proved (see [24] or [10]) that

any stabilizing controller of p_1 can be written in the form $c = (x + rd_1)/(y - rn_1)$ for some $r \in S$. Therefore p_1 and p_2 are simultaneously stabilizable if and only if $(x+rd_1)n_2+(y-rn_1)d_2 = xn_2+yd_2+r(d_1n_2-d_2n_1) \in U$ for some $r \in S$. If $xn_2+yd_2 = 0$, then p_1 and p_2 are both simultaneously stabilizable and simultaneously $\mathbb{R}_{+\infty}$-stabilizable, so we rule out this case. Assume that $xn_2 + yd_2 \neq 0$. By Definition 3.4 the equation above has a solution if and only if the plant $q = (d_1n_2-d_2n_1)/(xn_2+yd_2)$ is strongly stabilizable. We could have derived exactly the same computations for $\mathbb{R}_{+\infty}$-stabilizability by replacing S by $S(\mathbb{R}_{+\infty})$ and U by $U(\mathbb{R}_{+\infty})$ in our derivations. We have thus also that p_1 and p_2 are simultaneously $\mathbb{R}_{+\infty}$-stabilizable if and only if $q = (d_1n_2 - d_2n_1)/(xn_2 + yd_2)$ is strongly $\mathbb{R}_{+\infty}$-stabilizable. But now, by applying Corollary 4.4 the theorem is proved. □

In the proof we show that p_1 and p_2 are simultaneously stabilizable if and only if $q = (d_1n_2 - d_2n_1)/(xn_2 + yd_2)$ is strongly stabilizable. This intermediate plant q is constructed with the coprime decompositions $p_i = n_i/d_i$, $i = 1, 2$ together with the solutions $x, y \in S$ of $n_1x + d_1y = 1$. With a weak additional condition on the poles of p_1 and p_2, it is possible to put this equivalence between simultaneous stabilization of two plants and strong stabilization of a single plant in a new and more obvious form. Roughly speaking, two plants are simultaneously stabilizable if and only if their difference is strongly stabilizable.

THEOREM 4.6. *Let $p_i(s) \in \mathbb{R}(s)$, $i = 1, 2$ and suppose that $p_1(s)$ and $p_2(s)$ have no common poles on $\mathbb{R}_{+\infty}$. Then $p_1(s)$ and $p_2(s)$ are simultaneously stabilizable if and only if $p_1(s) - p_2(s)$ is strongly stabilizable.*

Proof. With p_1 and $p_2 \in \mathbb{R}(s)$, let $p_i = n_i/d_i$ be any coprime decomposition in $S(\mathbb{R}_{+\infty})$, $i = 1, 2$ and let $x, y \in S(\mathbb{R}_{+\infty})$ be such that $n_1x + d_1y = 1$. Then by Corollary 3.6 and Theorem 4.5, p_i are simultaneously stabilizable if and only if there exist $n_c, d_c \in S(\mathbb{R}_{+\infty})$ such that $n_cn_i + d_cd_i \in U(\mathbb{R}_{+\infty})$, $i = 1, 2$. By using the same argument as in the proof of Theorem 4.5, these two equations can be simultaneously fulfilled if and only if $(n_1d_2 - n_2d_1)r + (n_2x + d_2y) \in U(\mathbb{R}_{+\infty})$ has a solution for some $r \in S(\mathbb{R}_{+\infty})$. Such an equation has a solution if and only if (n_2x+d_2y) is nonzero and always has the same sign at the zeros of $(n_1d_2 - n_2d_1)$ on $\mathbb{R}_{+\infty}$ (this result is crucial and far reaching; a proof of it can be found in [24, p. 38]). Under the assumption that d_1 and d_2 have no common zeros on $\mathbb{R}_{+\infty}$ and with some additional algebra, this last condition can be shown to be equivalent to requiring that d_1d_2 always have the same sign at the zeros of $(n_1d_2 - n_2d_1)$. This in turn is equivalent to requiring $p_1 - p_2$ to be stabilizable by a stable controller. □

Theorem 4.6 is a stronger form of the results contained in [24] and in [30], which state, respectively, "if p_1 is stable, then p_1, p_2 are simultaneously stabilizable if and only if $p_1 - p_2$ is stabilizable by a stable controller" and "if p_1 and p_2 have no common poles in $\mathbb{C}_{+\infty}$, then they are simultaneously stabilizable if and only if $p_1 - p_2$ is stabilizable by a stable controller." Both of these results are contained in Theorem 4.6.

4.2. Stabilization of three plants and unit stabilization. We now investigate the case of three plants and its link with unit stabilization. In this subsection we consider only the case of plants that do not all intersect at the same point. This assumption is not generic (for example, strictly proper plants all intersect at infinity) and will be dropped in §4.4. We start with a crucial theorem.

THEOREM 4.7. *Let $p_i(s) \in \mathbb{R}(s)$ $i = 1, 2, 3$. Suppose that $p_1(s), p_2(s), p_3(s)$ have no common point of intersection in $\mathbb{C}_{+\infty}$ (i.e., there is no $s_0 \in \mathbb{C}_{+\infty}$ for which $p_1(s_0) = p_2(s_0) = p_3(s_0)$). Let $p_i = n_i/d_i$, $i = 1, 2, 3$ be any coprime decompositions*

and define $a_{ij} = n_i d_j - n_j d_i$, $(i,j = 1,2,3)$. *Then* $p_i(s)$, $i = 1,2,3$ *are simultaneously stabilizable if and only if there exist* $u_i \in U$, $i = 1,2,3$ *such that* $a_{12}u_3 + a_{23}u_1 + a_{31}u_2 = 0$.

Proof. Let $x, y \in S$ be solutions to $n_1 x + d_1 y = 1$ and define $b_i = n_i x + d_i y$, $i = 2,3$. It is easy to check that $b_2 a_{13} - b_3 a_{12} = a_{23}$. If $n_i n_c + d_i d_c = u_i$ for some $n_c, d_c \in S$, $i = 1,2,3$, then $a_{12}u_3 + a_{23}u_1 + a_{31}u_2 = 0$, hence the necessity is proved. For sufficiency, suppose that there exists $u_i \in U$, $i = 1,2,3$ such that $a_{12}u_3 + a_{23}u_1 + a_{31}u_2 = 0$. Using $b_2 a_{13} - b_3 a_{12} = a_{23}$, we have $a_{12}u_3 + (b_2 a_{13} - b_3 a_{12})u_1 + a_{31}u_2 = a_{12}(u_3 - b_3 u_1) + a_{31}(u_2 - b_2 u_1) = 0$. Since there is no $s_0 \in \mathbb{C}_{+\infty}$ for which $p_1(s_0) = p_2(s_0) = p_3(s_0)$, this implies in algebraic terms that a_{12} and a_{31} are coprime. Hence there exists some $r \in S$ for which $a_{31}r = u_3 - b_3 u_1$ and $a_{12}r = -u_2 + b_2 u_1$. Defining $r' = r/u_1$, we have that $a_{31}r' + b_3 = u_3/u_1 \in U$ and $a_{21}r' + b_2 = u_2/u_1 \in U$. But now, defining $n_c = x + r'd_1$ and $d_c = y - r'n_1$, the theorem is proved, since for d_c, n_c we have $n_i n_c + d_i d_c \in U$, $i = 1,2,3$. □

It is known (see [29] or [13]) that, modulo an additional condition, the three-plant problem can be reduced to one of finding a single controller that is stable, inverse stable (from here on we will refer to such controllers as unit controllers), and that stabilizes a single plant. Let us make this connection more obvious by using Theorem 4.7.

THEOREM 4.8. *Let* $p_i \in \mathbb{R}(s)$, $i = 1,2,3$ *and let* $p_i = n_i/d_i$, $i = 1,2,3$ *be arbitrary coprime decompositions in* S. *Suppose that* p_1 *avoids* p_2 *in* $\mathbb{C}_{+\infty}$. *Then* p_i, $i = 1,2,3$ *are simultaneously stabilizable if and only if* $(n_3 d_1 - n_1 d_3)/(n_2 d_3 - d_2 n_3)$ *is unit stabilizable, i.e., stabilizable by a unit controller.*

Proof. Since p_1 avoids p_2 in $\mathbb{C}_{+\infty}$ we have $n_1 d_2 - n_2 d_1 = u \in U$. Trivially, p_1, p_2, and p_3 have no common point of intersection in $\mathbb{C}_{+\infty}$, since p_1 and p_2 do not intersect in $\mathbb{C}_{+\infty}$. We may thus apply Theorem 4.7. Therefore, p_i, $i = 1,2,3$ are simultaneously stabilizable if and only if there exist $u_i \in U$, $i = 1,2,3$ such that $uu_3 + a_{23}u_1 + a_{31}u_2 = 0$. This last equation has a solution if and only if there exists u_1 and u_2 in U for which $a_{23}u_1 + a_{31}u_2 \in U$ or, equivalently, if and only if $(n_3 d_1 - n_1 d_3)/(n_2 d_3 - d_2 n_3)$ is unit stabilizable. □

Contrary to the similar result for strong stabilization of Theorem 4.6 we have no interpretation to propose for $(n_3 d_1 - n_1 d_3)/(n_2 d_3 - d_2 n_3)$ in terms of the plants p_1, p_2, and p_3.

As an illustration of the theorem, consider the plants $p_1(s) = 1, p_2(s) = -1/s$, and $p_3(s) = -(s-1)/s$. We can take for coprime decompositions $n_1 = 1, d_1 = 1$, $n_2 = -1/(s+1), d_2 = s/(s+1)$, and $n_3 = -(s-1)/(s+1), d_3 = s/(s+1)$. p_1 and p_2 have no intersections in $\mathbb{C}_{+\infty}$, since $n_1 d_2 - n_2 d_1 = 1 \in U$. We can apply Theorem 4.8. Therefore p_i, $i = 1,2,3$ are simultaneously stabilizable if and only if $(2s^2 + s - 1)/(s^2 - 2s)$ is unit stabilizable.

An important special case of the problem of the stabilizability of three plants is therefore equivalent to the stabilizability of a single plant by a unit controller. This can be proven rigorously in the sense that for any plant $p(s)$ it is possible to construct three plants $p_1(s)$, $p_2(s)$, and $p_3(s)$ such that $p(s)$ is stabilizable by a unit controller if and only if $p_i(s)$, $i = 1,2,3$ are simultaneously stabilizable. This equivalence is one of the reasons for investigating conditions under which a plant is stabilizable by a unit controller. For the same reason as before, we first examine the condition under which a single plant is $\mathbb{R}_{+\infty}$-stabilizable by a unit controller. This condition is rather simple.

THEOREM 4.9. *Let* $p(s) \in \mathbb{R}(s)$. *There exists a unit controller that* $\mathbb{R}_{+\infty}$*-stabilizes*

$p(s)$ if and only if $p(s)$ has the even interlacing property.

Proof. Necessity is trivial: Apply Theorem 4.2 to $p(s)$ and $p^{-1}(s)$. Sufficiency can be shown by modifying slightly the proof of Theorem 3.2 in [29], in which the author proves that a stable controller with no real unstable zeros exists for any plant that satisfies the even interlacing condition. □

Obviously, the even interlacing property is a necessary condition for stabilization of a plant by a unit controller. By similarity with the strong stabilization condition and with Corollary 4.4, this necessary condition was also conjectured to be sufficient. The conjecture is false, however, and we give a counterexample in §5. Before proceeding to this, we investigate additional conditions under which three plants are simultaneously $\mathbb{R}_{+\infty}$-stabilizable.

4.3. Alternative conditions for stabilization of three plants. In §4.2 we showed the connection between simultaneous stabilizability of three plants and stabilizability of a related plant with a unit controller. Here we provide new conditions for simultaneous stabilizability and $\mathbb{R}_{+\infty}$-stabilizability of three plants. We start with a theorem that is of independent interest. Roughly speaking, it says that three plants are simultaneously stabilizable if and only if there exist three *stable* plants that have pairwise the same intersections in $\mathbb{C}_{+\infty}$ as the original three plants.

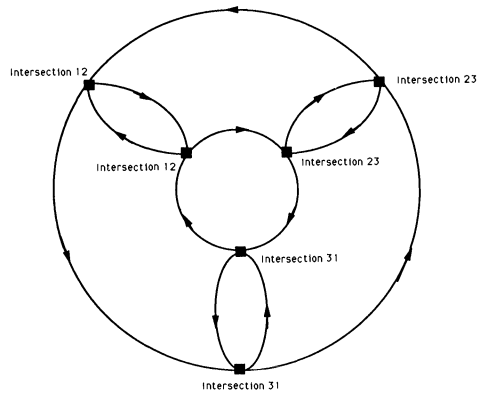
THEOREM 4.10. *Let $p_i(s) \in \mathbb{R}(s)$, $i = 1,2,3$. Suppose that $p_1(s), p_2(s), p_3(s)$ have no common point of intersection in $\mathbb{C}_{+\infty}$ (i.e., there is no $s_0 \in \mathbb{C}_{+\infty}$ for which $p_1(s_0) = p_2(s_0) = p_3(s_0)$). Then $p_i(s)$, $i = 1,2,3$ are simultaneously stabilizable if and only if there exist $p_i'(s) \in S$, $i = 1,2,3$ such that $p_i(s)$ and $p_j(s)$ have pairwise the same intersections in $\mathbb{C}_{+\infty}$ as $p_i'(s)$ and $p_j'(s)$ when $i,j = 1,2,3$.*

Proof. Let $p_i = n_i/d_i$, $i = 1,2,3$ be arbitrary coprime decompositions and define $a_{ij} = n_i d_j - n_j d_i$ $(i,j = 1,2,3)$. Suppose first that there exists $p_i'(s) \in S$, $i = 1,2,3$ such that $p_i(s)$ and $p_j(s)$ have pairwise the same intersections in $\mathbb{C}_{+\infty}$ as $p_i'(s)$ and $p_j'(s)$. In algebraic terms this means that $p_i' - p_j' = u_{ij} a_{ij}$ $(i,j = 1,2,3)$ for some units $u_{ij} \in U$ $(i,j = 1,2,3)$. Putting $u_1 = u_{23}, u_2 = u_{31}$, and $u_3 = u_{12}$ in Theorem 4.7, we get that $p_i(s)$, $i = 1,2,3$ are simultaneously stabilizable. To prove necessity, suppose that $p_i(s)$, $i = 1,2,3$ are simultaneously stabilizable and have no common intersection in $\mathbb{C}_{+\infty}$. Again, by Theorem 4.7 there exist $u_i \in U$, $i = 1,2,3$ such that $a_{12}u_3 + a_{23}u_1 + a_{31}u_2 = 0$. Take any $r_2 \in S$ and define $r_1 = r_2 + u_3 a_{12} \in S$ and $r_3 = r_2 - u_1 a_{23} \in S$. Then we have that $r_1 - r_2 = a_{12}u_3$, $r_2 - r_3 = a_{23}u_1$, but also $r_3 - r_1 = a_{31}u_2$. And thus $r_i \in S$, $i = 1,2,3$ are such that r_i and r_j have pairwise the same intersections in $\mathbb{C}_{+\infty}$ as $p_i(s)$ and $p_j(s)$ for $i,j = 1,2,3$. This ends the proof. □

As we argued in §3.2, the results that we obtain in this section are still valid for general regions Ω that satify the assumptions stated in §2. In particular, we may derive the counterpart of Theorem 4.10 for the region $\Omega = \mathbb{R}_{+\infty}$.

THEOREM 4.11. *Let $p_i(s) \in \mathbb{R}(s)$, $i = 1,2,3$. Suppose that $p_1(s), p_2(s), p_3(s)$ have no common point of intersection on $\mathbb{R}_{+\infty}$ (i.e., there is no $s_0 \in \mathbb{R}_{+\infty}$ for which $p_1(s_0) = p_2(s_0) = p_3(s_0)$). Then $p_i(s)$, $i = 1,2,3$ are simultaneously $\mathbb{R}_{+\infty}$-stabilizable if and only if there exist $p_i'(s) \in S(\mathbb{R}_{+\infty})$, $i = 1,2,3$ such that $p_i(s)$ and $p_j(s)$ have pairwise the same intersections on $\mathbb{R}_{+\infty}$ as $p_i'(s)$ and $p_j'(s)$ when $i,j = 1,2,3$.*

The interest of this last result is that, while we do not know a tractable test to check the condition in Theorem 4.10, we have one for the condition in Theorem 4.11. The existence of three rational functions with no poles on $\mathbb{R}_{+\infty}$ that *mimic* the pairwise intersections of three plants on $\mathbb{R}_{+\infty}$ relies on an interlacing property that we state hereafter.



GRAPH 1.3. *3-interlacing property.*

DEFINITION 4.12. *Let* $p_i(s) \in \mathbb{R}(s)$, $i = 1,2,3$. *Suppose that* $p_1(s), p_2(s), p_3(s)$ *have no common point of intersection on* $\mathbb{R}_{+\infty}$. *Then* $p_i(s)$, $i = 1,2,3$ *have the* 3-interlacing property *if the succession of their intersections on* $\mathbb{R}_{+\infty}$, *as* s *increases from zero to infinity, corresponds to a possible path in Graph* 1.3.

We can now prove our theorem.

THEOREM 4.13. *Let* $p_i(s) \in \mathbb{R}(s)$, $i = 1,2,3$. *Suppose that* $p_1(s), p_2(s), p_3(s)$ *have no common point of intersection on* $\mathbb{R}_{+\infty}$. *Then* $p_i(s)$, $i = 1,2,3$ *are simultaneously* $\mathbb{R}_{+\infty}$*-stabilizable if and only if they have the* 3*-interlacing property.*

Proof. Suppose that $p_1(s), p_2(s), p_3(s)$ have no common point of intersection on $\mathbb{R}_{+\infty}$. By Theorem 4.11, $p_i(s)$, $i = 1,2,3$ are simultaneously $\mathbb{R}_{+\infty}$-stabilizable if and only if there exist $p_i'(s) \in S(\mathbb{R}_{+\infty})$, $i = 1,2,3$ such that $p_i(s)$ and $p_j(s)$ have pairwise the same intersections on $\mathbb{R}_{+\infty}$ as $p_i'(s)$ and $p_j'(s)$. The fact that $p_i'(s)$ have no poles on $\mathbb{R}_{+\infty}$ implies that not all successions of pairwise intersections are possible, i.e., the succession of intersections between three continous functions from $\mathbb{R}_{+\infty}$ to $\mathbb{R}$ is not arbitrary. We claim that the successions that are possible are precisely those that represent a possible path in Graph 1.3. To prove this, note that at each point $s_0 \in \mathbb{R}_{+\infty}$ where the p_i' do not pairwise intersect we have $p_i'(s_0) > p_j'(s_0) > p_k'(s_0)$ for some $i,j,k = 1,2,3$. In this way we can associate, to each point $s_0 \in \mathbb{R}_{+\infty}$ where the plants p_i' do not pairwise intersect, one of the six orderings $p_1' < p_2' < p_3'$, $p_1' < p_3' < p_2'$, $p_2' < p_1' < p_3'$, $p_2' < p_3' < p_1'$, $p_3' < p_1' < p_2'$, or $p_3' < p_2' < p_1'$. If s_0 and s_1 are two points on $\mathbb{R}_{+\infty}$ such that $p_i'(s)$ have no pairwise intersections on $[s_0, s_1]$, then, because the $p_i'(s)$ are continuous, the ordering at s_0 and s_1 are the same. Hence, the ordering changes precisely at the pairwise intersections of the $p_i'(s)$. For example, the ordering $p_1' < p_2' < p_3'$ changes to $p_1' < p_3' < p_2'$ after an intersection between p_2' and p_3'. Note also that not all changes are admitted, for example, $p_1' < p_2' < p_3'$ cannot be changed to $p_3' < p_2' < p_1'$ after a single intersection. Representing the six possible orderings above in a graph together with all possible changes at the intersections yields the Graph 1.3. Necessity is proved. To prove sufficiency, it suffices to show that given a succession of pairwise intersections on $\mathbb{R}_{+\infty}$ that follows a path in Graph 1.3, it is always possible to construct three functions in $S(\mathbb{R}_{+\infty})$ that do not intersect simultaneously on $\mathbb{R}_{+\infty}$ and whose pairwise intersections are the given points. We do not give a technical, and tedious, proof of this here. Instead, we outline the sketch

of a constructive procedure. First translate the problem onto I by using the usual conformal equivalence. Then construct three continuous functions that satisfy the desired property. By careful use of the fact that polynomials are dense in the set of continuous functions on I, construct three polynomials that also satisfy this property. Notice then that polynomials are members of $S(I)$, so that by using the conformal equivalence again the theorem is proved. □

The case where the plants do intersect on $\mathbb{R}_{+\infty}$ is analysed in §4.4 below.

Theorem 4.13 and the 3-interlacing property are equivalent to an algebraic condition recently given in [30]. It was obtained independently by the authors. Again, it is a necessary condition for simultaneous stabilizability of three plants, since it is necessary and sufficient for $\mathbb{R}_{+\infty}$-stabilizability. In the conclusion of [30] it is conjectured that this condition is also sufficient for stabilizability, but we will prove in §5 that this is not true.

To illustrate the use of Theorem 4.13, we analyse an example given in the literature [13]. A natural question when analysing the simultaneous stabilizability of three plants is: "Given three plants that are simultaneously stabilizable, they are of course pairwise simultaneously stabilizable. Is the converse also true?" Unfortunately, the answer is no. Ghosh provided a counterexample to this: $p_1(s) = (s-7)/(s-4.6)$, $p_2(s) = (s-2)/(2s-2.6)$, and $p_3(s) = (s-6)/(4.8s-24.6)$ are pairwise simultaneously stabilizable, but it is shown in [13] that they are not simultaneously stabilizable. Application of our Theorem 4.13 easily shows that they are not even $\mathbb{R}_{+\infty}$-simultaneously stabilizable. The intersections between p_1 and p_2 are $\sigma_{12} = 1$ and $\sigma_{12} = 9$. For the other two pairwise intersections we get: $\sigma_{23} = 3$ and $\sigma_{23} = 4$, $\sigma_{31} = 7.34$ and $\sigma_{31} = 5.17$. Note that for these three plants all the intersections happen to be on $\mathbb{R}_{+\infty}$, which is by no means generic. Ordering the succesion of pairwise $\mathbb{R}_{+\infty}$-intersections we get: $\sigma_{12}, \sigma_{23}, \sigma_{23}, \sigma_{31}, \sigma_{31}, \sigma_{12}$. This does not correspond to a possible path in Graph 1.3. Hence $p_i(s)$, $i = 1, 2, 3$, do not have the 3-interlacing property and, by Theorem 4.13, the three plants are not simultaneously stabilizable.

As a final remark on Theorem 4.13, it is worth noting that our 3-interlacing property can be extended to more than three plants. If k plants are simultaneously stabilizable, then the same sequence of pairwise intersections on $\mathbb{R}_{+\infty}$ is achievable by the pairwise intersections of k $\mathbb{R}_{+\infty}$-stable plants. This provides a necessary condition for simultaneous stabilization of k plants. We do not develop this further here because we believe that the results contained in the next section overshadow the interest of stabilization conditions of k plants when the plants do not intersect. We end §4 by analysing the case where there exists some $s_0 \in \mathbb{R}_{+\infty}$ such that $p_i(s_0) = w_0$, $i = 1, \ldots, k$.

4.4. Simultaneous $\mathbb{R}_{+\infty}$ stabilization for intersecting plants. All the conditions in §§4.2 and 4.3 are for the case where the three plants have no common point of intersection on either $\mathbb{C}_{+\infty}$ or $\mathbb{R}_{+\infty}$. There exists an important special case for which this condition is not satisfied. When the plants are all strictly proper, they all take the value 0 at infinity so that they have a common point of intersection at infinity. It is this special structure that partly motivates the next result, which is the central result of this section. It shows that, for simultaneous $\mathbb{R}_{+\infty}$-stabilizability, the conditions are much simpler when the plants have a common point of intersection on $\mathbb{R}_{+\infty}$. Note that the theorem applies not just to the three-plant case but to the general k plants case.

THEOREM 4.14. *Let $p_i(s) \in \mathbb{R}(s)$, $i = 1, \ldots, k$ and suppose that there exists a value $s_0 \in \mathbb{R}_{+\infty}$ such that the plants intersect at s_0 (i.e., there exists some $s_0 \in \mathbb{R}_{+\infty}$*

and some $w_0 \in \mathbb{R}_\infty$ such that $p_i(s_0) = w_0$, $i = 1, \ldots, k$). Then the plants are simultaneously $\mathbb{R}_{+\infty}$-stabilizable if and only if they are pairwise simultaneously $\mathbb{R}_{+\infty}$-stabilizable.

Proof. Necessity is obvious. We prove sufficiency by showing that, under the assumptions that the k plants $p_i(s)$, $i = 1, \ldots, k$ intersect at $s_0 \in \mathbb{R}_{+\infty}$ and that they are pairwise simultaneously $\mathbb{R}_{+\infty}$-stabilizable, it is possible to find a rational function $q(s)$ that avoids them all on $\mathbb{R}_{+\infty}$. The result will then follow by Corollary 3.9.

For simplicity we assume that $s_0 = 0$, and we define $w_0 = p_i(s_0) = p_i(0)$; the proof for an arbitrary s_0 goes along the same line. We assume also that $w_0 \neq \infty$. If not, we can redefine $p_i' = 1/p_i$ and $w_0' = 0$. First use the bilinear transformation that maps $\mathbb{C}_{+\infty}$ onto $\overline{D}$. Under this transformation, we get $p_i'(z) = p_i((1+z)/(1-z))$. Since $p_i(0) = w_0$, we have $p_i'(-1) = w_0$ for $i = 1, \ldots, k$. In view of this, define $p_i''(z) = p_i'(z) - w_0$. It is clear that $p_i''(z)$ all have a zero at $z_0 = -1$. Also from our assumptions $p_i''(z)$, $i = 1, \ldots, k$ are real rational and are pairwise simultaneously I-stabilizable. To end the proof, it remains to show that $p_i''(z)$ are simultaneously I-stabilizable, i.e., that there exists a rational function that avoids $p_i''(z)$ on I.

To see this we define k continuous functions $v_i(z)$ from I to $\mathbb{R}$ by $v_i(z) = \arctan p_i''(z)$, $z \in I$. Here the inverse tangent function has to be taken with an *unwrapped argument*, i.e., the function $v_i(z)$ is made continuous from $\mathbb{R}_\infty$ to $\mathbb{R}$ as z increases from -1 to 1 by choosing an appropriate branch of the inverse tangent function at the real poles of $p(z)$. Since $p_i''(-1) = 0$, we may chose $v_i(-1) = 0$. Some manipulations show that a rational function $r(z)$ avoids $p_i''(z)$ on I, $i = 1, \ldots, k$, if and only if $v_i(z) - n\pi < \arctan r(z) < v_i(z) - (n-1)\pi$ for all $z \in I$, $i = 1, \ldots, k$ and for some $n \in \mathbb{N}$. In the sequel our objective is to construct such an $r(z)$. We therefore need an intermediate result.

We show that, because $p_i''(z)$ are pairwise simultaneously I-stabilizable, we have $\mid v_i(z) - v_j(z) \mid < \pi$, for all $z \in I$, $i, j = 1, \ldots, k$. Suppose, by contradiction, that for some i, j and some $z_0 \in I$ we have $\mid v_i(z_0) - v_j(z_0) \mid \geq \pi$. Then, since $\mid v_i(-1) - v_j(-1) \mid = 0$, and since $v_i(z)$ are continuous, there must exist $z_1 \in [-1, z_0]$ such that $\mid v_i(z_1) - v_j(z_1) \mid = \pi$. But then, given any rational function $r(z)$ and the continuous function $v(z) = \arctan r(z)$ from I to $\mathbb{R}$, there exists some $z_2 \in [-1, z_1]$ such that either $v_i(z_2) - v(z_2) = n\pi$ or $v_j(z_2) - v(z_2) = n\pi$ for some $n \in \mathbb{N}$. Assume $v_i(z_2) - v(z_2) = n\pi$. Then $v_i(z_2) = v(z_2) + n\pi$ and, taking the tangent of both sides, $p_i''(z_2) = r(z_2)$. This shows that every rational function intersects either $p_i''(z)$ or $p_j''(z)$ at some $z \in I$. This last statement contradicts the fact that $p_i''(z)$ and $p_j''(z)$ are simultaneously I-stabilizable and so we have proved that $\mid v_i(z) - v_j(z) \mid < \pi$, for all $z \in I$, $i, j = 1, \ldots, k$. We now construct a stabilizing controller.

Define $w(z) : z \to \min_{i=1,\ldots,k} v_i(z)$. $w(z)$ is a continuous function from I to $\mathbb{R}$. By the above argument, $\mid v_i(z) - v_j(z) \mid < \pi$, for all $z \in I$, $i, j = 1, \ldots, k$ and hence $v_i(z) - \pi < w(z) \leq v_i(z)$, for all $z \in I$, $i = 1, \ldots, k$. We define $w'(z) = w(z) - \epsilon$ with ϵ sufficiently small so that $v_i(z) - \pi < w'(z) < v_i(z)$, for all $z \in I$, $i = 1, \ldots, k$. Some algebraic manipulations, together with the fact that polynomials are uniformly dense in the set of continuous functions from I to $\mathbb{R}$, show that, given $w'(z)$ and $\epsilon > 0$ it is possible to find a rational function $q(z)$ such that $\mid w'(z) - \arctan q(z) \mid < \epsilon$, for all $z \in I$. But then, for sufficiently small ϵ, we have $v_i(z) - \pi < \arctan q(z) < v_i(z)$, for all $z \in I$, $i = 1, \ldots, k$. Taking the tangent of both sides, this last statement clearly shows that $q(z)$ avoids $p_i''(z)$ for $i = 1, \ldots, k$ and $z \in I$. This in turn implies by Corollary 3.9 that $p_i''(z)$, and hence $p_i'(z)$, are simultaneously I-stabilizable. The equivalence between the simultaneous I-stabilizability of the $p_i'(z)$ and that of the

$\mathbb{R}_{+\infty}$-stabilizability of the $p_i(s)$ ends the proof. □

Using this theorem, the next results are straightforward and their proofs are left to the reader.

COROLLARY 4.15. *Let $p_i(s) \in \mathbb{R}(s)$, $i = 1, \ldots, k$ and suppose that there exists a value $s_0 \in \mathbb{R}_{+\infty}$ such that the plants intersect at s_0. Then the plants are simultaneously $\mathbb{R}_{+\infty}$-stabilizable if and only if they are pairwise simultaneously stabilizable.*

COROLLARY 4.16. *Let $p_i(s) \in \mathbb{R}(s)$, $i = 1, \ldots, k$ and suppose that $p_i(s)$ have a common pole or a common zero on $\mathbb{R}_{+\infty}$. Then the plants are simultaneously $\mathbb{R}_{+\infty}$-stabilizable if and only if they are pairwise simultaneously $\mathbb{R}_{+\infty}$-stabilizable.*

COROLLARY 4.17. *Let $p_i(s) \in \mathbb{R}(s)$, $i = 1, \ldots, k$ be strictly proper (they all have a zero at infinity). The plants are simultaneously $\mathbb{R}_{+\infty}$-stabilizable if and only if they are pairwise simultaneously $\mathbb{R}_{+\infty}$-stabilizable.*

Note that in the above example of Ghosh the plants are not strictly proper.

These are only some of the possible corollaries of Theorem 4.14. Their main common interest is that, contrary to most of the results on simultaneous stabilisation, they provide tractable tests to decide whether k plants are simultaneously $\mathbb{R}_{+\infty}$-stabilizable. Most of the known results on simultaneous stabilization of more than two plants are only restatements of untractable conditions into other untractable conditions. Here we have provided tractable tests, since the simultaneous stabilizability of two plants can be tested by using only a finite number of rational operations (see [1]). On the other hand, the drawback of our conditions is that, even though they are necessary and sufficient for $\mathbb{R}_{+\infty}$-stabilizability, they are only necessary conditions for $\mathbb{C}_{+\infty}$-stabilizability. We show in §5 that the conditions that we have obtained are in general not sufficient and, as soon as k is greater than two, it is necessary to look at the behaviour of the plants in the whole extended right half complex plane and not just on the extended positive real axis.

5. Stabilization in the complex plane. In the previous section we have found necessary and sufficient conditions for $\mathbb{R}_{+\infty}$-stabilizability of a single plant by a stable controller (parity interlacing property) and by a unit controller (even interlacing property). We have also treated the case of simultaneous $\mathbb{R}_{+\infty}$-stabilization of three or more plants (3-interlacing condition in the case of three plants that do not intersect, and pairwise stabilizability in the case of k plants that intersect on $\mathbb{R}_{+\infty}$). All these conditions are, as we have shown, necessary conditions for stabilizability in the usual sense, i.e., $\mathbb{C}_{+\infty}$-stabilizability. One of these conditions has also been shown to be sufficient for $\mathbb{C}_{+\infty}$-stabilizability, namely, two plants are simultaneously stabilizable if and only if they are simultaneously $\mathbb{R}_{+\infty}$-stabilizable. It was hoped that this property would flow on to the case $k \geq 3$. The implicit conjecture "k plants are simultaneously stabilizable if and only if they are $\mathbb{R}_{+\infty}$-stabilizable" has obviously been a driving motivation for many of the partial results on simultaneous stabilization. In this section we give counterexamples showing that $\mathbb{R}_{+\infty}$-stabilizability does not, in general, imply $\mathbb{C}_{+\infty}$-stabilizability.

For convenience (mainly because $\overline{D}$ is a bounded set), we give the counterexamples of this section in $\overline{D}$ rather than in $\mathbb{C}_{+\infty}$. The counterpart of $\mathbb{R}_{+\infty}$ is then $I = \overline{D} \cap \mathbb{R}_{\infty}$. It must be clear, however, that *all our counterexamples have a counterpart in continuous time.* The equivalence can be shown by using the bilinear transformation, and we illustrate it for the first theorem.

We start with the easiest counterexample.

THEOREM 5.1. *Let $p_1(z) = 0$, $p_2(z) = z/(z+2)$, $p_3(z) = 2z/(z+2)$, and $p_4(z) = 2z/((z+2)(2-kz))$ be four discrete time systems. If $k > e^{26}$, then $p_i(z)$,*

$i = 1, \ldots, 4$ are simultaneously I-stabilizable but not simultaneously $\overline{D}$-stabilizable.

Proof. Recall that $I = [-1, 1]$. The plants have a common point of intersection at $z = 0$, since $p_i(0) = 0$, $i = 1, \ldots, 4$. It is easy to check that for any k they are pairwise stabilizable and hence, applying Theorem 4.14, they are simultaneously I-stabilizable. It remains to be shown that for $k > e^{26}$ they are not simultaneously $\overline{D}$-stabilizable. Suppose, by contradiction, that for some $k > e^{26}$ the plants are simultaneously $\overline{D}$-stabilizable. Then for this k, and by using the natural coprime decomposition of $p_i(z)$, there must exist $n_c, d_c \in S(\overline{D})$ such that $d_c \in U(\overline{D}), zn_c + (z+2)d_c \in U(\overline{D}), 2zn_c + (z+2)d_c \in U(\overline{D})$, and $2zn_c + (2-kz)(z+2)d_c \in U(\overline{D})$. We define $f = 2zn_c/d_c(z+2) + 2 \in S(\overline{D})$. By the above equations it is then clear that $f \in U(\overline{D}), f - 1 \in U(\overline{D})$, and $f - kz \in U(\overline{D})$. The first two equations imply that $f(z) \neq 0$ and $f(z) \neq 1$ for every $z \in D$. In addition to this, $f(z)$ is analytic in D and $f(0) = 2$. By applying Picard–Schottky's theorem ([2, p. 19]) we have that $\mid f(z) \mid \leq e^{24}$ for every $\mid z \mid \leq \frac{1}{2}$. But then $\mid f(z) \mid < k \mid z \mid$ for $\mid z \mid = \frac{1}{2}$. This last inequality implies by Rouché's theorem [22] that $f - kz$ has a zero in $\{z : \mid z \mid \leq \frac{1}{2}\}$. This leads to a contradiction, since $f - kz \in U(\overline{D})$, and thus the theorem is proved. □

We provide the counterpart for continuous-time stability by using the conformal mapping.

COROLLARY 5.2. *Let $p_1(s) = 0, p_2(s) = (s-1)/(s+1)$, $p_3(s) = 2(s-1)/(s+1)$, $p_4(s) = 2(s-1)/((2-k)s + (2+k))$ be four continuous time systems. If $k > e^{26}$ then $p_i(s)$, $i = 1, \ldots, 4$ are simultaneously $\mathbb{R}_{+\infty}$-stabilizable but they are not simultaneously $\mathbb{C}_{+\infty}$-stabilizable.*

Proof. The four plants are simultaneously $\mathbb{C}_{+\infty}$-stabilizable if and only if $p_1(z) = 0$, $p_2(z) = z$, $p_3(z) = 2z$ and $p_4(z) = 2z/(2-kz)$ are simultaneously $\overline{D}$-stabilizable. This, in turn, implies that the four plants are simultaneously $\mathbb{C}_{+\infty}$-stabilizable if and only if $p_1(z) = 0$, $p_2(z) = z/(z+2)$, $p_3(z) = 2z/(z+2)$, and $p_4(z) = 2z/((z+2)(2-kz))$ are simultaneously $\overline{D}$-stabilizable. The impossibility of this is proved in Theorem 5.1. □

The next counterexample is slightly stronger. It applies to the case of three plants. This result also answers negatively the question addressed in the conclusion of [30].

THEOREM 5.3. *Let n be a positive integer, and let $p_{1,n}(z) = 0$, $p_{2,n}(z) = nz/(z+2)$, and $p_{3,n}(z) = -1/(nz(z+2))$ be three discrete time plants. For every n, $p_{i,n}(z)$, $i = 1, 2, 3$ are simultaneously I-stabilizable. There exists, however, an n such that $p_{i,n}(z)$, $i = 1, 2, 3$ are not simultaneously $\overline{D}$-stabilizable.*

Proof. It can be checked that for any positive integer n these three plants are simultaneously I-stabilizable; this part is left to the reader (the result follows from Theorem 4.13). The fact that they are not simultaneously $\overline{D}$-stabilizable for all n is more difficult to prove. We suppose in the sequel that for every n they are simultaneously $\overline{D}$-stabilizable and we produce a contradiction.

Notice first, since $(z+2) \in U(\overline{D})$ that $p_{i,n}(z)$ are simultaneously $\overline{D}$-stabilizable for every integer n if and only if $p'_{1,n}(z) = 0$, $p'_{2,n}(z) = nz$, and $p'_{3,n}(z) = -1/nz$ are simultaneously $\overline{D}$-stabilizable for every n. This in turn is possible if and only if for each n there exist $n_{c,n}(z), d_{c,n}(z) \in S(\overline{D})$ such that $d_{c,n}(z) \in U(\overline{D})$, $n_{c,n}(z)nz + d_{c,n}(z) \in U(\overline{D})$, and $n_{c,n}(z) - d_{c,n}(z)nz \in U(\overline{D})$. Since $d_{c,n}(z) \in U(\overline{D})$, we may define $h_n(z) \triangleq n_{c,n}(z)/d_{c,n}(z) \in S(\overline{D})$ to be the solution associated to n. We then have that $h_n(z)nz + 1 \in U(\overline{D})$ and $h_n(z) - nz \in U(\overline{D})$ for every n. In the next part we show that the existence, for every n, of a simultaneous solution $h_n(z)$ to these two equations is impossible.

Since $h_n(z)nz+1 \in U(\overline{D})$, we can define $g_n(z) = (h_n(z)nz-n^2z^2)/(h_n(z)nz+1) \in S(\overline{D})$. These functions are analytic in D, they have no zeros in $D\backslash\{0\}$ and they take the value 1 only twice in D, namely, at $z = j/n$ and $z = -j/n$. By the generalised form of Montel's normal family criterion ([16, p. 70]) this implies that the sequence $(g_n(z))$ is a normal family in $D\backslash\{0\}$. Hence, going to a subsequence, we can assume that $g_n(z)$ converges uniformly on compact subsets of $D\backslash\{0\}$. There are only two possible cases: either $g_n(z)$ tends locally uniformly to infinity, or $g_n(z)$ tends locally uniformly to an analytic function in $D\backslash\{0\}$. We show in what follows that both these cases lead to a contradiction.

Case 1. $g_n(z)$ tends locally uniformly to infinity, i.e., the functions $1/g_n(z)$ tend locally to zero on every compact set of $D\backslash\{0\}$. Consider the compact set $\{z : |z| = \frac{1}{2}\}$. Given $\epsilon > 0$, we have $|1/g_n(z)| \leq \epsilon/2 = \epsilon|z|$ for every $n \geq n_0(\epsilon)$ and $|z| = 1/2$. By definition of $g_n(z)$ we know that $nzh_n(z)(1-1/g_n(z)) = -(1+n^2z^2/g_n(z))$. Using this equality together with the bounds obtained above we get $|h_n(z)/n| \leq (\epsilon/8)/\frac{1}{2}(1-\frac{1}{4})$ for $n \geq n_0(\epsilon) + n_0(\frac{1}{2})$ and $\{z :| z |= \frac{1}{2}\}$. For some large integer n we thus have $|h_n(z)/n| < \frac{1}{2}$ when $|z| = \frac{1}{2}$, i.e., $|h_n(z)/n| < |z|$ when $|z| = \frac{1}{2}$. The functions $h_n(z)/n$ are analytic in $\{z : |z| \leq \frac{1}{2}\}$ and hence, by Rouché's theorem, $h_n(z)/n - z$ has a zero in $\{z :| z |\leq \frac{1}{2}\}$ for some integer n. But this contradicts the fact that $h_n(z) - nz \in U(\overline{D})$ and thus Case 1 cannot occur.

Case 2. $g_n(z)$ tends locally uniformly to an analytic function in $D\backslash\{0\}$. Then $g_n(z)$ are uniformly bounded on compact subsets of $D\backslash\{0\}$. Say, $|g_n(z)| \leq M$ for $|z| = \frac{1}{2}$. We have defined $g_n(z) = (h_n(z)nz - n^2z^2)/(h_n(z)nz + 1)$ and thus also $g_n(z) = 1 - (1 + n^2z^2)/(h_n(z)nz + 1)$. This last equation, together with the bound on $g_n(z)$, implies that $|(1 + n^2z^2)/(h_n(z)nz + 1)| \leq M + 1$ for $|z| = \frac{1}{2}$. This in turn implies that $|n^2/(h_n(z)nz + 1)| \leq (M + 1)/(\frac{1}{4} - 1/n^2)$ for $|z| = \frac{1}{2}$ and $n > 3$. The function $n^2/(h_n(z)nz + 1)$ is analytic in D and hence, by the Maximum Modulus Theorem, the bound obtained above holds throughout the disc of radius $\frac{1}{2}$. In particular, it holds at $z = 0$ so that we must have $n^2 \leq (M + 1)/(\frac{1}{4} - 1/n^2)$ for $n > 3$. But this inequality is obviously violated when $n > 2\sqrt{M + 2}$. A contradiction is obtained and thus Case 2 cannot occur. □

We end this paper by providing an example of a plant that has the even interlacing property but that is not $\overline{D}$-stabilizable by a unit controller. Recall that in §4.2 we established that a plant $p(z)$ is I-stabilizable by a unit controller if and only if $p(z)$ has the even interlacing property on I.

THEOREM 5.4. *Let $p_n(z) = z/(1 + n^2z^2)$. $p_n(z)$ has the even interlacing property for every positive integer n. There exists, however, an n such that $p_n(z)$ is not unit $\overline{D}$-stabilizable.*

Proof. Suppose, by contradiction, that for every integer n there exists a unit $\overline{D}$-stabilizer of $p_n(z) = z/(1 + n^2z^2)$. Then, for every positive integer n, there exist $n_{c,n}, d_{c,n} \in U(\overline{D})$ such that $zn_{c,n} + (1 + n^2z^2)d_{c,n} = u_n \in U(\overline{D})$. Since $n_{c,n}, d_{c,n} \in U(\overline{D})$ this implies that $u_n/d_{c,n} = z(n_{c,n}/d_{c,n}) + (1 + n^2z^2) = nz((n_{c,n}/nd_{c,n}) + nz) + 1 \in U(\overline{D})$. Define $h_n = (n_{c,n}/nd_{c,n}) + nz$; then, for every n, h_n defined above is such that $h_nnz + 1 \in U(\overline{D})$ and $h_n - nz \in U(\overline{D})$. This has been proved to be impossible in the proof of Theorem 5.3 and thus the theorem is proved. □

6. Conclusion. In this paper we have analysed some aspects of the simultaneous stabilization question.

Our first contribution was to show that the problem of internal stabilization of k plants p_i, $i = 1, \ldots, k$ is equivalent to what we have called an avoidance problem: "Under what condition on $p_i(s)$, $i = 1, \ldots, k$ is it possible to find $q(s)$ such that

$p_i(s) \neq q(s)$, $\forall s \in \mathbb{C}_{+\infty}$, $i = 1, \ldots, k$?" Our first message is clear: *stabilization = avoidance.* This restatement of the problem does not answer any question, but provides new insights and new proof techniques for the establishment of other results.

The second part dealt with a subproblem of the simultaneous stabilization problem. Given two plants p_1 and p_2, we showed that there exists a controller c such that the closed-loop transfer functions associated with p_1 and p_2 are stable if and only if there exist a controller c such that the closed-loop transfer functions associated to p_1 and p_2 have no *real* unstable poles. The same property is proved for the strong stabilization problem. Motivated by these results, we have developed in that part a complete answer to the question: "Given k plants p_i, $i = 1, \ldots, k$ when is it possible to find a single controller c such that all the transfer functions have no real unstable poles?" Although such a question may seem to be of limited practical interest, we have given some motivations for it.

The third part gave answers to some of the questions raised in part two and elsewhere. In particular, we showed that, unlike the case of two plants, the existence of a simultaneous stabilizing controller for more than two plants cannot be guaranteed by the existence of a controller such that the closed-loop transfer functions have no real unstable poles.

To conclude, let us stress the fact that our results provide a much better understanding of the original simultaneous stabilization problem for more than two plants but that the problem is...still unanswered.

Acknowledgments. We wish to thank F. Callier, P. Delsarte, C. Hollot, and two anonymous reviewers for their comments on a first version of this paper.

Note added in proof. Since the submission of this paper, research on simultaneous stabilization has progressed rapidly. Many of the theorems presented here have been extended and some of the proofs have been simplified. In particular, the 3-interlacing condition of §4 and all proofs §5 are given in a new simplified form in V. Blondel, *Simultaneous Stabilization of Linear Systems*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1993.

REFERENCES

[1] B. D. O. Anderson, *A note on the Youla–Bongiorno–Lu Condition*, Automatica, 12 (1976), pp. 387–388.

[2] L. Ahlfors, *Conformal invariants*, McGraw-Hill Series in Higher Mathematics, McGraw-Hill, New York, 1973.

[3] A. Alos, *Stabilization of a class of plants with possible loss of outputs or actuator failures*, IEEE Trans. Automat. Control, 28 (1983), pp. 231–233.

[4] V. Blondel, *A Problem from Control Theory, Solvability of Equations over an Euclidean Domain*, M.Sc. thesis, Dept. of Mathematics, Imperial College, London, U.K., 1990.

[5] ———, *A counterexample to a simultaneous stabilization condition for systems with identical unstable poles and zeros*, Systems Control Lett., 17 (1991), pp. 339–341.

[6] V. Blondel, G. Campion, and M. Gevers, *A Sufficient Condition for Simultaneous Stabilization*, IEEE Trans. Automat. Control, 38 (1993), pp. 1264–1266.

[7] V. Blondel, *Simultaneous Stabilization*, Ph.D. thesis, Dept. of Electrical Engineering, Université Catholique de Louvain, Louvain, Belgium, 1992.

[8] P. Dorato, H. Park, and Y. Li, *An algorithm for interpolation with units in H^∞, with applications to feedback stabilization*, Automatica, 25 (1989), pp. 427–430.

[9] E. Emre, *Simultaneous stabilization with fixed closed loop characteristic polynomial*, IEEE Trans. Automat. Control, 28 (1983), pp. 103–104.

[10] G. Gelfond, *The Solution of Equation with Integer*, Golden Gate Books, San Francisco, 1961.

[11] B. Ghosh and C. Byrnes, *Simultaneous stabilization and pole-placement by nonswitching dynamic compensation*, Trans. Automat. Control, 28 (1983), pp. 735–741.

[12] B. GHOSH, *Some new results on the simultaneous stabilizability of a family of single input single output systems*, Systems Control Lett., 6 (1985), pp. 39–45.
[13] ———, *Transcendental and interpolation methods in simultaneous stabilization and simultaneous partial pole placement problems*, SIAM J. Control Optim., 24 (1986), pp. 1091–1109.
[14] ———, *An approach to simultaneous system design. Part* 1, SIAM J. Control Optim., 24 (1986), pp. 480–496.
[15] ———, *An approach to simultaneous system design. Part* 2, SIAM J. Control Optim., 26 (1988), pp. 919–963.
[16] G. GOLUZIN, *Geometric theory of functions of a complex variable*, Trans. Math. Monographs, American Math. Soc., 26 (1969).
[17] J. HELTON, *Worst case analysis in the frequency domain*, IEEE Trans. Automat. Control, 30 (1985), pp. 1154–1170.
[18] P. KABAMBA AND C. YANG, *Simultaneous controller design for linear time invariant systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 106–111.
[19] P. KHARGONEKAR AND A. TANNENBAUM, *Non-Euclidean metrics and the robust stabilization of systems with parametric uncertainty*, IEEE Trans. Automat. Contr., 30 (1985), pp. 1005–1013.
[20] H. KWAKERNAAK, *A condition for robust stabilizability*, Systems Control Lett., 2 (1985), pp. 1005–1013.
[21] H. MAEDA AND M. VIDYASAGAR, *Some results on simultaneous stabilization*, Systems Control Lett., 5 (1984), pp. 205–208.
[22] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1986.
[23] R. SAEKS AND J. MURRAY, *Fractional representation, algebraic geometry and the simultaneous stabilization problem*, IEEE Trans. Automat. Control, 27 (1982), pp. 895–903.
[24] M. VIDYASAGAR, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[25] ———, *Some results on simultaneous stabilization with multiple domains of stability*, Automatica, 23 (1987), pp. 535–540.
[26] M. VIDYASAGAR AND N. VISWANADHAM, *Algebraic design techniques for reliable stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 1085–1095.
[27] K. WEI AND B. BARMISH, *An iterative design procedure for simultaneous stabilization of* MIMO *systems*, Automatica, 24 (1988), pp. 643–652.
[28] K. WEI, *Simultaneous pole assignment for a class of linear time invariant siso systems*, in Proceedings of the 28th Conference on Decision and Control, Tampa, FL, pp. 1247–1252, 1989.
[29] ———, *Stabilization of a linear plant via a stable compensator having no real unstable zeros*, Systems and Control Lett., 15 (1990), pp. 259–264.
[30] ———, *The Solution of a Transcendental Problem and Its Application in Simultaneous Stabilization Problems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1305–1315.
[31] ———, *Solvability of a transcendental problem in system theory*, in Proceedings of the 30th Conference on Decision and Control, Brighton, UK, pp. 1933–1938, 1991.
[32] D. YOULA, J. BONGIORNO, AND C. LU, *Single-loop feedback stabilization of linear multivariable plants*, Automatica, 10 (1974), pp. 159–173.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 591–611, May 1994

ON GENERALIZED SECOND-ORDER DERIVATIVES AND TAYLOR EXPANSIONS IN NONSMOOTH OPTIMIZATION*

W. L. CHAN†, L. R. HUANG†, AND K. F. NG†

Abstract. A representation of Cominetti and Correa's generalized second-order directional derivative [SIAM J. Control Optim., 28 (1990), pp. 789–809] is given and then applied to obtain a Taylor theorem type result. A conjecture in [*SIAM J. Control Optim.*, 28 (1990), pp. 789–809] concerning functions of the form $\max_{1\le i\le n} g_i(x)$ is proved under a strengthened assumption, but not true otherwise.

Key words. Dini-directional derivative, Clarke's directional derivative, generalized second-order directional derivative, Taylor expansion, nonsmooth analysis

AMS subject classifications. 26A27, 26E15, 49A52, 49B27, 49D37

Introduction. Since the pioneering works of F. Clarke and B. N. Pshenichnyi, generalized directional derivatives have been studied, and successfully applied in various fields (e.g., [1]–[8], [12]–[18]), especially in optimization and control theory. The study of generalized second-order directional derivatives is more recent; one of such derivatives is defined in Cominetti and Correa [2] by

$$f^{\infty}(x;u,v) := \lim \sup_{\substack{y\to x\\ s,t\downarrow 0}} \frac{1}{ts}\{f(y+tu+sv) - f(y+tu) - f(y+sv) + f(y)\}.$$

On the basis of their work, in Proposition 1.4 we represent $f^{\infty}(x;u,v)$ in the form of the upper limit of the rates of changes of the Dini-directional derivatives. This representation enables us to establish second-order Taylor expansions (Theorems 3.2 and 3.3) for nonsmooth functions. These extend the corresponding results of Cominetti and Correa who assumed the C^1-condition. In §5 we apply our results to a large class of functions (e.g., convex and concave functions) that are not covered by [2], Prop. 4.1. Applications to optimization theory are presented in §6.

In [2], a conjecture was made about the possible validity of $h^{\infty}(x;u,v) = \max_{1\le i\le n} D^2 g_i(x;u,v)$, where each g_i is C^2 and $D^2 g_i$ denotes the second-order directional derivative. Example 2.3 shows that the conjecture is incorrect and an affirmative answer is given in Corollary 2.5 and Corollary 2.7 under similar strengthened conditions.

1. Dini-directional derivatives, Clarke's directional derivatives, and generalized second-order directional derivatives. Let X be a locally convex space and $f : X \to \mathbb{R}$ a function. We consider the extended real field $\bar{\mathbb{R}} = \mathbb{R}\cup\{-\infty, +\infty\}$ with the usual operations, order, and topology familiar in convex analysis. Denote the upper and lower Dini-directional derivatives by

$$D^+ f(x;v) := \lim \sup_{t\downarrow 0} \frac{1}{t}(f(x+tv) - f(x)),$$

$$D_+ f(x;v) := \lim \inf_{t\downarrow 0} \frac{1}{t}(f(x+tv) - f(x)),$$

and the upper and lower Clarke's directional derivatives at x along the direction $v \in X$ by

$$f^0(x;v) := \lim \sup_{\substack{y\to x\\ t\downarrow 0}} \frac{1}{t}(f(y+tv) - f(y))$$

* Received by the editors March 12, 1992; accepted for publication (in revised form) September 21, 1992.
† Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

and

$$f_0(x;v) := \lim \inf_{\substack{y\to x\\ t\downarrow 0}} \frac{1}{t}(f(y+tv)-f(y)).$$

If $X=\mathbb{R}$ and $v=1$, we shall write $D^+f(x)$ for $D^+f(x;v)$ and similarly for $D_+f(x)$, $f^0(x)$. We shall often make use of the elementary computation rules for lim sup and lim inf without further comments, e.g., if $f=f_1-f_2$, then

$$D_+f(x)\le D_+f_1(x)-D_+f_2(x),$$

provided that the two terms on the right are finite. Also $D_+f(x)\le D^+f_1(x)-D^+f_2(x)$ with similar provisions (see, for instance, [17], p. 108).

Furthermore, as in [2], [4], and [5] we define the upper and lower generalized second-order directional derivatives at x in the direction $(u,v)\in X\times X$ by

$$f^\infty(x;u,v) := \lim \sup_{\substack{y\to x\\ t,s\downarrow 0}} \frac{1}{st}\{f(y+tu+sv)-f(y+tu)-f(y+sv)+f(y)\}$$

and

$$f_\infty(x;u,v) := \lim \inf_{\substack{y\to x\\ t,s\downarrow 0}} \frac{1}{st}\{f(y+tu+sv)-f(y+tu)-f(y+sv)+f(y)\}.$$

For the sake of convenience, we list some of its properties in the following proposition. For more detailed properties of f^∞ we refer to [2], [4].

PROPOSITION 1.1 [2]. *Let $f: X\to\mathbb{R}$ and $x\in X$. Then*

(i) *The map $(u,v)\mapsto f^\infty(x;u,v)$ is symmetric and sublinear on each variable separately.*

(ii) *The map $y\mapsto f^\infty(y;u,v)$ is upper semi-continuous at x for every $(u,v)\in X\times X$.*

(iii) $f^\infty(x;u,-v)=f^\infty(x;-u,v)=(-f)^\infty(x;u,v)=-f_\infty(x;u,v)$.

Before studying the relationships between the above directional derivatives, we give a few lemmas which will often be used in the sequel. Lemma 1.2 has appeared in [2], Lems. 1.4 and 1.5.

LEMMA 1.2. *Let $f: X\to\mathbb{R}$ be a continuous function, $x,v\in X$, and $t>0$. Then there exists $\alpha\in(0,t)$ such that*

$$\frac{f(x+tv)-f(x)}{t}\le D_+f(x+\alpha v;v).$$

Consequently,

$$\lim \sup_{y\to x} D_+f(y;v)=\lim \sup_{y\to x} D^+f(y;v)=\lim \sup_{y\to x} f^0(y;v)=f^0(x;v).$$

Remark. If let $f=-g$, then we have

$$\frac{g(x+tv)-g(x)}{t}\ge D^+g(x+\alpha v;v),$$

and so the corresponding results for $f_0(x;v)$.

From Lemma 1.2 we have the following.

LEMMA 1.3. *Suppose that* $f : X \to \mathbb{R}$ *is continuous and* $x, u, v \in X$. *Then for any* $t_0 > 0, t \in (0, t_0)$ *and* $s \in \mathbb{R}$ *there exists* $\alpha \in (0, t)$ *such that*

$$\begin{aligned} &\frac{1}{t}[f(x+sv+tu) - f(x+sv) - f(x+tu) + f(x)] \\ &\le D^+f(x+\alpha u+sv;u) - D^+f(x+\alpha u;u) \end{aligned} \tag{1.1}$$

and

$$\begin{aligned} &\frac{1}{t}[f(x+sv+tu) - f(x+sv) - f(x+tu) + f(x)] \\ &\le D_+f(x+\alpha u+sv;u) - D_+f(x+\alpha u;u) \end{aligned} \tag{1.2}$$

if $D^+f(\cdot\,;u)$ *and* $D_+f(\cdot\,;u)$ *are finite on the segments* $(x, x+t_0u)$ *and* $(x+sv, x+sv+t_0u)$.

Remark. If we let $f = -g$, then we have

$$\begin{aligned} &\frac{1}{t}[g(x+sv+tu) - g(x+sv) - g(x+tu) + g(x)] \\ &\ge D_+g(x+\alpha u+sv;u) - D_+g(x+\alpha u;u) \end{aligned} \tag{1.1$'$}$$

and

$$\begin{aligned} &\frac{1}{t}[g(x+sv+tu) - g(x+sv) - g(x+tu) + g(x)] \\ &\ge D^+g(x+\alpha u+sv;u) - D^+g(x+\alpha u;u). \end{aligned} \tag{1.2$'$}$$

Proof. Let us fix an arbitrary $s \in \mathbb{R}$ and denote the left number of (1.1) by

$$\frac{\Phi(t) - \Phi(0)}{t}$$

where $\Phi(t) := f(x+sv+tu) - f(x+tu) = \Phi_1(t) - \Phi_2(t)$ with the obvious meaning of Φ_1, Φ_2. If $t_0 > 0$ and $t \in (0, t_0)$, then by Lemma 1.2, there exists an $\alpha \in (0, t)$ such that

$$\frac{\Phi(t) - \Phi(0)}{t} \le D_+\Phi(\alpha),$$

where $D_+\Phi(\alpha)$ denotes $D_+\Phi(\alpha; 1)$ for short. By assumption $D^+\Phi_1(\alpha), D^+\Phi_2(\alpha)$ are finite, and it follows that

$$\begin{aligned} \frac{\Phi(t) - \Phi(0)}{t} &\le D_+\Phi(\alpha) \le D^+\Phi_1(\alpha) - D^+\Phi_2(\alpha) \\ &= D^+f(x+\alpha u+sv;u) - D^+f(x+\alpha u;u). \end{aligned}$$

This proves (1.1), and similarly one can prove (1.2) because

$$D_+\Phi(\alpha) \le D_+\Phi_1(\alpha) - D_+\phi_2(\alpha)$$

as the two terms on the right are finite. □

Recall that f is regular at x [1] if the one-sided directional derivative

$$f'(x;v) = \lim_{t \downarrow 0} \frac{1}{t}(f(x+tv) - f(x)),$$

exists and $f'(x;v) = f^0(x;v)$ for all v.

PROPOSITION 1.4. *Let $f : X \to R$ be a continuous function. Let $x, u, v \in X$ and suppose that $f^0(\cdot\,; u), D^+f(\cdot\,; u)$ and $D_+f(\cdot\,; u)$ are finite near x. Then one has*

$$\begin{aligned}&(f^0(\cdot\,; u))^0(x; v)\\ &\le f^\infty(x; u, v) = (D^+f(\cdot\,; u))^0(x; v) = (D_+f(\cdot\,; u))^0(x; v);\end{aligned} \tag{1.3}$$

that is,

$$\begin{aligned}&\limsup_{\substack{y\to x\\ s\downarrow 0}} \frac{1}{s}(f^0(y+sv; u) - f^0(y; u))\\ &\le f^\infty(x; u, v) = \limsup_{\substack{y\to x\\ s\downarrow 0}} \frac{1}{s}(D_+f(y+sv; u) - D_+f(y; u))\\ &= \limsup_{\substack{y\to x\\ s\downarrow 0}} \frac{1}{s}(D^+f(y+sv; u) - D^+f(y; u)).\end{aligned} \tag{1.4}$$

Dually one also has

$$\begin{aligned}&(f_0(\cdot\,; u))_0(x; v)\\ &\ge f_\infty(x; u, v) = (D_+f(\cdot\,; u))_0(x; v) = (D^+f(\cdot\,; u))_0(x; v)\end{aligned} \tag{1.5}$$

if $f_0(\cdot\,; u), D_+f(\cdot\,; u)$, and $D^+f(\cdot\,; u)$ are finite near x.

Furthermore, if f is regular near x, then the inequality in (1.3) *becomes an equality.*

Proof. We need only prove (1.3) as (1.5) will then follow by considering $-f = g$ (the assertion for the regular case is evident from (1.4) because then $D_+f(y+sv; u) = f^0(y+sv; u)$ for all y near x and small v). By Lemma 1.2 we have

$$\limsup_{z\to y} D^+f(z+sv; u) = f^0(y+sv; u).$$

Thus, since $f^0(\cdot\,; u)$ is finite near x, it follows from the subadditivity of lim sup that

$$f^0(y+sv; u) - f^0(y; u) \le \limsup_{z\to y}(D^+f(z+sv; u) - D^+f(z; u)).$$

This implies that

$$\begin{aligned}&(f^0(\cdot\,; u))^0(x, v)\\ &= \limsup_{\substack{y\to x\\ s\downarrow 0}} \frac{1}{s}(f^0(y+sv; u) - f^0(y; u))\\ &\le \limsup_{\substack{y\to x\\ s\downarrow 0}} \limsup_{z\to y} \frac{1}{s}(D^+f(z+sv; u) - D^+f(z; u))\\ &\le \limsup_{\substack{y\to x\\ s\downarrow 0}} \frac{1}{s}(D^+f(y+sv; u) - D^+f(y; u))\\ &= (D^+f(\cdot\,; u))^0(x; v),\end{aligned} \tag{1.6}$$

showing the inequality in (1.3).

On the other hand, since $D^+f(\cdot\,; u)$ and $D_+f(\cdot\,; u)$ are finite near x, one has, by the subadditivity of lim sup,

$$\begin{aligned}&D^+f(y+sv; u) - D^+f(y; u)\\ &\le \limsup_{t\downarrow 0} \frac{1}{t}[f(y+sv+tu) - f(y+sv) - f(y+tu) + f(y)]\end{aligned}$$

and also

$$\begin{aligned} &D_+f(y+sv;u)-D_+f(y;u)\\ &\le \limsup_{t\downarrow 0}\frac{1}{t}[f(y+sv+tu)-f(y+sv)-f(y+tu)+f(y)]. \end{aligned}$$

These imply that

$$\begin{aligned} &(D^+f(\cdot\,;u))^0(x;v)\\ &=\limsup_{\substack{y\to x\\ s\downarrow 0}}\frac{1}{s}(D^+f(y+sv;u)-D^+f(y;u))\\ &\le\limsup_{\substack{y\to x\\ s,t\downarrow 0}}\frac{1}{st}[f(y+sv+tu)-f(y+sv)-f(y+tv)+f(y)]\\ &=f^\infty(x;u,v) \end{aligned} \tag{1.7}$$

and, similarly,

$$\begin{aligned} &(D_+f(\cdot\,;u))^0(x;v)\\ &=\limsup_{\substack{y\to x\\ s\downarrow 0}}\frac{1}{s}(D_+f(y+sv;u)-D_+f(y;u))\\ &\le\limsup_{\substack{y\to x\\ t,s\downarrow 0}}\frac{1}{ts}[f(y+sv+tu)-f(y+sv)-f(y+tu)+f(y)]\\ &=f^\infty(x;u,v). \end{aligned} \tag{1.8}$$

By definition and (1.2) of Lemma 1.3,

$$\begin{aligned} &f^\infty(x;u,v)\\ &=\limsup_{\substack{y\to x\\ t,s\downarrow 0}}\frac{1}{ts}[f(y+sv+tu)-f(y+sv)-f(y+tu)+f(y)]\\ &\le\limsup_{\substack{y\to x\\ t,s\downarrow 0}}\frac{1}{s}(D_+f(y+\alpha u+sv;u)-D_+f(y+\alpha u;u)] \qquad \alpha\in(0,t)\\ &=\limsup_{\substack{y\to x\\ s\downarrow 0}}\frac{1}{s}(D_+f(y+sv;u)-D_+f(y;u))\\ &=(D_+f(\cdot\,;u))^0(x;v), \end{aligned} \tag{1.9}$$

where we have written α for $\alpha=\alpha(y,s,u,v)$ for the sake of simplicity in notations. Similarly,

$$f^\infty(x;u,v)\le(D^+f(\cdot\,;u))^0(x;v). \tag{1.10}$$

Together with (1.6), (1.7), (1.8), and (1.9), we have (1.3). □

Remark. There are examples of Lipschitz functions on an interval say $[0,b]$ that fail to be right-differentiable at infinitely many points near 0. Thus the representation given in the preceding proposition is valid, but cannot be expressed in the form of (3) in [2], Prop. 1.3. For example, write

$$(0,1/2\pi]=\bigcup_{k=1}^{\infty}[x_{k+1},x_k],\qquad x_k=1/2k\pi.$$

Define $f(0)=0, f(x_k)=0$ and

$$f(x)=(x-x_{k+1})(x_k-x)\sin(x-x_{k+1})^{-1}$$

if $x \in (x_{k+1}, x_k)$. Then $f'_+(x)$ does not exist at each x_k.

In view of Proposition 1.4 we introduce the following generalized second-order directional derivative in line of Clarke's derivatives as an alternative to $f^\infty(x; u, v)$.

DEFINITION 1.5. *Let $f : X \to \mathbb{R}, x, v \in X$, and suppose that $f^0(x; v)$ and $f_0(\cdot\,; v)$ are finite near x. Then the upper and lower generalized second-order directional derivatives are defined, respectively, by*

$$f^{00}(x; u, v) := \lim \sup_{\substack{y \to x \\ t \downarrow 0}} \frac{1}{t}(f^0(y + tu; v) - f^0(y; v))$$

and

$$f_{00}(x; u, v) := \lim \inf_{\substack{y \to x \\ t \downarrow 0}} \frac{1}{t}(f_0(y + tu; v) - f_0(y; v)).$$

It is easy to see that the function $u \mapsto f^{00}(x; u, v)$ is sublinear and the function $x \mapsto f^{00}(x; u, v)$ is upper semi-continuous. Furthermore, if f is continuous and $f^0(x; v)$, $f_0(x; v)$ are finite near x, then from the above proposition we have

$$f_\infty(x; u, v) \le f_{00}(x; u, v) \le f^{00}(x; u, v) \le f^\infty(x; u, v);$$

and $f^{00}(x; u, v) = f^\infty(x; u, v)$ if f is regular near x.

In §6 we shall give applications of f^∞ and f^{00} in the second-order necessary optimality condition for constrained problem.

2. On Cominetti and Correa's conjecture. In this section we study second-order directional derivative of the function h of the form

$$h(x) = \max\{g_1(x), g_2(x), \ldots, g_n(x)\} \qquad (x \in X)$$

where each g_i is a real-valued function on X. Note that $h = f \circ g$ if one writes $g = (g_1, g_2, \ldots, g_n)$ and defines

$$f(a) = \max_{i \in I}\{a_i\}, \quad \text{for any } a = (a_1, \ldots, a_n) \in \mathbb{R}^n$$

where $I := \{1, 2, \ldots, n\}$. Let $I(a)$ denote the subset of I consisting of all i for which $f(a) = a_i$.

For $x, u, v \in X$ we shall write $H(x; u, v) \underline{\propto} 0$ to denote the following condition:

$$(g'_i(x; u) - g'_j(x; u))(g'_i(x; v) - g'_j(x; v)) \le 0$$

for all $i, j \in I(g(x))$, and $H(x; u, v) \propto 0$ to denote the condition that the strict inequality holds for all distinct $i, j \in I(g(x))$.

Suppose each g_i is a C^2-function with the usual second-order directional derivative at x with respect to the directions u, v denoted by $D^2 g_i(x; u, v)$. Cominetti and Correa conjectured in [2] that if $\{g'_i(x); i \in I(g(x))\}$ is affinely independent, then the following formula holds

$$h^\infty(x; u, v) = \max_{i \in I(g(x))} D^2 g_i(x; u, v) \tag{2.1}$$

if $H(x; u, v) \underline{\propto} 0$. This is incorrect as shown by Example 2.3 below, but true if the condition is strengthened to $H(x; u, v) \propto 0$ (Corollary 2.5).

In the following we first consider a property related to the set $I(a)$.

LEMMA 2.1. *Suppose that X is a locally convex space and g is arbitrary continuous function of X into $\mathbb{R}^n$ denoted by $g = (g_1, \ldots, g_n)$. Then for the above f and any $x \in X$, there exists a neighborhood W of x such that*

$$I(g(y)) \subseteq I(g(x))$$

for all $y \in W$.

Proof. We fix $i \in I(g(x))$. Then, for each $j \in I \backslash I(g(x))$,

$$g_j(\cdot) < g_i(\cdot)$$

at x and hence on a neighborhood W_j of x. Do this for each such j and let W denote the intersection of W_j's. Then W has the required property: if $y \in W$ and $j \notin I(g(x))$, then $g_j(y) < g_i(y)$ showing that $j \notin I(g(y))$. □

LEMMA 2.2. *Suppose that $g'(\cdot) = (g_1'(\cdot), \ldots, g_n'(\cdot))$ is continuous near x and $g_1'(x), \ldots, g_n'(x)$ are affinely independent. Then either there exists a neighborhood W of x such that the following condition $H(y; u, v) \underline{\propto} 0$ holds for each $y \in W$:*

$$(g_i'(y; v) - g_j'(y; v))(g_i'(y; u) - g_j'(y; u)) \leq 0, \quad \forall i, j \in I(g(y))$$

or

$$h^{\infty}(x; u, v) = +\infty.$$

Proof. By the continuity of g' at x it is easy to show that there exists a neighborhood W_1 of x such that $g_1'(y), \ldots, g_n'(y)$ are affinely independent for all $y \in W_1$. Now if for each neighborhood U of x, there exists $y \in U \cap W_1$ so that the condition $H(y; u, v) \underline{\propto} 0$ is not satisfied, then by [2], Prop. 3.9, $h^{\infty}(y; u, v) = +\infty$. Hence by the upper semicontinuity of $h^{\infty}(\cdot\,; u, v)$,

$$h^{\infty}(x; u, v) = +\infty. \qquad \square$$

Example 2.3. Let $g = (g_1, g_2, g_3)$ with the C^2-functions $g_1(x, y, z) = \xi(x) + x + y, g_2(x, y, z) = x + 2y$ and $g_3(x, y, z) = 4(x - y) + z$ for all $x, y, z \in \mathbb{R}$, where

$$\xi(x) = \begin{cases} x^5 \sin \dfrac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Let $\bar{x} = (0, 0, 0)$. Since $g_1, g_2, g_3 = 0$ at $\bar{x}, I(g(\bar{x})) = \{1, 2, 3\}$. Further, $g_1'(\bar{x}) = (1, 1, 0), g_2'(\bar{x}) = (1, 2, 0)$, and $g_3'(\bar{x}) = (4, -4, 1)$. Thus, $g_1'(\bar{x}), g_2'(\bar{x})$, and $g_3'(\bar{x})$ are linearly independent. Let

$$u = (1, 0, 0) \quad \text{and} \quad v = (1, 1, 0).$$

Then

$$g_1'(\bar{x}; u) - g_2'(\bar{x}; u) = 1 - 1 = 0 \quad \text{and} \quad g_1'(\bar{x}; v) - g_2'(\bar{x}; v) = 2 - 3 < 0.$$

Similarly we can verify, for all other pairs of distinct i, j, that $g_1'(\bar{x}; u) - g_j'(\bar{x}; u)$ and $g_i'(\bar{x}; v) - g_j'(\bar{x}; v)$ are of opposite signs (or zero); that is, the Cominetti and Correa's condition $H(\bar{x}; u, v) \underline{\propto} 0$ is satisfied. But, in constrast to their conjecture, (2.1) does not hold. In fact, we will prove that

$$h^{\infty}(\bar{x}; u, v) = +\infty. \tag{2.2}$$

Let $P_n := (x_n, y_n, 0)$ with

$$x_n = -\frac{1}{2n\pi + \frac{\pi}{2}} \quad \text{and} \quad y_n = \frac{1}{\left(2n\pi + \frac{\pi}{2}\right)^5} = -x_n^5.$$

Then

$$g_1(P_n) = -x_n^5 + x_n + y_n = g_2(P_n).$$

This implies that $I(g(P_n)) = \{1, 2\}$ because $x_n \leq 2y_n$. Further,

$$g_1'(P_n) = \left(1 - 5\left(2n\pi + \frac{\pi}{2}\right)^{-4}, 1, 0\right)$$

and so

$$g_1'(P_n; u) - g_2'(P_n; u) = -5\left(2n\pi + \frac{\pi}{2}\right)^{-4} < 0$$

and

$$g_1'(P_n; v) - g_2'(P_n; v) = 2 - 5\left(2n\pi + \frac{\pi}{2}\right)^{-4} - 3 < 0.$$

Thus the condition $H(P_n; u, v) \underline{\propto} 0$ does not hold for all n. Since $P_n \to \bar{x}$, it follows from Lemma 2.2 that (2.2) must hold.

The above example actually shows that for $n \geq 2$ (if $n = 2$, we ignore g_3), the condition $H(x; u, v) \underline{\propto} 0$ is not sufficient for

$$h^\infty(x; u, v) = \max_{i \in I(g(X))} D^2 g_i(x; u, v).$$

We shall show however that the strengthened condition $H(x; u, v) \propto 0$ will be sufficient. Before our proof, we recall an elementary fact that if each g_i is directionally differentiable at x, then one has

$$h'(x; u) = \max_{i \in I(g(x))} g_i'(x; u) \tag{2.3}$$

for all $u \in X$.

PROPOSITION 2.4. *Suppose that each g_i is a C^1-function (that is, continuous Gâteaux differentiable function) at x. If for all $i, j \in I(g(x)), i \neq j$, one has*

$$[g_i'(x; u) - g_j'(x; u)][g_i'(x; v) - g_j'(x; v)] < 0, \tag{2.4}$$

then

$$h^\infty(x; u, v) \leq \max_{i \in I(g(X))} g_i^\infty(x; u, v).$$

Proof. By Proposition 1.4, we take a net $(z_\nu, \lambda_\nu)_\nu \in X \times \mathbb{R}_+$ written for short (z, λ) with $z \to x$ and $\lambda \downarrow 0$ such that

$$\begin{aligned} h^\infty(x; u, v) &= \lim_{\substack{z \to x \\ \lambda \downarrow 0}} \frac{1}{\lambda}(h'(z + \lambda u; v) - h'(z; v)) \\ &= \lim_{\substack{z \to x \\ \lambda \downarrow 0}} \frac{1}{\lambda}\left(\max_{i \in I(g(z+\lambda u))} g_i'(z + \lambda u; v) - \max_{i \in I(g(z))} g_i'(z; v)\right). \end{aligned}$$

In view of Lemma 2.1, we can assume that

$$I(g(z)), I(g(z+\lambda u)) \subseteq I(g(x)). \tag{2.5}$$

Since I is a finite set and considering a subnet if necessary we can assume without loss of generality that

$$h'(z;v) = \max_{i\in I(g(z))} g_i'(z;v) = g_1'(z;v) \quad \text{say,}$$

and

$$h'(z+\lambda u;v) = \max_{i\in I(g(z+\lambda u))} g_i'(z+\lambda u;v) = g_{i_0}'(z+\lambda u;v)$$

for some $i_0 \in I(g(z+\lambda u))$ and for all (z,λ).

We claim that there exists a subnet (z_s,λ_s) of (z,λ) such that

$$g_{i_0}'(z_s;v) - g_1'(z_s;v) \le 0.$$

In this case we will then obtain

$$\begin{aligned} h^\infty(x;u,v) &= \lim_s \frac{1}{\lambda_s}[g_{i_0}'(z_s+\lambda u;v) - g_{i_0}'(z_s;v) \\ &\qquad + g_{i_0}'(z_s;v) - g_1'(z_s;v)] \\ &\le \limsup_s \frac{1}{\lambda_s}[g_{i_0}'(z_s+\lambda u;v) - g_{i_0}'(z_s;v))] \\ &\le g_{i_0}^\infty(x;u,v). \end{aligned}$$

By (2.5), $i_0 \in I(g(x))$ and so we are done.

If our claim is false, then by considering a subnet if necessary, we assume that for all (z,λ)

$$g_{i_0}'(z;v) - g_1'(z;v) > 0 \tag{2.6}$$

where

$$i_0 \in I(g(z+\lambda u)) \quad \text{and} \quad 1 \in I(g(z)) \tag{2.7}$$

for all (z,λ). It follows that $g_{i_0}'(x;v) - g_1'(x;v) \ge 0$ and from (2.4) that the strict inequality must hold and

$$g_{i_0}'(x;u) - g_1'(x;u) < 0. \tag{2.8}$$

Since g_i is C^1, the formula (2.8) can be rewritten as

$$\lim_{\substack{y\to x \\ t\downarrow 0}} \frac{1}{t}[g_{i_0}(y+tu) - g_{i_0}(y) - g_1(y+tu) + g_1(y)] < 0$$

and so we can choose a neighborhood W of x and $\delta > 0$ such that

$$(g_{i_0}(y+tu) - g_{i_0}(y) - g_1(y+tu) + g_1(y)) < 0 \tag{2.9}$$

for all $y \in W$ and $0 < t < \delta$. Without loss of generality, we can assume that

$$(z,\lambda) \in W \times (0,\delta).$$

From (2.6) and the choice of 1, we see that $i_0 \notin I(g(z))$ and so $g_{i_0}(z) < g_1(z)$ for all (z, λ). Thus, together with (2.9) we conclude that

$$\begin{aligned} &g_{i_0}(z+\lambda u) - g_1(z+\lambda u) \\ &< g_{i_0}(z+\lambda u) - g_{i_0}(z) - g_1(z+\lambda u) + g_1(z) < 0. \end{aligned}$$

But this is impossible since

$$i_0 \in I(g(z+\lambda u)). \qquad \square$$

COROLLARY 2.5. *Suppose that each g_i is a C^2-function at $x, 1 \le i \le n$, and the derivatives $g_i'(x)$ are affinely independent. If for all $i, j \in I(g(x)), i \ne j$, one has*

$$(g_i'(x;u) - g_j'(x;u))(g_i'(x;v) - g_j'(x;v)) < 0,$$

then

$$h^\infty(x;u,v) = \max_{i \in I(g(X))} D^2 g_i(x;u,v).$$

Proof. Since each g_i is a C^2-function,

$$g_i^\infty(x;u,v) = D^2 g_i(x;u,v).$$

It follows from Proposition 2.4 that

$$h^\infty(x;u,v) \le \max_{i \in I(g(X))} D^2 g_i(x;u,v).$$

But the assumption on affinely independence ensures

$$h^\infty(x;u,v) \ge \max_{i \in I(g(X))} D^2 g_i(x;u,v)$$

[2], Prop. 3.7 and 3.8. $\square$

For normed spaces, we have another sufficient condition result for the similar representation of h^∞.

PROPOSITION 2.6. *Let X be a normed space, $u, x \in X$ and $g_i, 1 \le i \le n$, be C^2-functions at x. Suppose that W is a neighborhood of x such that for all $y \in W$ and $i, j \in I(g(y))$, one has $g_i'(y;u) = g_j'(y;u)$. Then*

$$h^\infty(x;u,u) = \max_{i \in I(g(x))} D^2 g_i(x;u,u).$$

Proof. Note first that since $g_i, 1 \le i \le n$, are C^2-functions, $g_i'(\cdot\,;u)$ are continuous on some neighborhood $W_1 \subseteq W$ of x (see [1], p. 32, cor.). Consequently by Lemma 2.1 and (2.3) $h'(\cdot\,;u)$ is also continuous. Next we show that

$$D^+(h'(\cdot\,;u))(z;u) = \max_{i \in I(g(z))} D^2 g_i(z;u,u) \tag{2.10}$$

for any $z \in W_1$. To do this, we choose a subnet $t_\nu > 0$ written for short t such that

$$\begin{aligned} D^+(h'(\cdot\,;u))(z;u) &= \lim_{t \downarrow 0} \frac{1}{t}\{h'(z+tu;u) - h'(z;u)\} \\ &= \lim_{t \downarrow 0} \frac{1}{t}\left\{ \max_{i \in I(g(z+tu))} g_i'(z+tu;u) - \max_{i \in I(g(z))} g_i'(z;u) \right\}. \end{aligned}$$

By Lemma 2.1 we can assume that $I(g(z+tu)) \subseteq I(g(z))$ and $z+tu \in W_1$. Since I is a finite set and considering a subnet if necessary, we can assume without loss of generality that there exists $i_z \in I(g(z+tu)) \subseteq I(g(z))$ such that

$$\max_{i \in I(g(z+tu))} g_i'(z+tu;u) = g_{i_z}'(z+tu;u)$$

for all t. By assumption, $g_i'(z;u) = g_{i_z}'(z;u)$ so

$$\max_{i \in I(g(z))} g_i'(z;u) = g_{i_z}'(z;u),$$

it follows that

$$\begin{aligned} D^+(h'(\cdot\,;u))(z;u) &= \lim_{t \downarrow 0} \frac{1}{t}\{g_{i_z}'(z+tu;u) - g_{i_z}'(z;u)\} \\ &= D^2 g_{i_z}(z;u,u). \end{aligned}$$

Now if (2.10) is not true, there must exist $i \in I(g(z))$ such that

$$D^2 g_i(z;u,u) > D^2 g_{i_z}(z;u,u);$$

since $g_i'(y;u) = g_{i_z}'(y;u)$ by assumption, it follows that

$$g_i'(z+\tau u;u) - g_{i_z}'(z+\tau u;u) > 0$$

for all small enough $\tau > 0$. Now we choose a small enough t from our net $\{t_\nu\}$ and recall that $i_z \in I(g(z+tu))$. But

$$\int_0^t (g_i - g_{i_z})'(z+\tau u;u)\,d\tau = g_i(z+tu) - g_{i_z}(z+tu) > 0,$$

contradicting the given assumption. Thus (2.10) is proved. On the other hand, we have by Proposition 1.4 that

$$\begin{aligned} h^\infty(x;u,u) &= (h'(\cdot\,;u))^0(x;u) = \limsup_{z \to x} D^+(h'(\cdot\,;u))(z;u) \\ &= \limsup_{z \to x} \max_{i \in I(g(z))} D^2 g_i(z;u,u) = D^2 g_{i_0}(x;u,u) \end{aligned}$$

for some $i_0 \in I(g(x))$, where the last equality is valid because of Lemma 2.1 and the fact that I is a finite set. Since

$$h^\infty(x;u,u) \geq D^+(h'(\cdot\,;u))(x;u)$$

and by (2.10)

$$D^+(h'(\cdot\,;u))(x;u) = \max_{i \in I(g(x))} D^2 g_i(x;u,u),$$

it follows that

$$h^\infty(x;u,u) = \max_{i \in I(g(x))} D^2 g_i(x;u,u). \qquad \square$$

COROLLARY 2.7. *Let X be a normed space, $u, x \in X$ and g_i, $1 \leq i \leq n$, be C^2-functions at x. Suppose further that $g_i'(x)$ are affinely independent. Then*

$$h^\infty(x;u,u) = \max_{i \in I(g(x))} D^2 g_i(x;u,u)$$

if and only if there exists a neighborhood W of x such that

$$g_i'(y;u) = g_j'(y;u)$$

for all $y \in W$ and $i, j \in I(g(y))$.

Proof. The sufficiency follows from Proposition 2.6. Conversely if $h^\infty(x;u,u) = \max_{i\in I(g(x))} D^2 g_i(x;u,u)$, then $h^\infty(x;u,u)$ is finite and hence, by Lemma 2.2, there exists a neighborhood W of x such that the condition $H(y;u,u) \underline{\propto} 0$ holds for each $y \in W$. This implies immediately that

$$g_i'(y;u) - g_j'(y;u) = 0$$

for all $y \in W$ and $i, j \in I(g(y))$. □

3. Generalized second-order Taylor expansion. Suppose that X and f are as in §1, we define the generalized Hessian [2] of f at x by

$$\partial^2 f(x)(u) := \{x^* \in X^*; \langle x^*, v\rangle \le f^\infty(x;u,v) \text{ for all } v \in X\},$$

where the symbol X^* denotes the dual space of X. It is easy to see that $\partial^2 f(x)(u)$ is a closed convex subset of X^* with respect to the w^*-topology. If f is twice C-differentiable at x [2], that is, $f^\infty(x;\cdot\,,v)$ (or equivalently $f^\infty(x;v,\cdot)$) is lower semi-continuous for each $v \in X$, then one has

$$f^\infty(x;u,v) = \sup\langle \partial^2 f(x)(u), v\rangle. \tag{3.1}$$

Now for a, u, v, and $x \in X$, we consider the following new kinds of first- and second-order directional derivatives, respectively, defined by

$$f_a^0(x;u) := \limsup_{\substack{\lambda\to 0\\ s\downarrow 0}} \frac{1}{s}\{f(x+\lambda a+su) - f(x+\lambda a)\},$$

$$f_{+a}^0(x;u) := \limsup_{s,\lambda\downarrow 0} \frac{1}{s}\{f(x+\lambda a+su) - f(x+\lambda a)\},$$

$$f_{-a}^0(x;u) := \limsup_{\substack{\lambda\uparrow 0\\ s\downarrow 0\\ \lambda+s\le 0}} \frac{1}{s}\{f(x+\lambda a+su) - f(x+\lambda a)\},$$

$$f_{0,a}(x;u) := -(-f)_a^0(x;u), \qquad f_{0,+a}(x;u) := -(-f)_{+a}^0(x;u)$$

and

$$f_a^\infty(x;u,v) := \limsup_{\substack{\lambda\to 0\\ s,t\downarrow 0}} \frac{1}{st}\{f(x+\lambda a+tu+sv) - f(x+\lambda a+tu) - f(x+\lambda a+sv) + f(x+\lambda a)\},$$

$$f_{\infty,a}(x;u,v) := \liminf_{\substack{\lambda\to 0\\ s,t\downarrow 0}} \frac{1}{st}\{f(x+\lambda a+tu+sv) - f(x+\lambda a+tu) - f(x+\lambda a+sv) + f(x+\lambda a)\},$$

(f_a^0 and f_a^∞ are different from f^0 and f^∞ as here we only consider

$$x + \lambda a \to x$$

along the direction a).

In terms of f_a^0, f_{+a}^0, and f_{-a}^0, we have the following.

LEMMA 3.1. *Suppose that $f : X \to \mathbb{R}$ is a continuous function. Then one has*

$$f_{+a}^0(x;a) = \lim \sup_{\lambda \downarrow 0} D^+ f(x+\lambda a;a),$$
$$f_{-a}^0(x;a) = \lim \sup_{\lambda \uparrow 0} D^+ f(x+\lambda a;a),$$

and

$$f_a^0(x;a) = \lim \sup_{\lambda \to 0} D^+ f(x+\lambda a;a).$$

Proof. Let $x, a \in X$, and $\lambda < 0, s > 0$. By Lemma 1.2 there exists $\alpha \in (0,s)$ such that

$$\frac{1}{s}\{f(x+\lambda a+sa) - f(x+\lambda a)\} \le D^+ f(x+\lambda a+\alpha a;a).$$

Hence, by the definition of f_{-a}^0 we have

$$f_{-a}^0(x;a) \le \lim \sup_{\lambda \uparrow 0} D^+ f(x+\lambda a;a).$$

But

$$\begin{aligned}\lim \sup_{\lambda \uparrow 0} D^+ f(x+\lambda a;a) &= \lim \sup_{\lambda \uparrow 0} \lim \sup_{s \downarrow 0} \frac{1}{s}\{f(x+\lambda a+sa) - f(x+\lambda a)\} \\ &\le \lim \sup_{\substack{\lambda \uparrow 0 \\ s \downarrow 0 \\ \lambda + s \le 0}} \frac{1}{s}\{f(x+\lambda a+sa) - f(x+\lambda a)\} = f_{-a}^0(x;a).\end{aligned}$$

So we have $f_{-a}^0(x;a) = \lim \sup_{\lambda \uparrow 0} D^+ f(x+\lambda a;a)$. Similarly, we have $f_{+a}^0(x;a) = \lim \sup_{\lambda \downarrow 0} D^+ f(x+\lambda a;a)$. Thus, one has

$$f_a^0(x;a) = \lim \sup_{\lambda \to 0} D^+ f(x+\lambda a;a). \qquad \square$$

The following theorem provides an answer to the question of Cominetti and Correa [2] about Taylor's expansion.

THEOREM 3.2. *Let $f : [x,y] \to \mathbb{R}$ be a continuous function on a line segment in a locally convex space X. Suppose that $D^+ f(\cdot\,; y-x)$ is finite, upper semi-continuous on (x,y) and $f_{+(y-x)}^0(x; y-x)$, $f_{-(y-x)}^0(y; y-x)$ are finite. Then there exists $t_0 \in (0,1)$ such that*

$$\begin{aligned}(3.2)\quad \frac{1}{2} f_{y-x}^\infty(x+t_0(y-x); y-x, y-x) &\ge f(y) - f(x) - f_{+(y-x)}^0(x; y-x) \\ &\ge \frac{1}{2} f_{\infty, y-x}(x+t_0(y-x); y-x, y-x).\end{aligned}$$

Hence, we also have

$$\begin{aligned}(3.3)\quad \frac{1}{2} f_{y-x}^\infty(x+t_0(y-x); y-x, y-x) + f_{y-x}^0(x; y-x) &\ge f(y) - f(x) \\ &\ge f_{0, y-x}(x; y-x) + \frac{1}{2} f_{\infty, y-x}(x+t_0(y-x); y-x, y-x).\end{aligned}$$

The following theorem is a corollary of Theorem 3.2.

THEOREM 3.3. *Suppose that the assumptions in Theorem* 3.2 *hold. If in addition* f *is defined on* X *and is twice* C*-differentiable at each point of* (x, y)*, then one has*

$$f(y) - f(x) - f^0_{+(y-x)}(x; y-x) \in \frac{1}{2}\overline{\langle \partial^2 f(x + t_0(y-x))(y-x), y-x\rangle}; \tag{3.4}$$

and also

$$f(y) - f(x) \in \overline{\langle \partial f(x), y-x\rangle + \frac{1}{2}\langle \partial^2 f(x + t_0(y-x))(y-x), y-x\rangle} \tag{3.5}$$

if $\partial f(x)$ *is nonempty and* $f^0(x; y-x) = \sup_{x^* \in \partial f(x)} \langle x^*, y-x\rangle$*, where* $\partial f(x)$ *denotes the Clarke's subdifferential and the "bar" denotes the closure of the set. The bar is superfluous if* f *is* $C^{1,1}$ [2] *on* (x, y).

Indeed, granting Theorem 3.2, we have

$$\begin{aligned} &\frac{1}{2} f^\infty(x + t_0(y-x); y-x, y-x) + f^0(x; y-x) \\ &\geq f(y) - f(x) \geq f_0(x; y-x) + \frac{1}{2} f_\infty(x + t_0(y-x); y-x, y-x) \end{aligned}$$

by (3.3). Thus, by (3.1) and our assumptions for any $\varepsilon > 0$ there exist $x_1^* \in \partial f(x)$ and $x_2^* \in \partial^2 f(x + t_0(y-x))(y-x)$ such that

$$f(y) - f(x) \leq \left\langle x_1^* + \frac{1}{2} x_2^*, y-x \right\rangle + \varepsilon.$$

Similarly, since $f_\infty(x + t_0(y-x); y-x, y-x) = -f^\infty(x + t_0(y-x); y-x, x-y)$ and $f_0(x; y-x) = -f^0(x; x-y)$, there exist

$$z_1^* \in \partial f(x) \quad \text{and} \quad z_2^* \in \partial^2 f(x + t_0(y-x))(y-x)$$

such that

$$f(x) - f(y) \leq \left\langle z_1^* + \frac{1}{2} z_2^*, x-y \right\rangle + \varepsilon.$$

Hence we can choose $\lambda \in (0, 1)$ such that

$$f(y) - f(x) = \left\langle (\lambda z_1^* + (1-\lambda)x_1^*) + \frac{1}{2}(\lambda z_2^* + (1-\lambda)x_2^*), y-x \right\rangle + (1-\lambda)\varepsilon - \lambda\varepsilon.$$

Since $\partial f(x)$ and $\partial^2 f(x + t_0(y-x))(y-x)$ are convex,

$$\lambda z_1^* + (1-\lambda)x_1^* \in \partial f(x) \quad \text{and} \quad \lambda z_2^* + (1-\lambda)x_2^* \in \partial^2 f(x + t_0(y-x))(y-x).$$

We then have

$$f(y) - f(x) \in \overline{\langle \partial f(x), y-x\rangle + \frac{1}{2}\langle \partial^2 f(x + t_0(y-x))(y-x), y-x\rangle}$$

as required to show for (3.5). Similarly one can prove (3.4). Thus it remains only to prove Theorem 3.2.

4. Detailed proof of Theorem 3.2. This section is entirely devoted to the proof of Theorem 3.2. Let $\gamma : [0,1] \to \mathbb{R}$ be defined by $\gamma(t) = f(x + t(y-x))$. Then it is easy to see that

$$\begin{gathered} D^+ f(x + t(y-x); y-x) = D^+\gamma(t;1), \\ f^0_{-(y-x)}(y; y-x) = \gamma^0_-(1;1) \qquad (= \gamma^0_{-1}(1;1)), \\ f^0_{+(y-x)}(x; y-x) = \gamma^0_+(0;1) \qquad (= \gamma^0_{+1}(0;1)), \\ f^0_{y-x}(x; y-x) = \gamma^0(0;1), \qquad f_{0,y-x}(x; y-x) = \gamma_0(0;1) \end{gathered}$$

and

$$\begin{gathered} f^\infty_{y-x}(x + t_0(y-x); y-x, y-x) = \gamma^\infty(t_0;1,1), \\ f_{\infty,y-x}(x + t_0(y-x); y-x, y-x) = \gamma_\infty(t_0;1,1). \end{gathered}$$

Then Theorem 3.2 can be rewritten as follows.

THEOREM 4.1. *Let $\gamma : [0,1] \to \mathbb{R}$ be a continuous function. Suppose that $D^+\gamma(\cdot\,;1)$ is finite, upper semi-continuous on $(0,1)$ and $\gamma^0_-(1;1), \gamma^0_+(0;1)$ are finite. Then there exists $t_0 \in (0,1)$ such that*

$$\frac{1}{2}\gamma^\infty(t_0;1,1) \geq \gamma(1) - \gamma(0) - \gamma^0_+(0;1) \geq \frac{1}{2}\gamma_\infty(t_0;1,1) \tag{4.1}$$

and so

$$\frac{1}{2}\gamma^\infty(t_0;1,1) + \gamma^0(0;1) \geq \gamma(1) - \gamma(0) \geq \gamma_0(0;1) + \frac{1}{2}\gamma_\infty(t_0;1,1). \tag{4.2}$$

To show Theorem 4.1, we need the following.

LEMMA 4.2. *Suppose that the function $h : [0,1] \to \mathbb{R}$ is upper semi-continuous on* (0,1) *with $h(0) = h(1)$ and*

$$\limsup_{t \downarrow 0} h(t) = h(0), \limsup_{t \uparrow 1} h(t) = h(1).$$

Then one of the following properties holds.

(i) *h attains a local maximum at some $t_0 \in (0,1)$;*

(ii) *There exists $t_0 \in (0,1)$ such that h is decreasing on $[0,t_0)$ and increasing on $(t_0,1]$.*

Proof. By assumptions, h is upper semi-continuous on $[0,1]$. Suppose (i) does not hold. Then h is neither decreasing on $[0,1)$ nor increasing on $(0,1]$, for otherwise the assumptions of the lemma would imply that h is a constant function. Thus there exist $t_1, t_2 \in [0,1)$ with $t_1 < t_2$ and $h(t_1) < h(t_2)$. We then claim that h is increasing on $(t_2,1]$. In fact if there exist $t_3, t_4 \in (t_2,1]$ with $t_3 < t_4$ such that $h(t_3) > h(t_4)$, the upper semi-continuity of h on $[t_1,t_4]$ will imply that (i) holds at some interior point of $[t_1,t_4]$.

Let t_0 denote the greatest lower bound of the nonempty set $T := \{t \in (0,1); h \text{ is increasing on } (t,1]\}$. Then $t_0 \leq t_2 < 1$ and also $t_0 \neq 0$ because h is not increasing on $(0,1]$. Note further that h is decreasing on $[0,t_0)$, for otherwise one can show as above that there exist $\bar{t}_1, \bar{t}_2 \in [0,t_0)$ with $\bar{t}_1 < \bar{t}_2, h(\bar{t}_1) < h(\bar{t}_2)$ and hence that h is increasing on $(\bar{t}_2,1]$, contradicting the definition of t_0. It is now clear that t_0 has the properties required in (ii). □

Now we prove Theorem 4.1. Define the function h: [0,1] $\to \mathbb{R}$ by

$$h(t) := \gamma(t) - \gamma(1) + (1-t)\xi(t) + (1-t)^2[\gamma(1) - \gamma(0) - \gamma^0_+(0;1)],$$

where

$$\xi(t) := \begin{cases} \gamma_+^0(0;1) & t = 0, \\ D^+\gamma(t;1) & 0 < t < 1, \\ \gamma_-^0(1;1) & t = 1. \end{cases}$$

Then by the finiteness assumption of $\gamma_-^0(1;1)$ and $\gamma_+^0(0;1)$ it follows from Lemma 3.1 that $h(0) = h(1) = 0$,

$$\limsup_{t\downarrow 0} h(t) = \limsup_{t\downarrow 0} D^+\gamma(t;1) - \gamma_+^0(0;1) = 0 = h(0)$$

and

$$\limsup_{t\uparrow 1} h(t) = 0 = h(1).$$

Further h is upper semi-continuous on $(0,1)$ since $D^+\gamma(\cdot\,;1)$ is assumed upper semi-continuous on $(0,1)$. Thus, Lemma 4.2 is applicable to h and so there exists $t_0 \in (0,1)$ such that either (i) h attains a local maximum at t_0 or (ii) h is decreasing on $[0,t_0)$ and increasing on $(t_0,1]$.

(I) Suppose (i) holds. Then we have (a) $0 \geq D^+h(t_0;1)$ and (b) $0 \leq h^0(t_0;1)$ by a well-known result of Clarke [1], p. 38. (As h is not necessarily Lipschitz, we briefly indicate a proof here: t_0 is a local minimum point for $-h$ so $0 \leq (-h)^0(t_0;-1) = h^0(t_0;1)$). Note that, by subadditivity,

$$\begin{aligned} h^0(t_0;1) \leq \gamma^0(t_0;1) - D^+\gamma(t_0;1) + (1-t_0)(D^+\gamma(\cdot\,;1))^0(t_0;1) \\ - 2(1-t_0)[\gamma(1) - \gamma(0) - \gamma_+^0(0;1)], \end{aligned}$$

where the first two terms can be cancelled out because

$$\limsup_{t\to t_0} D^+\gamma(t;1) = \gamma^0(t_0;1)$$

by Lemma 1.2 and $\limsup_{t\to t_0} D^+\gamma(t;1) \leq D^+\gamma(t_0;1)$ by the upper semi-continuity assumption of $D^+\gamma(\cdot\,;1)$. Hence (b) and (1.3) of Proposition 1.4 imply that

$$\gamma(1) - \gamma(0) - \gamma_+^0(0;1) \leq \frac{1}{2}(D^+\gamma(\cdot\,;1))^0(t_0;1) = \frac{1}{2}\gamma^\infty(t_0;1,1). \tag{4.3}$$

This verifies one inequality required in (4.1). The other inequality in (4.1) follows similarly from (a) because, by (1.5) of Proposition 1.4, one has

$$D_+(D^+\gamma(\cdot\,;1))(t_0;1) \geq \gamma_\infty(t_0;1,1)$$

and, by elementary computation rules for D^+ and D_+, that

$$\begin{aligned} D^+h(t_0;1) \geq D^+\gamma(t_0;1) - D^+\gamma(t_0;1) + (1-t_0)D_+(D^+\gamma(\cdot\,;1))(t_0;1) \\ - 2(1-t_0)[\gamma(1) - \gamma(0) - \gamma_+^0(0;1)]. \end{aligned} \tag{4.4}$$

(II) We next consider the case when (ii) holds: h is decreasing on $[0,t_0)$ and increasing on $(t_0,1]$. Take a sequence $t_n \uparrow t_0$ and note that (ā) $0 \geq D^+\gamma(t_n;1)$ for each n and (b̄) $0 \leq h^0(t_0;1)$. As done above (b̄) ensures that (4.3) holds while (ā) implies that

$$\gamma(1) - \gamma(0) - \gamma_+^0(0;1) \geq \frac{1}{2}D_+(D^+\gamma(\cdot\,;1))(t_n;1) \geq \frac{1}{2}\gamma_\infty(t_n;1,1)$$

because (4.4) holds with t_0 replaced by t_n. Since $\gamma_\infty(\cdot\,;1,1)$ is lower semi-continuous (Proposition 1.1), we have the other inequality required in (4.1) in addition to (4.3). □

5. Corollaries of Theorem 3.2 and Theorem 3.3.

COROLLARY 5.1 [2], Prop. 4.1. *Suppose that $f : X \to \mathbb{R}$ is continuously Gâteaux differentiable and twice C-differentiable on a segment $[x, y] \subseteq X$. Then there exists $t_0 \in (0, 1)$ such that*

$$f(y) - f(x) - f'(x; y - x) \in \frac{1}{2}\overline{\langle \partial^2 f(x + t_0(y - x))(y - x), y - x\rangle}.$$

If f is $C^{1,1}$ on $[x, y]$, then the closure can be ignored.

Proof. Since f is continuously Gâteaux differentiable at each point of $[x, y]$, it satisfies the assumptions in Theorem 3.2. Now apply Theorem 3.3. □

COROLLARY 5.2. *Suppose that $f : X \to \mathbb{R}$ is continuous at each point of a segment $[x, y]$. Then f satisfies* (3.2) *in each of the following cases*:

(i) *$D^+f(\cdot; y - x)$, $f^\infty_{y-x}(\cdot; y - x, y - x)$, and $f_{\infty,y-x}(\cdot; y - x, y - x)$ are finite on (x, y) and $f^0_{+(y-x)}(x; y - x)$, $f^0_{-(y-x)}(y; y - x)$ are finite;*

(ii) *$D^+f(z; y - x) = f^0_{y-x}(z; y - x)$ at each point of (x, y) and $f^0_{+(y-x)}(x; y - x)$, $f^0_{-(y-x)}(x; y - x)$ are finite;*

(iii) *f is regular in the Clarke's sense at each point of (x, y) and $f^0_{+(y-x)}(x; y - x)$, $f^0_{-(y-x)}(y; y - x)$ are finite.*

Proof. Suppose that (i) is true. Let $\gamma(t) := f(x + t(y - x)), t \in (0, 1)$. Clearly, it suffices to show that $D^+\gamma(\cdot; 1)$ is upper semi-continuous on $(0, 1)$. Now $D^+\gamma(t; 1)$ and $\gamma^\infty(t; 1, 1)$ are finite for any $t \in (0, 1)$. Take a finite number $K > \gamma^\infty(t; 1, 1)$. Then, by (1.4) of Proposition 1.4, there exists $\delta > 0$ such that

$$K > \frac{1}{\lambda}\{D^+\gamma(t' + \lambda; 1) - D^+\gamma(t'; 1)\}$$

whenever $|t' - t| < \delta$ and $0 < \lambda < \delta$. Passing to the limits as $\lambda \downarrow 0$ and $t' \to t$, it follows that

$$0 \geq \limsup_{\substack{t' \to t \\ \lambda \downarrow 0}} \{D^+\gamma(t' + \lambda; 1) - D^+\gamma(t'; 1)\}$$

and so

$$0 \geq \limsup_{\lambda \downarrow 0} D^+\gamma(t + \lambda; 1) - D^+\gamma(t; 1).$$

Thus,

$$D^+\gamma(t; 1) \geq \limsup_{\lambda \downarrow 0} D^+\gamma(t + \lambda; 1) = \limsup_{t' \downarrow t} D^+\gamma(t'; 1). \tag{5.1}$$

Similarly, since $\gamma_\infty(t; 1, 1)$ is finite, one can apply (1.5) of Proposition 1.4 to show that

$$0 \leq \liminf_{\substack{t' \to t \\ \lambda \downarrow 0}} \{D^+\gamma(t' + \lambda; 1) - D^+\gamma(t'; 1)\}.$$

Letting $\tau' = t' + \lambda$, we then obtain

$$\begin{aligned} 0 &\geq \limsup_{\substack{\tau' \to t \\ \lambda \downarrow 0}} \{D^+\gamma(\tau' - \lambda; 1) - D^+\gamma(\tau'; 1)\} \\ &\geq \limsup_{\lambda \downarrow 0} D^+\gamma(t - \lambda; 1) - D^+\gamma(t; 1) \end{aligned}$$

and so

$$D^+\gamma(t;1) \geq \limsup_{\lambda \downarrow 0} D^+\gamma(t-\lambda;1) = \limsup_{t' \uparrow t} D^+\gamma(t';1). \tag{5.2}$$

Together with (5.1) we have

$$D^+\gamma(t;1) \geq \limsup_{t' \to t} D^+\gamma(t';1),$$

showing that $D^+\gamma(\cdot\,;1)$ is upper semi-continuous on $(0,1)$.

In the case (ii) $D^+f(\cdot\,;y-x)$ is upper semi-continuous on (x,y) since $f^0_{y-x}(z;y-x)$ is clearly so. Consequently Theorem 3.2 is applicable.

For the case (iii), let $z = x + t(y-x), t \in (0,1)$. Then, by the regularity of f, Lemmas 1.2 and 3.1, one has

$$\begin{aligned} f'(z;y-x) = f^0(z;y-x) &= \limsup_{z' \to z} f'(z';y-x) \\ &\geq \limsup_{t' \to t} f'(x+t'(y-x);y-x) = f^0_{y-x}(z;y-x) \end{aligned}$$

showing that

$$f'(z;y-x) = f^0_{y-x}(z;y-x)$$

for any $z \in (x,y)$. Thus, the result holds from the case (ii). □

COROLLARY 5.3. *Let $-f : [x,y] \to \mathbb{R}$ satisfy the assumptions in Theorem* 3.2. *Then there exists $t_0 \in (0,1)$ such that*

$$\begin{aligned} \frac{1}{2} f^\infty_{y-x}(x+t_0(y-x);y-x,y-x) &\geq f(y) - f(x) - f_{0,+(y-x)}(x;y-x) \\ &\geq \frac{1}{2} f_{\infty,y-x}(x+t_0(y-x);y-x,y-x) \end{aligned} \tag{5.1}$$

and so (3.3) *holds, where* $f_{0,+(y-x)}(x;y-x) = -(-f)^0_{+(y-x)}(x;y-x)$.

Proof. By Theorem 3.2, we have

$$\begin{aligned} \frac{1}{2}(-f)^\infty_{y-x}(x+t_0(y-x);y-x,y-x) &\geq f(x) - f(y) - (-f)^0_{+(y-x)}(x;y-x) \\ &\geq \frac{1}{2}(-f)_{\infty,y-x}(x+t_0(y-x);y-x,y-x) \end{aligned}$$

and so, by elementary results similar to (iii) of Proposition 1.1,

$$\begin{aligned} \frac{1}{2} f^\infty_{y-x}(x+t_0(y-x);y-x,y-x) &\geq f(y) - f(x) - f_{0,+(y-x)}(x;y-x) \\ &\geq \frac{1}{2} f_{\infty,y-x}(x+t_0(y-x);y-x,y-x). \end{aligned}$$

This implies immediately that

$$\begin{aligned} &\frac{1}{2} f^\infty_{y-x}(x+t_0(y-x);y-x,y-x) + f^0_{y-x}(x;y-x) \geq f(y) - f(x) \\ &\quad \geq f_{0,y-x}(x;y-x) + \frac{1}{2} f_{\infty,y-x}(x+t_0(y-x);y-x,y-x). \end{aligned}$$ □

Remark 1. By [1], Prop. 2.3.6, it follows from part (ii) of Corollary 5.2 and Corollary 5.3 that a convex function satisfies (3.2) and a concave function satisfies (5.1), respectively, and both satisfy (3.3).

Remark 2. In each of the cases (i)–(iii), it is well known that f can fail to have Gâteaux derivative at some points so [2], Prop. 4.1 is not applicable.

6. Some applications in optimization.

DEFINITION 6.1. *Let $f : X \to \mathbb{R}$ and $x \in X$. $\partial^2 f(x)$ will be said to be positively definite* [2] *if $f_\infty(x; u, u) > 0$ for every $u \in X, u \neq 0$. Furthermore, a function $f : X \to \mathbb{R}$ is called twice uniformly locally Lipschitzian at x* [2] *if there exist neighborhoods X_0 of x and U of zero such that $f^\infty(X_0; U, U)$ is bounded in $\mathbb{R}$. This condition implies in particular that f is twice C-differentiable at each point x_0 in X_0 because then, for each $u \in U$, the sublinear map $v \mapsto f^\infty(x_0; u, v)$ is bounded on U and hence continuous on X.*

PROPOSITION 6.2. *Let $x \in X = \mathbb{R}^n, f : X \to \mathbb{R}$ be locally Lipschitz near x and twice uniformly locally Lipschitzian at x. If $f^0_{+u}(x; u) \geq 0$ for all $u \in X$, then a sufficient condition for x to be a strict local minimum point of f is that $\partial^2 f(x)$ is positively definite.*

Proof. By assumption, take a constant $M > 1$ and neighborhoods X_0 of x and U of zero such that

$$|f^\infty(X_0; U, U)| < M \tag{6.1}$$

and that f on X_0 is Lipschitz. Let $B := \{u \in X; \|u\| = 1\}$ and $u \in B$. By the strict positivity of $f_\infty(x; u, u)$ and the lower semi-continuity of $f_\infty(\cdot\,; u, u)$, one has a convex neighborhood $W(u)$ of x contained in X_0 and $1 > \delta(u) > 0$ such that

$$f_\infty(y; u, u) > \delta(u)$$

for all $y \in W(u)$. Let $1 > \lambda > 0$ with $\lambda u \in U$ and $U(u) = [\lambda\delta(u)/8M]U$. For any $v \in u + U(u), y \in W(u)$, it follows from Proposition 1.1 that

$$\begin{aligned} f_\infty(y; v, v) &= f_\infty(y; u + (v - u), u + (v - u)) \\ &\geq f_\infty(y; u, u) + f_\infty(y; v - u, v - u) + 2f_\infty(y; u, v - u) \\ &> \delta(u) - \frac{\lambda^2\delta(u)^2}{8^2M^2} \cdot M - \frac{\delta(u)}{4} \geq \frac{\delta(u)}{2} > 0. \end{aligned}$$

Since $X = \mathbb{R}^n$, by the compactness of B we can choose m neighborhoods $u_1 + U(u_1), \ldots, u_m + U(u_m)$ whose union covers B. Let

$$W = \bigcap_{i=1}^{m} W(u_i) \quad \text{and} \quad \delta = \min_{1 \leq i \leq m} \{\delta(u_i)\}.$$

Then for any $v \in B, y \in W$,

$$f_\infty(y; v, v) > \delta/2;$$

consequently $f_\infty(y; v, v) > 0$ for all $v \in X$ and $y \in W$. In view of the Assumption (6.1), it follows from part (i) of Corollary 5.2, that for any $y \in W, y \neq x$, there exists $t_0 \in (0, 1)$ such that

$$f(y) - f(x) \geq f^0_{+(y-x)}(x; y - x) + \frac{1}{2} f_\infty(y + t_0(x - y); y - x, y - x) > 0$$

because $f^0_{+(y-x)}(x; y - x) \geq 0$ and $y + t_0(x - y) \in W$. Therefore x is a strictly local minimal point. □

Remark. The preceding proposition can be deduced from [2], Prop. 5.2 because the twice uniformly locally Lipschitzian of f implies $f \in C^{1,1}$ [18]. We are indebted to the referee for the reference [18].

Let $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}^n$ be locally Lipschitz functions, C be a closed subset of X.

Now we consider the minimization problem with constraint

$$(\mathbb{P}) \qquad \min\{f(x); x \in Q\},$$

where $Q := \{x \in C \text{ and } g(x) \leq 0\}$. If x_0 is a solution of problem $(\mathbb{P})$, then by Clarke's Theorem [1], Thm. 6.1.1, there exists a multiplier $(\lambda, \gamma) \in \mathbb{R}^1 \times \mathbb{R}^n$ with $\lambda, \gamma_i \geq 0, 1 \leq i \leq n$, and $\lambda + \sum_{i=1}^n \gamma_i = 1$ such that

$$\gamma g(x_0) = 0 \quad \text{and} \quad 0 \leq L^0(x_0; u) \tag{6.2}$$

for any $u \in X$, where

$$L(x) := \lambda f(x) + \gamma g(x) + \alpha d_Q(x)$$

and α is a Lipschitzian constant for both f and g on a neighborhood of x_0.

PROPOSITION 6.3. *Suppose that x_0 is a solution of the problem $(\mathbb{P})$. Let $A := \{v; \gamma g(v) \geq 0\}$ with the contingent cone $T_A(x_0)$ [3]. Then*

(i) $L^\infty(x_0; u, u) \geq 0$, *for any u in $T_A(x_0)$ with $D_+L(x_0; u) = 0$;*

(ii) $L^{00}(x_0; u, u) \geq 0$, *for any u in $T_A(x_0)$ with $L^0(x_0; u) = 0$.*

Proof. (i) Since $f(x_0) \leq f(x)$ for any $x \in Q$, by [1], Prop. 2.4.3, $f + \alpha d_Q$ attains a local minimum at x_0. Let $u \in T_A(x_0)$ with $D_+L(x_0; u) = 0$ and take sequences $u_i \to u$ and $t_i \downarrow 0$ with $x_0 + t_i u_i \in A$. Therefore, one has

$$\begin{aligned} & L(x_0 + t_i u_i) - L(x_0) \\ &= \lambda f(x_0 + t_i u_i) + \gamma g(x_0 + t_i u_i) + \alpha d_Q(x_0 + t_i u_i) - \lambda f(x_0) \\ &= \lambda\{f(x_0 + t_i u_i) + \alpha d_Q(x_0 + t_i u_i) - f(x_0)\} + \gamma g(x_0 + t_i u_i) \\ &\qquad + (1 - \lambda)\alpha d_Q(x_0 + t_i u_i) \geq 0. \end{aligned}$$

By Lemma 1.2 there exists $\tau_i \in (0, t_i)$ such that

$$D_+L(x_0 + \tau_i u; u) \geq \frac{1}{t_i}(L(x_0 + t_i u) - L(x_0)) \geq 0.$$

Therefore $\limsup_{\tau \downarrow 0} \frac{1}{\tau} D_+L(x_0 + \tau u; u) \geq 0$. Since $D_+L(x_0; u) = 0$, it follows from Proposition 1.4 that

$$\begin{aligned} L^\infty(x_0; u, u) &= \limsup_{\substack{y \to x_0 \\ t \downarrow 0}} \frac{1}{t}(D_+L(y + tu; u) - D_+L(y; u)) \\ &\geq \limsup_{t \downarrow 0} \frac{1}{t}(D_+L(x_0 + tu; u) - D_+L(x_0; u)) \\ &= \limsup_{t \downarrow 0} \frac{1}{t} D_+L(x_0 + tu; u) \geq 0. \end{aligned}$$

(ii) By Definition 1.5 and similar proof of part (i), one has

$$\begin{aligned} L^{00}(x_0; u, u) &= \limsup_{\substack{y \to x_0 \\ t \downarrow 0}} \frac{1}{t}(L^0(y + tu; u) - L^0(y; u)) \\ &\geq \limsup_{t \downarrow 0} \frac{1}{t}(L^0(x_0 + tu; u) - L^0(x_0; u)) \\ &\geq \limsup_{t \downarrow 0} \frac{1}{t} D_+L(x_0 + tu; u) \geq 0. \end{aligned}$$

□

Acknowledgments. This work was partially supported by a Direct Grant for Research from the Research Grants Council of Hong Kong. The authors also express their gratitude to the referees for helpful comments and suggestions.

REFERENCES

[1] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[2] R. COMINETTI AND R. CORREA, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.
[3] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.
[4] J.-B. HIRIART-URRUTY, S.-J. STRODIOT, AND V. HIEN HGUYEN, *Generalized Hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data*, Appl. Math. Optim., 11 (1984), pp. 43–56.
[5] A. AUSLENDER AND R. COMINETTI, *A comparative study of multifunction differentiability with applications in mathematical programming*, Math. Oper. Res. 16 (1991), pp. 240–258.
[6] A. BEN-TAL AND J. ZOWE, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming Stud., 19 (1982), pp. 39–76.
[7] R. W. CHANEY, *Second-order sufficiency conditions for nondifferentiable programming problems*, SIAM J. Control Optim., 20 (1982), pp. 20–33.
[8] ———, *Second-order necessary conditions in constrained semismooth optimization*, SIAM J. Control Optim., 25 (1987), pp. 1072–1081.
[9] A. D. IOFFE, *Approximate subdifferentials and applications. I: The finite dimensional theory*, Trans. Amer. Math. Soc. 281 (1984), pp. 389–416.
[10] ———, *Calculus of Dini subdifferentials of functions and contingent coderivatives of set-valued maps*, Nonlinear Anal. Theory Methods Appl. 8 (1984), pp. 517–539.
[11] M. PAPPALARDO, *Tangent cones and derivatives*, J. Optim. Theory Appl. 70 (1991), pp. 97–107.
[12] B. N. PSHENICHNYI, *Necessary Conditions for an Extremum*, Marcel Dekker, New York, 1983.
[13] R. T. ROCKAFELLAR, *Directionally Lipschitzian functions and subdifferential Calculus*, Proc. London Math. Soc. (3) 39 (1979), pp. 331–355.
[14] ———, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math. Soc. 37 (1980), pp. 257–280.
[15] ———, *First- and second-order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc. 307 (1988), pp. 75–108.
[16] ———, *Generalized second derivatives of convex functions and saddle functions*, Trans. Amer. Math. Soc. 322 (1990), pp. 51–77.
[17] A. J. WHITE, *Real Analysis: An Introduction*, Addison-Wesley, Reading, MA, 1968.
[18] R. COMINETTI, *Equivalence between the classes of $C^{1,1}$ and twice locally Lipschitzian functions*, Ph.D. thesis, Université Blaise Pascal, Aubiere, France, 1989.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 612–634, May 1994

002

OPTIMAL CONTROL ON THE L^∞ NORM OF A DIFFUSION PROCESS*

GUY BARLES†, CHRISTIAN DAHER‡, AND MARC ROMANO§

Abstract. Stochastic control problems are considered, where the cost to be minimized is either a running maximum of the state variable or more generally a running maximum of a function of the state variable and the control. In both cases it is proved that the value function, which must be defined on an augmented state space to take care of the non-Markovian feature of the running maximum, is the unique viscosity solution of the associated Bellman equation, which turns out to be, in the second case, a variational inequality with an oblique derivative boundary condition. Most of this work consists of proving the convergence of L^p approximations and this is done by purely partial differential equation (PDE) methods.

Key words. stochastic control, running maximum, variational inequality, viscosity solutions

AMS subject classifications. 93E20, 35R05, 35Q80

Introduction. This paper is concerned with stochastic control problems where the cost to be minimized is given by a running maximum of the state variable over time. For the sake of simplicity, we will consider here the following two model problems. Let X_s be a state variable in $\mathbb{R}^n$ whose dynamic is governed by the controlled stochastic differential equation

$$\begin{aligned} dX_s &= b(X_s, s, \alpha_s)ds + \sigma(X_s, s, \alpha_s)dW_s \qquad 0 \le t \le s \le T, \\ X_t &= x \in \mathbb{R}^n, \end{aligned}$$

where $(\alpha_s)_s$ denotes the control process, which takes its values in a compact metric space A, and b, σ are given continuous functions. We are interested in control problems whose cost to be minimized is given either by

$$J(x,t,(\alpha_s)_s) = E\left[\psi\left(\sup_{s\in[t,T]} |X_s|\right)\right] \tag{1}$$

or, more generally, by

$$J(x,t,(\alpha_s)_s) = E\left[\psi\left(\sup_{s\in[t,T]} f(X_s,s,\alpha_s)\right)\right], \tag{2}$$

where ψ and $f \ge 0$ are given continuous functions.

Our goal is to explain how to determine the value function

$$V(x,t) = \inf_{(\alpha_s)_s} J(x,t,(\alpha_s)_s). \tag{3}$$

This work was originally motivated by problems arising in finance theory, in particular lookback options pricing models, in which the terminal pay-off of the contingent claim is not measurable with respect to the basic's terminal price, but is defined as a function of the L^p-norm (with possibly $p = \infty$) of the whole sample path of the price process. We take this opportunity for recasting together old and new ideas for attacking these kinds of problems.

We begin by recalling that the first difficulty of (1) and (2) is that the form of the cost function does not allow a straightforward application of the dynamic programming principle:

* Received by the editors December 16, 1991; accepted for publication (in revised form) September 29, 1992. This work was done in the Caisse Autonome de Refinancement Research and Development Department.

† Faculté des Sciences et Techniques, Université de Tours, Parc de Grandmont, 37200 Tours, France.

‡ Caisse Autonome de Refinancement, 2, square de Luynes, 75007 Paris, France.

§ Ecole Normale Supérieure, 45, Rue d'Ulm, 75005 Paris, France.

typically in (1), the non-Markovian feature of the process $Z_s = \sup_{\tau\in[t,s]} |X_\tau|$ for (1) (and $Z_s = \sup_{\tau\in[t,s]} f(\tau, X_\tau, \alpha_\tau)$ for (2)) prohibits the direct use of the standard martingale approach to deduce the Hamilton–Jacobi–Bellman (HJB) equation for V. To circumvent this difficulty, it is classical to reformulate (3) by adding one extra state variable, namely Z_s, that, roughly speaking, "carries the past information." With this well-known trick, the new cost function—denoted by $u^\infty(x, z, t)$—recovers a classical form and $V(x, t)$ is nothing but $u^\infty(x, x, t)$. This device apparently cannot be avoided in a stochastic context: indeed, readers familiar with probability theory are acquainted with the fact that, for instance, the law of the running maximum of the Brownian motion is computed *jointly* with the law of the Brownian motion itself (see, for example, [26] and references therein).

The second difficulty is that Z does not solve a stochastic differential equation with regular coefficients; to solve it by partial differential equation (PDE) arguments, we replace the L^∞-norm appearing in Z by an L^p approximation. As already mentioned, this kind of problem has itself an interest in finance, in particular the $p = 1$ case. Section 1.2 investigates the characterization of the approximate value function (u^p) as unique viscosity solution of the associated HJB equation on the augmented state space.

Then the passage to the limit $(p \to +\infty)$ gives us the equation satisfied by u^∞; this turns out to be an HJB equation set in a cone with homogeneous Neumann—or more precisely oblique derivative—boundary conditions. Actually, we meet again (at the PDE level) the strong link existing between the law of the supremum of a diffusion process and the law of its reflexion, but with completely different context and tools than those encountered in probability theory (see, e.g., [26] and references therein).

It is worth mentioning that the convergence of u^p to u^∞ is not obvious; this is a striking difference with [1], where a similar method is used. We first must prove the uniform convergence of u^p to u^∞, and we obtain it by a method that *relies on purely PDE arguments* (§1.4).

We conclude the paper with the applications to finance. After stating an extension of the previous results to the case of control with optimal stopping (§3.1), we show how to apply this to price American calls on stock options (§3.2).

Our work relies on the notion of viscosity solutions introduced by Crandall and Lions [12]; we refer to the "User's Guide" [11] for a complete presentation of this notion of weak solutions for degenerate elliptic and parabolic PDEs. The links between viscosity solutions and stochastic optimal control have been first cleared out in the works of Lions [29], [30] and [31].

Similar problems, but in the deterministic case and not with the present generality, were studied in [7]. It is worth mentioning that in [7] the limiting equation is a quasi-variational inequality in the original state space variables; however, in their case, the L^p approximation reduces to a standard problem by commuting the $1/p$ of the L^p norms and the infimum.

After this work was completed, we learned that Barron [6] and Heinricher and Stockbridge [18] considered the same types of problem: in [6], the case of (1) is completely treated with an even greater generality on the diffusion processes since the functions b and σ may depend on the Z process (our methods can also treat this more general case). The main differences with our work is that [6] uses deep probabilistic results for obtaining the convergence of u^p to u^∞ while we use a far simpler PDE method to get it; moreover [6] does not consider the case of (2). In [18], the problem (1) is introduced and some special cases are solved by dynamic programming arguments.

1. The model case.

1.1. Reformulation of the basic problem. We will consider in this part the control problem with a cost function given by (1). We recall that X's dynamics are governed by the

following stochastic differential equation:

$$dX_s = b(X_s, s, \alpha_s)ds + \sigma(X_s, s, \alpha_s)dW_s \qquad 0 \le t \le s \le T, \tag{4}$$

$$X_t = x \in \mathbb{R}^n, \tag{5}$$

where b and σ are continuous functions defined on $\mathbb{R}^n \times [0,T] \times A$ with values, respectively, in $\mathbb{R}^n$ and $\mathbb{M}^{n\times m}$ (the space of $n \times m$ matrices). A is a compact metric space; $(\Omega, \mathcal{F}, (\mathcal{F}_s)_{s\in[0,T]}, (W_s)_{s\in[0,T]})$ is a standard Brownian motion in $\mathbb{R}^m$; and $(\alpha_s)_s$ is a previsible A-valued process.

For the reasons explained in the Introduction, we first augment the state space to come down to a classical cost shape. The new state space will be denoted by $Q = \mathbb{R}^n \times \mathbb{R}_+^* \times [0,T]$. Second, we approximate the L^∞ norm by L^p norms. Thus, we introduce, for all $1 \le p \le \infty$, a second state variable in $\mathbb{R}_+$ that is defined, given an initial condition $(x,z,t) \in Q$, by (if $p < \infty$)

$$Z_s^p = \left(z^p + \int_t^s |X_\tau|^p d\tau\right)^{1/p}, \tag{6}$$

and by (if $p = \infty$)

$$Z_s^\infty = \max\left\{z, \sup_{\tau\in[t,s]} |X_\tau|\right\}, \tag{7}$$

X being the solution of (4), (5).

The cost function is given by

$$J^p(x,z,t,(\alpha_s)_s) = E[\psi(Z_T^p)], \tag{8}$$

where ψ is a bounded, Lipschitz-continuous $\mathbb{R}$-valued function. The superscript emphasizes the dependence of J with respect to p. We want to minimize J^p over all admissible controls, as follows:

$$u^p(x,z,t) = \inf_{(\alpha_s)_s} J^p(x,z,t,(\alpha_s)_s). \tag{9}$$

Throughout the paper, we make the following assumption.

ASSUMPTION 1. *For $\phi = f, b_i\ (1 \le i \le n)$ and $\sigma_{ij}\ (1 \le i \le n, 1 \le j \le m), \phi$ is continuous, $\phi(.,.,\alpha) \in W^{1,\infty}(\mathbb{R}^n \times (0,+\infty))$ for any $\alpha \in A$ and*

$$\sup_{\alpha\in A} \|\phi(.,.,\alpha)\|_{1,\infty} < \infty. \tag{10}$$

We recall that $W^{1,\infty}(\mathbb{R}^n \times (0,+\infty))$ is the set of functions z such that $z, D_x z, \partial z/\partial t$ are in $L^\infty(\mathbb{R}^n \times (0,+\infty))$ and that this space is equipped with the norm

$$\|z\|_{1,\infty} = \|z\|_\infty + \|D_x z\|_\infty + \left\|\frac{\partial z}{\partial t}\right\|_\infty.$$

1.2. The control problem for $1 \le p < \infty$. In this part, we focus on the case $1 \le p < \infty$ and prove that the value function of the problem is bounded, continuous, and the unique viscosity solution of the related HJB equation.

Differentiating (6) with respect to time yields

$$dZ_s = \frac{Z_s}{p}\left(\frac{|X_s|}{Z_s}\right)^p ds, \tag{11}$$

$$Z_t = z \geq 0, \tag{12}$$

which is nothing but an ordinary differential equation. We easily show that, for almost every sample path of X, (11), (12) has a unique solution and that the apparent singularity for $z = 0$ causes, in fact, no problem. The control problem (9) has thus, with respect to the system (X_s, Z_s), a (quasi-) standard shape and the dynamic programming principle hence applies.

The related HJB equation can be written

$$\mathcal{H}^p\left(x, z, t, \frac{\partial u}{\partial t}, D_x u, D_z u, D^2_{xx} u\right) = 0 \quad \text{in } Q, \tag{13}$$

where we set, for $\xi = (x, z, t, q_t, q_x, q_z, M) \in Q \times \mathbb{R}^{n+2} \times \mathcal{S}^n$,[1]

$$\mathcal{H}^p(\xi) = \mathcal{L}(\xi) = \frac{z}{p}\left(\frac{|x|}{z}\right)^p q_z$$

and

$$\mathcal{L}(\xi) = \mathcal{L}(x, t, q_t, q_x, M) = -q_t + \sup_{\alpha \in A}\left\{-\frac{1}{2}\operatorname{Tr}((\sigma\sigma^T)(x, t, \alpha)M) - b(x, t, \alpha)q_x\right\},$$

with terminal condition

$$u(x, z, T) = \psi(z) \quad \text{in } \mathbb{R}^n \times \mathbb{R}^*_+. \tag{14}$$

Throughout the paper, we use indifferently the notation $\mathcal{L}(x, t, \partial u/\partial t, D_x u, D^2_{xx} u), \mathcal{L}(x, t)$, or $\mathcal{L}u$ both for emphasizing the important dependence of $\mathcal{L}$ with respect to the function or the current point and also for simplicity of notations.

We can state the following theorem.

THEOREM 1.2.1. *We have*

(i) u^p *defined by* (9) *is bounded and Lipschitz-continuous in* x *and* z, *uniformly for* $t \in [0, T]$,

(ii) u^p *is the unique bounded viscosity solution of* (13) *and* (14).

Proof of Theorem 1.2.1. (i) u^p is bounded since ψ is. Let (x, z, t) and (x', z', t) be in Q and denote (X, Z) and (X', Z') the related processes. It is well known that, from (10),

$$\sup_{(\alpha_s)_s} E\left[\sup_{s \in [t,T]} |X_s - X'_s|\right] \leq k|x - x'|,$$

for some constant k. Now

$$Z_T = \left(\int_0^T (z\chi_{(s<t)}t^{-1/p} + |X_s|\chi_{(s\geq t)})^p ds\right)^{1/p} \quad \text{a.s.}$$

[1] $\mathcal{S}^n$ is the space of $n \times n$ symmetric matrices.

By the Minkowski inequality for L^p norms

$$|Z_T - Z'_T| \leq \left(\int_0^T (|z - z'|\chi_{(s<t)} t^{-1/p} + |X_s - X'_s|\chi_{(s\geq t)})^p ds \right)^{1/p}$$

and

$$|Z_T - Z'_T| \leq |z - z'| + \left(\int_t^T |X_s - X'_s|^p ds \right)^{1/p}$$

and the conclusion follows easily.

(ii) For the proof, see the Appendix. □

1.3. The HJB equation for u^∞. In this section, we analyze what becomes of the HJB equation (13) as we let p go to ∞. Since limit operations with viscosity solutions will be used throughout this paper, we start by giving the main definitions and the basic theorem related to the so-called half-relaxed limits. We use the standard following notations: for any sequence of functions F^p, define

$$\limsup_p{}^* F^p(\xi) = \limsup_{\substack{\xi' \to \xi \\ p \to \infty}} F^p(\xi'),$$

and

$$\liminf_{p*} F^p(\xi) = \liminf_{\substack{\xi' \to \xi \\ p \to \infty}} F^p(\xi').$$

Following is the basic theorem.

THEOREM 1.3.1. *Let $(u^p)_{p\in\mathbb{N}}$ be a sequence of uniformly locally bounded viscosity solutions of the equation*

$$G^p(x, u, Du, D^2u) = 0 \quad \text{in } \bar{\Omega},$$

where Ω is some domain in $\mathbb{R}^N$ and $(G^p)_p$ is a sequence of uniformly locally bounded functions defined in $\bar{\Omega} \times \mathbb{R} \times \mathbb{R}^N \times \mathcal{S}^n$ and satisfying

$$G^p(x, u, p, M) \leq G^p(x, u, p, N) \quad \text{if } M \geq N,$$

for any $x \in \bar{\Omega}$, $u \in \mathbb{R}$, $p \in \mathbb{R}^N$, $M, N \in \mathcal{S}^n$. Then $u^* = \limsup^* u^p$ (respectively, $u_* = \liminf_* u^p$) is a viscosity subsolution (respectively, supersolution) of $G_* = 0$ (respectively, $G^* = 0$), where $G_* = \liminf_* G^p$ (respectively, $G^* = \limsup^* G^p$).

We refer to [11] or [5] for a detailed presentation of this method of passage to the limit and, in particular, for the proof of Theorem 1.3.1, based on ideas introduced in [2].

After these prerequisites, let us show how this result applies to the problem we are dealing with. First we must compute

$$\mathcal{H}^* = \limsup^* \mathcal{H}^p \quad \text{and} \quad \mathcal{H}_* = \liminf_* \mathcal{H}^p.$$

This is done by carefully analyzing each case. First we do it for $t < T$. Keeping the notation of the previous section, we get

$$\text{if } |x| < z, \quad \mathcal{H}^* = \mathcal{L} \quad \text{and} \quad \mathcal{H}_* = \mathcal{L},$$

$$\text{if } |x| = z \begin{cases} \text{and} & -q_z < 0, & \mathcal{H}^* = \mathcal{L} & \text{and} & \mathcal{H}_* = -\infty, \\ \text{and} & -q_z = 0, & \mathcal{H}^* = +\infty & \text{and} & \mathcal{H}_* = -\infty, \\ \text{and} & -q_z > 0, & \mathcal{H}^* = +\infty & \text{and} & \mathcal{H}_* = \mathcal{L}, \end{cases}$$

$$\text{if } |x| > z \begin{cases} \text{and} & -q_z < 0, & \mathcal{H}^* = -\infty & \text{and} & \mathcal{H}_* = -\infty, \\ \text{and} & -q_z = 0, & \mathcal{H}^* = +\infty & \text{and} & \mathcal{H}_* = -\infty, \\ \text{and} & -q_z > 0, & \mathcal{H}^* = +\infty & \text{and} & \mathcal{H}_* = +\infty. \end{cases}$$

Now, let

$$u^* = \limsup{}^* u^p \quad \text{and} \quad u_* = \liminf{}_* u^p$$

(we will show in the next section that actually $u^* = u_* = u^\infty$). From Theorem 1.3.1 we know that u^* is a *sub*solution of $\mathcal{H}_*$ and u_* is a *super*solution of $\mathcal{H}^*$. Moreover, we see that

$$\mathcal{H}_*\left(x, z, t, \frac{\partial u}{\partial t}, D_x u, D_z u, D^2_{xx} u\right) \le 0$$

$$\iff \begin{cases} \mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right) \le 0 & \text{if } |x| < z, \\ \min\left\{\mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right), -D_z u\right\} \le 0 & \text{if } |x| = z, \\ -D_z u \le 0 & \text{if } |x| > z, \end{cases}$$

and

$$\mathcal{H}^*\left(x, z, t, \frac{\partial u}{\partial t}, D_x u, D_z u, D^2_{xx} u\right) \ge 0$$

$$\iff \begin{cases} \mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right) \ge 0 & \text{if } |x| < z, \\ \max\left\{\mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right), -D_z u\right\} \ge 0 & \text{if } |x| = z, \\ -D_z u \ge 0 & \text{if } |x| > z, \end{cases}$$

with all inequalities to be taken in the viscosity sense. For $t = T$, exactly the same computations yield to

$$\begin{cases} \min\left\{u - \psi, \mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right)\right\} \le 0 & \text{if } |x| < z, \\ \min\left\{u - \psi, \mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right), -D_z u\right\} \le 0 & \text{if } |x| = z, \\ \min\{u - \psi, -D_z u\} \le 0 & \text{if } |x| > z, \end{cases}$$

and

$$\begin{cases} \max\left\{u - \psi, \mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right)\right\} \ge 0 & \text{if } |x| < z, \\ \max\left\{u - \psi, \mathcal{L}\left(x, t, \dfrac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right), -D_z u\right\} \ge 0 & \text{if } |x| = z, \\ \max\{u - \psi, -D_z u\} \ge 0 & \text{if } |x| > z, \end{cases}$$

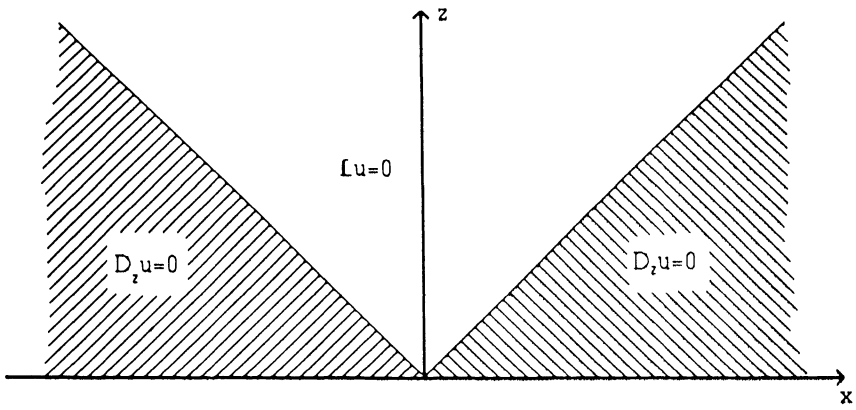


FIG. 1.

The results of [4] imply that, since the "$\mathcal{L}$" inequalities cannot hold at $t = T$ (essentially because the (x, z, t) dynamic exits through this boundary), these properties reduce to

$$\begin{cases} u - \psi \leq 0 & \text{if } |x| < z, \\ \min\{u - \psi, -D_z u\} \leq 0 & \text{if } |x| \geq z, \end{cases}$$

and

$$\begin{cases} u - \psi \geq 0 & \text{if } |x| < z, \\ \max\{u - \psi, -D_z u\} \geq 0 & \text{if } |x| \geq z, \end{cases}$$

Now let

$$\begin{aligned} \Omega &= \{(x, z) \text{ s.t. } |x| < z\} \times (0, T), \\ \bar{\Omega} &= \{(x, z) \text{ s.t. } |x| \leq z\} \times [0, T], \\ \partial_0\Omega &= \{(x, z) \text{ s.t. } |x| = z\} \times [0, T]. \end{aligned}$$

(See Fig. 1.) Readers familiar with viscosity solutions will recognize on $\partial_0\Omega$ a Neumann—or, more precisely, an oblique derivative—boundary condition in the viscosity sense (see [11]). We show now that this is indeed the case.

THEOREM 1.3.2. *u^* (respectively u_*) is a viscosity subsolution (respectively, supersolution) of*

$$\mathcal{L}\left(x, t, \frac{\partial u}{\partial t}, D_x u, D^2_{xx} u\right) = 0 \quad \text{in } \Omega, \tag{15}$$

$$-D_z u = 0 \quad \text{on } \partial_0\Omega, \tag{16}$$

$$u(x, z, T) = \psi(z) \quad \text{in } \{|x| < z\}. \tag{17}$$

Proof of Theorem 1.3.2. We only prove the result in the subsolution case and for $t < T$, the other cases being treated by similar arguments. Let $\phi \in C^2(\bar{\Omega})$ and $(\bar{x}, \bar{z}, \bar{t})$ a strict global maximum point of $(u^* - \phi)$. If $(\bar{x}, \bar{z}, \bar{t}) \in \Omega$, there is no problem, so we will suppose that $(\bar{x}, \bar{z}, \bar{t}) \in \partial_0\Omega$.

Consider a $C^2(\bar{Q})$-extension of ϕ, still denoted by ϕ. Set, for $k > 0$,

$$\tilde{\phi}_k(x,z,t) = \phi(x,z,t) + k(|x| - z)^{+4},$$

and let (x_k, z_k, t_k) be a maximum point of $u^* - \tilde{\phi}_k$. Since $(\bar{x}, \bar{z}, \bar{t})$ is a global strict maximum point of $u^* - \phi$ in $\bar{\Omega}$ and since $(|x| - z)^+ \equiv 0$ in Ω, then (x_k, z_k, t_k) necessarily lies in $\bar{Q}\backslash\Omega$. Moreover, if we let $k \to \infty$, then (x_k, z_k, t_k) tends to $(\bar{x}, \bar{z}, \bar{t})$.

If there are infinitely many $(x_k, z_k, t_k) \notin \partial_0\Omega$, then, extracting if necessary a subsequence, we have $-D_z\tilde{\phi}_k(x_k, z_k, t_k) \leq 0$, because u^* is a subsolution of $\mathcal{H}_*$. Now we notice that $D_z\tilde{\phi}_k \leq D_z\phi$. Thus, letting $k \to \infty$, from the continuity of $D\phi$, we get $D_z\phi(\bar{x}, \bar{z}, \bar{t}) \geq 0$.

Otherwise, we extract a subsequence that remains on $\partial_0\Omega$, and conclude by passing to the limit in the inequality $\min\{\mathcal{L}\phi, -D_z\phi\} \leq 0$. □

We now state the uniqueness theorem.

THEOREM 1.3.3. *Let u (respectively, v) be an upper semicontinuous (u.s.c.) subsolution (respectively, a lower semicontinuous (l.s.c.) supersolution) of* (15)–(17). *Then*

$$u \leq v \quad \text{in } \bar{\Omega}.$$

This result is proved in Dupuis and Ishii [16], [17], where uniqueness results for oblique derivative problems in domains with corners are obtained in a more general setting. In our case, the proof can be highly simplified because of the particular form of our problem, but we skip it, however, since the arguments to obtain it are easy.

By standard arguments (cf. [11]), Theorems 1.3.2 and 1.3.3 imply Theorem 1.3.4.

THEOREM 1.3.4. *u^p converges as $p \to \infty$ locally uniformly to the unique* BUC *solution $u = u^* = u_*$ of* (15)–(17).

1.4. Convergence of u^p as $p \to \infty$. The goal of this section is to show that u is actually u^∞. This will complete the proof that u^∞ is continuous and the only solution of its related HJB equation, namely (15)–(17). The following theorem states our result.

THEOREM 1.4.1. $\lim_{p\to\infty} u^p = u^\infty$ *uniformly on compact subsets of* $\bar{\Omega}$.

The main step in the proof of this result is the following lemma.

LEMMA 1.4.1. *Let $(Z_s^p)_s$ and $(Z_s^\infty)_s$ be, respectively, given by* (6) *and* (7), *then we have*

$$E(|Z_T^p - Z_T^\infty|) \to 0 \tag{18}$$

uniformly on compact subsets of $\bar{\Omega}$ and uniformly in $(\alpha_s)_s$. In particular,

$$E(|\|X\|_{L^p(t,T)} - \|X\|_{L^\infty(t,T)}|) \to 0, \tag{19}$$

uniformly on compact subsets of $\mathbb{R}^n$ and uniformly in $(\alpha_s)_s$.

Proof of Theorem 1.4.1. We first complete the proof of Theorem 1.4.1 by using Lemma 1.4.1. For $(x,z,t) \in \bar{\Omega}$, we estimate $u^p - u^\infty$,

$$\begin{aligned} |u^p(x,z,t) - u^\infty(x,z,t)| &= |\inf_{(\alpha_s)_s} E[\psi(Z_T^p)] - \inf_{(\alpha_s)_s} E[\psi(Z_T^\infty)]| \\ &\leq \sup_{(\alpha_s)_s} E|\psi(Z_T^p) - \psi(Z_T^\infty)|, \end{aligned}$$

and by the Lipschitz continuity of ψ, we have

$$|u^p(x,z,t) - u^\infty(x,z,t)| \leq C \sup_{(\alpha_s)_s} E|Z_T^p - Z_T^\infty|, \tag{20}$$

and we conclude by using Lemma 1.4.1. □

Proof of Lemma 1.4.1. We first want to point out that the proof below uses only purely PDE arguments. To prove (18), we use the device that consists in splitting the variables and we introduce another control problem: for $p, q \in (1, \infty), \xi = (x, z, z', t) \in D = \mathbb{R}^n \times \mathbb{R}_+^* \times \mathbb{R}_+^* \times [0, T]$, let X be the solution of (4) subject to initial condition $X_t = x$. Then we consider the control problem

$$\Upsilon^{p,q}(\xi) = \sup_{(\alpha_s)_s} E[|Z_T^p - Z_T^q|]. \tag{21}$$

Now let $\Upsilon^* = \limsup^*_{p,q\to\infty} \Upsilon^{p,q}$ and set

$$\mathcal{O} = \{\xi \text{ such that } |x| < z \text{ and } |x| < z'\}.$$

We want to show the following inequality:

$$\Upsilon^*(\xi) \le |z - z'| \quad \text{on } \bar{\mathcal{O}}. \tag{22}$$

This will prove Lemma 1.4.1: indeed, since $\Upsilon^* \ge 0$, this inequality implies

$$\Upsilon^*(x, z, z, t) = 0 \quad \text{in } \mathbb{R}^n \times \mathbb{R}_+ \times [0, T],$$

and (18) is an easy consequence of this property. To prove (22), we follow the method already used in §1.3.

Arguing as in the proof of Theorem 1.2.1 we easily show that $\Upsilon^{p,q}$ is Lipschitz continuous and is a viscosity solution of

$$\begin{gathered} \tilde{\mathcal{L}}u - \tfrac{z}{p}\left(\tfrac{|x|}{z}\right)^p D_z u - \tfrac{z'}{q}\left(\tfrac{|x'|}{z'}\right)^q D_{z'} u = 0 \quad \text{in } D, \\ u(x, z, z', T) = |z - z'| \quad \text{in } \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}_+, \end{gathered}$$

where

$$\tilde{\mathcal{L}}u = -\frac{\partial u}{\partial t} + \inf_{\alpha \in A}\left\{-\frac{1}{2} Tr(\sigma\sigma^T(x, t, \alpha) D_{xx}^2 u) - b(x, t, \alpha) D_x u\right\}. \tag{23}$$

Arguing along the lines of §1.3, Υ^* is a viscosity subsolution of

$$\tilde{\mathcal{L}}u = 0 \quad \text{in } \mathcal{O}, \tag{24}$$

$$-D_z u = 0 \quad \text{if } |x| = z, \tag{25}$$

$$-D_{z'} u = 0 \quad \text{if } |x| = z'. \tag{26}$$

Moreover, Υ^* has a sublinear growth at infinity since

$$E(Z_T^p), E(Z_T^q) \le C(1 + |x| + |z| + |z'|)$$

for some constant C depending only on b and σ. Now we remark that $w(x, z, z', t) = |z - z'|$ is a supersolution of (24)–(26) and so are the regularizations $w^\delta(x, z, z', t) = (|z - z'|^2 + \delta^2)^{1/2}$ of w. To actually show that $\Upsilon^* \le w$ in $\mathcal{O}$, we introduce the function

$$\begin{aligned} \Phi(x, z, z', t) = \Upsilon^*(x, z, z', t) - w^\delta(x, z, z', t) - \alpha e^{\gamma(T-t)}(|x|^2 + 1) + \cdots \\ + \eta[\phi(z) + \phi(z')] - \beta[|z|^2 + |z'|^2], \end{aligned}$$

where $\alpha, \beta, \delta, \eta$ are small parameters devoted to tend to zero and chosen later. Then, γ is a constant large enough chosen in such a way to have

$$\tilde{\mathcal{L}}(\alpha e^{\gamma(T-t)}(|x|^2+1)) > 0 \quad \text{in } \mathcal{O}, \tag{27}$$

and finally ϕ is a bounded C^1 strictly increasing function in $\mathbb{R}$. Since Υ^* has a sublinear growth at infinity, as soon as $\alpha, \beta > 0$, the maximum of Φ is achieved at some point $\bar{\xi} = (\bar{x}, \bar{z}, \bar{z}', \bar{t})$. Our goal is to show that $\bar{t} = 0$ for a suitable choice of the parameters α, β, and η. Since Υ^* is a viscosity subsolution of (24), $\bar{\xi} \notin \mathcal{O}$ because of (27). Then, if $|\bar{x}| = \bar{z}$, since the $\tilde{\mathcal{L}}$-inequality cannot hold again because of (27), we have necessarily

$$-D_z w^\delta(\bar{x}, \bar{z}, \bar{z}', \bar{t}) + \eta\phi'(\bar{z}) - 2\beta\bar{z} \leq 0. \tag{28}$$

However, because of the sublinear growth of Υ^*, $|\bar{x}| = O(1/\alpha)$ and therefore $|\bar{z}| = O(1/\alpha)$. If we fix $\eta > 0$, then for β small enough $\eta\phi'(\bar{z}) - 2\beta\bar{z} > 0$. Moreover,

$$-D_z w^\delta(\bar{x}, \bar{z}, \bar{z}', \bar{t}) = \frac{\bar{z}' - \bar{z}}{w^\delta(\bar{x}, \bar{z}, \bar{z}', \bar{t})} = \frac{\bar{z}' - \bar{x}}{w^\delta(\bar{x}, \bar{z}, \bar{z}', \bar{t})} \geq 0.$$

Finally, (28) cannot hold. Hence $\bar{t} = 0$ and we have

$$\begin{aligned}\Upsilon^*(x, z, z', t) \leq w^\delta(x, z, z', t) - \alpha \exp^{\gamma(T-t)}(|x|^2+1) + \cdots \\ + \eta[\phi(z) + \phi(z')] - \beta[|z|^2 + |z'|^2] + 2\eta\|\phi\|_\infty,\end{aligned}$$

And letting successively β, η, α, and finally δ to zero gives the desired conclusion. □

Remark. We want to point out that we do not use any comparison argument for the problem (24)–(26), and the fact that the domain $\mathcal{O}$ exhibits corners therefore does not really matter.

2. The general case. In this part, we consider the case where the cost function is given by

$$J(x, t, (\alpha_s)_s) = E\left[\psi\left(\sup_{s\in[t,T]} f(X_s, s, \alpha_s)\right)\right]. \tag{29}$$

We recall that $f \geq 0$ is assumed to satisfy Assumption 1 and throughout this section we also assume the following assumption holds

ASSUMPTION 2.

$$\{f(x, t, \alpha); \alpha \in A\} = [f^-(x, t), f^+(x, t)]$$

for any $x \in \mathbb{R}^n, t \in [0, T]$ *and where* $f^+(x, t) = \sup_{\alpha\in A} f(x, t, \alpha), f^-(x, t) = \inf_{\alpha\in A} f(x, t, \alpha)$.

At the end of the section we will explain why this assumption is necessary and, in some sense, natural in our approach.

Since most of the arguments developed in §1 are similar to those used here, we will only focus on the main subtleties this problem requires compared to the preceding one.

The new state variable must clearly be as follows: if $p < \infty$,

$$Z_s^p = \left(z^p + \int_t^s f(X_\tau, \tau, \alpha_\tau)^p d\tau\right)^{1/p} \tag{30}$$

and, if $p = \infty$,

$$Z_s^\infty = \max\left\{ z, \sup_{\tau\in[t,s]} f(X_\tau, \tau, \alpha_\tau)\right\}, \tag{31}$$

X being as always the solution of (4), (5). The cost function of the augmented problem is again given by

$$J^p(x, z, t, (\alpha_s)_s) = E[\psi(Z_T^p)] \tag{32}$$

We want to minimize J^p over all admissible controls

$$u^p(x, z, t) = \inf_{(\alpha_s)_s} J^p(x, z, t, \alpha). \tag{33}$$

Differentiating (30) with respect to time yields

$$dZ_s = \frac{Z_s}{p}\left(\frac{f(X_s, s, \alpha_s)}{Z_s}\right)^p.$$

Thus, if we set, for $\xi = (x, z, t, q_t, q_x, q_z, M) \in Q \times \mathbb{R}^{n+2} \times \mathcal{S}^n$ and $\alpha \in A$,

$$\mathcal{L}^\alpha(\xi) = -q_t - \tfrac{1}{2}\operatorname{Tr}((\sigma\sigma^T)(x, t, \alpha)M) - b(x, t, \alpha)q_x,$$

the Hamiltonian related to the control problem (33) is, for $p < \infty$,

$$\mathcal{H}^p(\xi) = \sup_{\alpha\in A}\left\{\mathcal{L}^\alpha(\xi) - \frac{z}{p}\left(\frac{f(x, s, \alpha)}{z}\right)^p q_z\right\}.$$

We state without proof the following result, which is a straightforward generalization of Theorem 1.2.1.

THEOREM 1.2.1. *Under Assumption* 1, *we have*

(i) u^p *is bounded and Lipschitz continuous in* x *and* z, *uniformly for* $t \in [0, T]$.

(ii) u^p *is the unique bounded viscosity solution of*

$$\mathcal{H}^p\left(x, z, t, \frac{\partial u}{\partial t}, D_x u, D_z u, D_{xx}^2 u\right) = 0 \quad \text{in } Q,$$
$$u(x, z, T) = \psi(z) \quad \text{in } \mathbb{R}^n \times \mathbb{R}_+^*.$$

Looking at the equation, which is a priori satisfied by u^∞, and showing that actually u^p tends to u^∞ is here a little more difficult than in §1, mostly because of the dependency of the cost with respect to α.

Again we take the half-relaxed limits

$$\mathcal{H}^* = \limsup{}^* \mathcal{H}^p \quad \text{and} \quad \mathcal{H}_* = \liminf{}_* \mathcal{H}^p$$

and

$$u^* = \limsup{}^* u^p \quad \text{and} \quad u_* = \liminf{}_* u^p.$$

Define

$$A_{x,z,t} = \{\alpha \in A \text{ such that } f(x, t, \alpha) \le z\},$$

TABLE 1

	$q_z < 0$	$q_z = 0$	$q_z > 0$
$z > f^+(x,t)$	$\mathcal{L}(x,z,t)$	$\mathcal{L}(x,z,t)$	$\mathcal{L}(x,z,t)$
$z = f^+(x,t)$	$+\infty$	$+\infty$	$\mathcal{L}(x,z,t)$
$f^- < z < f^+$	$+\infty$	$+\infty$	$\mathcal{L}(x,z,t)$
$z = f^-(x,t)$	$+\infty$	$+\infty$	$\mathcal{L}(x,z,t)$
$z < f^-(x,t)$	$+\infty$	$+\infty$	$-\infty$

TABLE 2

	$q_z < 0$	$q_z = 0$	$q_z > 0$
$z > f^+(x,t)$	$\mathcal{L}_*(x,z,t)$	$\mathcal{L}_*(x,z,t)$	$\mathcal{L}_*(x,z,t)$
$z = f^+(x,t)$	$\mathcal{L}_*(x,z,t)$	$\mathcal{L}_*(x,z,t)$	$\mathcal{L}_*(x,z,t)$
$f^- < z < f^+$	$+\infty$	$\mathcal{L}_*(x,z,t)$	$\mathcal{L}_*(x,z,t)$
$z = f^-(x,t)$	$+\infty$	$-\infty$	$-\infty$
$z < f^-(x,t)$	$+\infty$	$-\infty$	$-\infty$

and

$$A'_{x,z,t} = \{\alpha \in A \text{ such that } f(x,t,\alpha) < z\};$$

finally we set

$$\mathcal{L}(x,z,t) = \sup_{\alpha \in A_{x,z,t}} \mathcal{L}^\alpha.$$

Again we use in this part simplified notation by dropping the dependency of $\mathcal{L}$ with respect to the derivatives of u. It is clear enough that $\mathcal{L}$ is u.s.c. and that the l.s.c. envelope of $\mathcal{L}$ is given by

$$\mathcal{L}_*(x,z,t) = \sup_{\alpha \in A'_{x,z,t}} \mathcal{L}^\alpha.$$

We detailed only the situation when $t < T$. A simple (although not immediate) computation gives the values of $\mathcal{H}^*$ and $\mathcal{H}_*$. Table 1 gives the values for $\mathcal{H}^*$. Table 2 gives the values for $\mathcal{H}_*$.

We know by Theorem 1.3.1 that u^* (respectively, u_*) is a subsolution (respectively, supersolution) of $\mathcal{H}_* \leq 0$ (respectively, $\mathcal{H}^* \geq 0$). Table 3 gives those values.

For $t = T$, by using arguments analogous to the ones of the first section, we have in the viscosity sense

$$\max\{u(x,z,T) - \psi(z), -D_z u\} = 0 \quad \text{in } \{z \leq f^+(x,T)\},$$

and

$$u(x,z,T) = \psi(z) \quad \text{in } \{z > f^+(x,T)\}.$$

We set

$$\begin{aligned}
\Omega_1 &= \{(x,z,t) \text{ such that } z > f^+(x,t)\},\\
\Omega_2 &= \{(x,z,t) \text{ such that } f^-(x,t) \leq z \leq f^+(x,t)\},\\
\Omega &= \bar{\Omega}_1 \cup \Omega_2,\\
\partial_0\Omega &= \{(x,z,t) \text{ such that } z = f^-(x,t)\}.
\end{aligned}$$

TABLE 3

	$\mathcal{H}^* \geq 0$	$\mathcal{H}_* \leq 0$
$z > f^+(x,t)$	$\mathcal{L}(x,z,t) \geq 0$	$\mathcal{L}_*(x,z,t) \leq 0$
$z = f^+(x,t)$	$\max\{\mathcal{L}(x,z,t), -q_z\} \geq 0$	$\mathcal{L}_*(x,z,t) \leq 0$
$f^- < z < f^+$	$\max\{\mathcal{L}(x,z,t), -q_z\} \geq 0$	$\max\{\mathcal{L}_*(x,z,t), -q_z\} \leq 0$
$z = f^-(x,t)$	$\max\{\mathcal{L}(x,z,t), -q_z\} \geq 0$	$-q_z \leq 0$
$z < f^-(x,t)$	$-q_z \geq 0$	$-q_z \leq 0$

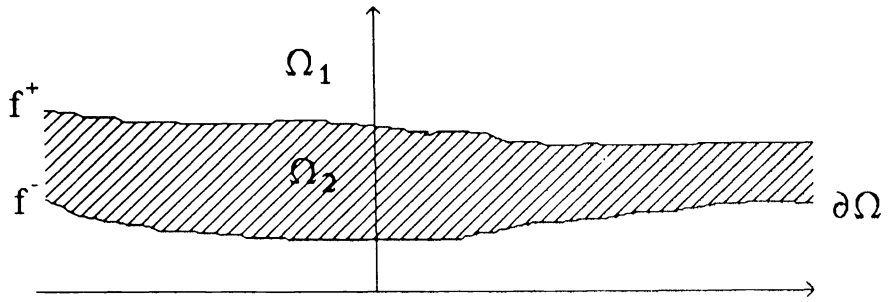


FIG. 2.

See Fig. 2.

Our first result consists in showing that we have indeed an oblique derivative boundary condition on $\partial_0\Omega$.

THEOREM 2.2. *u^* (respectively, u_*) is a viscosity subsolution (respectively, supersolution) of the mixed PDE-VI problem*

$$\sup_{\alpha\in A} \mathcal{L}^\alpha u = 0 \quad \text{in } \Omega_1, \tag{34}$$

$$\max\{\sup_{\alpha\in A_{x,z,t}} \mathcal{L}^\alpha u, -D_z u\} = 0 \quad \text{in } \Omega_2, \tag{35}$$

$$-D_z u = 0 \quad \text{on } \partial_0\Omega, \tag{36}$$

$$\max\{u(x,z,T) - \psi(z), -D_z u\} = 0 \quad \text{in } \{f^-(x,T) \leq z \leq f^+(x,T)\}, \tag{37}$$

$$u(x,z,T) = \psi(z) \quad \text{in } \{z > f^+(x,T)\}. \tag{38}$$

Proof of Theorem 2.2. Again we only prove the result in the subsolution case and for $t < T$, the other cases being treated by similar arguments. Let $\phi \in C^2(\bar{\Omega})$ and let $(\bar{x},\bar{z},\bar{t})$ be a strict global maximum point of $(u^* - \phi)$. If $(\bar{x},\bar{z},\bar{t}) \in \Omega$, there is no problem, so we will suppose that $(\bar{x},\bar{z},\bar{t}) \in \partial_0\Omega$. The additional difficulty here is that $\partial_0\Omega$ is just a Lipschitz continuous boundary; therefore we need to introduce a sequence of smooth approximations $(f_\varepsilon^-)_\varepsilon > 0$ built in such a way to have $f_\varepsilon^- \leq f$.

Consider a $C^2(\bar{Q})$-extension of ϕ, still denoted by ϕ. Set, for $k > 0$ and $\varepsilon > 0$,

$$\tilde{\phi}_k^\varepsilon(x,z,t) = \phi(x,z,t) + k(f_\varepsilon^-(x,t) - z)^{+4},$$

and let $(x_k^\varepsilon, z_k^\varepsilon, t_k^\varepsilon)$ be a maximum point of $u^* - \tilde{\phi}_k^\varepsilon$. Since $(\bar{x},\bar{z},\bar{t})$ is a global strict maximum point of $u^* - \phi$ in $\bar{\Omega}$ and since $(f_\varepsilon^-(x,t) - z)^+ \equiv 0$ in Ω (recall that $f_\varepsilon^- \leq f$), then $(x_k^\varepsilon, z_k^\varepsilon, t_k^\varepsilon)$

necessarily lies in $\bar{Q}\backslash\Omega$. Moreover, if we let $\varepsilon \to 0$ and then $k \to \infty$, $(x_k^\varepsilon, z_k^\varepsilon, t_k^\varepsilon)$ tends to $(\bar{x}, \bar{z}, \bar{t})$.

If there are infinitely many $(x_k^\varepsilon, z_k^\varepsilon, t_k^\varepsilon) \notin \partial_0\Omega$, then, extracting if necessary a subsequence, we have $-D_z\tilde{\phi}_k(x_k^\varepsilon, z_k^\varepsilon, t_k^\varepsilon) \leq 0$, because u^* is a subsolution of $\mathcal{H}_*$. Now we note that $D_z\tilde{\phi}_k^\varepsilon \leq D_z\phi$. Thus, letting $k \to \infty$, from the continuity of $D\phi$, we get $D_z\phi(\bar{x}, \bar{z}, \bar{t}) \geq 0$.

Otherwise, we extract a subsequence that remains on $\partial_0\Omega$; noting that, in this case, $(f_\varepsilon^-(x_k^\varepsilon, t_k^\varepsilon) - z_k^\varepsilon)^+ = 0$, we conclude by passing to the limit in the inequality $\min\{\mathcal{L}\phi, -D_z\phi\} \geq 0$. □

Now we turn to our uniqueness result for this problem,

THEOREM 2.3. *Assume that Assumption* 1 *holds. If* u *is an u.s.c. bounded subsolution and* v *is a l.s.c. bounded supersolution of* (34)–(38) *such that* $u(T) \leq v(T)$. *Then*

$$u \leq v \quad \text{in } \bar{\Omega}.$$

Proof of Theorem 2.3. We face here two types of difficulties: the first one is connected to the variational inequality feature while the second one is related to the a priori possible discontinuity of the Hamiltonian $\sup_{\alpha\in A_{x,z,t}} \mathcal{L}^\alpha$ in z. The main remark that allows us to solve the second difficulty is that the Hamiltonian is in fact increasing in z.

Let $\phi : [0, \infty) \to \mathbb{R}$ be bounded, strictly decreasing and such that

$$\phi'(z) = -1 \quad \text{for all } 0 \leq z \leq K = \sup_{x,t,\alpha} f(x, t, \alpha) + 1.$$

For $\beta, \gamma, \delta, \epsilon > 0$, consider the test function

$$\Phi(x, y, z, w, t, s) = \beta(|x|^2 + |y|^2 + (w - K)^{+2}) + \gamma(\phi(z) + \phi(w)) + \exp(C(T - t))\left(\frac{|x - y|^2}{\epsilon} + \frac{|t - s|^2}{\epsilon}\right) + \frac{|z - w|^2}{\delta},$$

where C is some constant large enough and let (x, y, z, w, t) be a maximum point of $u(x, z, t) - v(y, w, s) - \Phi(x, y, z, w, t, s)$.

Case 1. $(x, z) \in \Omega_2$.

Since u is a viscosity subsolution of (34)–(38), we have $\max\{\mathcal{L}(x, z, t)\Phi, -D_z\Phi\} \leq 0$. A straightforward computation gives

$$-D_z\Phi = -\frac{2(z - w)}{\delta} - \gamma\phi'(z) \leq 0;$$

thus, if K_1 is a bound on the Lipschitz constant of f^+,

$$K_1(|x - y| + |t - s|) + f^+(y, s) \geq f^+(x, t) \geq z \geq w - \frac{\delta\gamma}{2}\phi'(z) \geq w + \frac{\delta\gamma}{2}. \tag{39}$$

We recall that $|x - y|, |t - s| = o(\sqrt{\epsilon})$ and $|z - w| = o(\sqrt{\delta})$; therefore, choosing $\epsilon \ll \delta$, we can always suppose the inequality

$$|x - y| \leq -K_1^{-1}\frac{\delta\gamma}{4}\phi'(z) \tag{40}$$

to hold. Then, from (39) and (40), we get $f^+(y, t) > w$; therefore $(y, w) \in \Omega_2$ and $(w - K)^+ = 0$. However,

$$-(-D_w\Phi) = -\frac{2(z - w)}{\delta} + \gamma\phi'(w) = -D_z\Phi + \gamma\phi'(z) + \gamma\phi'(w) < 0,$$

and this inequality implies that the $\mathcal{L}$-inequality holds for v. Thus, proceeding as in the Appendix, we obtain

$$u(x,z,t) - v(y,w,t) \le (I)$$

where

$$(I) = - \sup_{\alpha \in A_{x,z,t}} \left\{ -\tfrac{1}{2} \operatorname{Tr}(\sigma\sigma^T)(x,t,\alpha) M - b(x,t,\alpha) D_x \Phi \right\} + \sup_{\beta \in A_{y,w,s}} \left\{ -\tfrac{1}{2} \operatorname{Tr}(\sigma\sigma^T)(y,t,\beta) N + b(y,t,\beta) D_y \Phi \right\}$$

Next observe that, if $K_1 \ge |f|_{0,1}$, then

$$z \ge w + K_1(|x-y| + |t-s|) \Rightarrow A'_{x,z,t} \supseteq A_{y,w,s}. \tag{41}$$

Hence (39) and our choice of the parameters imply that (41) holds and therefore

$$(I) \le - \sup_{\alpha \in A_{y,w,s}} \left\{ -\tfrac{1}{2} \operatorname{Tr}(\sigma\sigma^T)(x,t,\alpha) M - b(x,t,\alpha) D_x \Phi \right\} + \sup_{\beta \in A_{y,w,s}} \left\{ -\tfrac{1}{2} \operatorname{Tr}(\sigma\sigma^T)(y,t,\beta) N + b(y,t,\beta) D_y \Phi \right\}$$

and it is then a purely routine job to proceed as in Theorem 1.21(ii).

Case 2. $(x,z) \in \Omega_1$ and $((y,w) \in \Omega_1$ or $D_w \Phi < 0)$.

Then immediately

$$\mathcal{L}(x,z,t)\Phi \le 0 \quad \text{and} \quad \mathcal{L}(y,w,s)\Phi \ge 0,$$

and one proceeds as in Case 1, still choosing the parameters so that (40) holds.

Case 3. $(x,z) \in \Omega_1$ and $(y,w) \in \Omega_2$ and $D_w \Phi \ge 0$.

We are going to show that this case cannot happen. Since

$$D_w \Phi = -\frac{2(z-w)}{\delta} + \gamma \phi'(w) \ge 0,$$

and $(y,w) \in \Omega_2$,

$$z \le w - \frac{\delta\gamma}{2}.$$

However,

$$(x,z) \in \Omega_1 \Rightarrow z > f^+(t,x),$$
$$(y,w) \in \Omega_2 \Rightarrow w \le f^+(t,y).$$

By Lipschitz continuity of f^+,

$$K_1|x-y| \ge w - z \ge \frac{\delta\gamma}{2}.$$

However, this last inequality cannot hold since (40) does.

Theorem 2.4 concludes this part.

THEOREM 2.4. $\lim_{p\to\infty} u^p = u^\infty$ *uniformly on compact subsets of* $\bar{\Omega}$.

Proof of Theorem 2.4. The ideas of the proof of the analogous result in §1 do not extend here since it is false here that Z_T^p converges to Z_T^∞ *uniformly* with respect to $(\alpha_s)_s$; this is only true when we work with a fixed control. Therefore we will use different types of arguments. We proceed in two steps.

Step 1. $\lim_{p\to\infty} u^p \le u^\infty$.

By the very definition of u^p, we have

$$u^p(x,z,t) \le J^p(x,z,t,(\alpha_s)_s) = E[\psi(Z_T^p)]$$

for any control $(\alpha_s)_s$. We fix the control and we pass to the limit in this inequality by using the convergence of Z_T^p to Z_T^∞, we get

$$\lim_{p\to\infty} u^p(x,z,t) \le E[\psi(Z_T^\infty)].$$

Finally, taking the infimum over $(\alpha_s)_s$ in the right-hand side yields the desired inequality.

Step 2. $\lim_{p\to\infty} u^p \ge u^\infty$.

To obtain this result, we are going to prove that u^∞ is a viscosity subsolution of (34)–(38). Indeed, if this claim is true, the inequality will be an immediate consequence of Theorem 2.3 since $\lim_{p\to\infty} u^p$ is a viscosity solution of (34)–(38).

Now we prove the claim. The only difficulty comes from the region $f^-(x,t) \le z \le f^+(x,t)$. We must prove that in this region

$$-D_z u \le 0,$$

and

$$\mathcal{L}_*(x,z,t)\left(x,z,t,\frac{\partial u}{\partial t}, D_x u, D_z u, D^2_{xx} u\right) \le 0,$$

in the viscosity sense.

We are going to use a dynamic programming principle type argument. If we choose a control $\alpha_s \equiv \alpha$ on the time interval $(t, t+h), \alpha$ being also deterministic, then classical arguments (cf. Krylov [27]) implies

$$u^\infty(x,z,t) \le E_{x,z,t}[u^\infty(X_{t+h}, Z_{t+h}, t+h)]. \tag{42}$$

for any $h > 0$.

To prove that u^∞ is decreasing with respect to z if $z \le f^+(x,t)$, we consider $z < z' \le f^+(x,t)$. If $z' \le f^-(x,t)$, we have obviously $u(x,z,t) = u(x,z',t)$. Otherwise by Assumption 2 there exists $\alpha \in A$ such that $z' = f(x,t,\alpha)$. We use this constant control in (42); we obtain

$$u^\infty(x,z,t) \le E_{x,z,t}\left[u^\infty\left(X_{t+h}, \sup\left(z, \sup_{(t,t+h)} f(X_s,s,\alpha)\right), t+h\right)\right],$$

and we let h go to zero, $\sup(z, \sup_{(t,t+h)} f(X_s,s,\alpha))$ converges to $\sup(z, f(x,t,\alpha)) = z'$ and therefore

$$u^\infty(x,z,t) \le (u^\infty)^*(x,z',t).$$

Recall that we know that u^∞ is Lipschitz continuous in x and z but we do not know it is continuous in t. However, since the above inequality is true for any (x,z,t) it obviously yields $(u^\infty)^*(x,z,t) \le (u^\infty)^*(x,z',t)$ and therefore $-D_z u \le 0$.

Now if $z > f(x,t,\alpha)$, we still consider the constant control $\alpha_s \equiv \alpha$ on the interval $(t, t+h)$. By (42), we have

$$u^\infty(x,z,t) \leq E[u^\infty(X_{t+h}, Z_{t+h}, t+h)],$$

which implies

$$(u^\infty)^*(x,z,t) \leq E[(u^\infty)^*(X_{t+h}, Z_{t+h}, t+h)],$$

since the right-hand side is u.s.c. Classical arguments yield

$$\mathcal{L}^\alpha(x,z,t,\frac{\partial u}{\partial t}, D_x u, D_z u, D^2_{xx} u) \leq 0,$$

in the viscosity sense. Taking the infimum on $\alpha \in A'_{x,z,t}$ gives the result. □

We conclude this section by explaining why Assumption 2 is necessary. We consider the deterministic case when $b \equiv 0, \sigma \equiv 0, A = \{0,1\}, f(x,t,\alpha) = \alpha$, and $\psi(t) = \inf(t(t-1), 0)$. In this simple case, we can compute everything: consider z such that $0 < z < \frac{1}{4}$

$$u^\infty(x,z,t) = \inf(\psi(z), \psi(1)) = \psi(z),$$

since $Z_t = \sup(z,1) = z$, if we use the control $\alpha = 1, Z_t = z$ otherwise. On the other hand

$$u^p(x,z,t) = \inf_{s\in(0,T-t)} \psi((z^p+s)^{1/p})$$

and if $T-t$ is, say, larger than 1 there exists $s \in (0, T-t)$ such that $(z^p+s)^{1/p} = \frac{1}{2}$ and therefore

$$u^p(x,z,t) = \psi(1/2) = \inf_{\mathbb{R}} \psi.$$

Therefore $u^p(x,z,t)$ cannot converge to $u^\infty(x,z,t)$ since if $0 < z < \frac{1}{4}$ and if $T-t > 1$

$$u^\infty(x,z,t) = \psi(z) > \psi(\tfrac{1}{2}) = u^p(x,z,t).$$

The L^p problem appears as a "relaxed problem" of the L^∞ one; therefore the convexity assumption of the set $\{f(x,t,\alpha); \alpha \in A\}$ is natural in this context.

3. Extensions and applications.

3.1. Optimal stopping. All the results of §§1 and 2 can be extended readily to the case of optimal stopping time problems. We just formulate in this section the result corresponding to the problem of pricing American options on stocks that we use in the next section. We use the same notation as in the first part and we define for $(x,z,t) \in \Omega$, a control $(\alpha_s)_s$ and a stopping time θ such as $t \leq \theta \leq T$ almost surely, the cost function by

$$J(x,z,t,(\alpha_s)_s,\theta) = E[\psi(X_\theta, Z_\theta, \theta)],$$

where $\psi \in W^{1,\infty}(\mathbb{R}^n \times \mathbb{R}_+ \times [0,T])$. We recall that in this framework the control consists both in the process $(\alpha_s)_s$ and in the stopping time θ. The value function is then given by

$$u(x,z,t) = \inf_{(\alpha_s)_s,\theta} J(x,z,t,(\alpha_s)_s,\theta).$$

Our result is the following theorem.

THEOREM 3.1.1. *u is the unique viscosity of the variational inequality*

$$\max\left\{\mathcal{L}(x,z,t,\frac{\partial u}{\partial t}, D_x u, D^2_{xx} u), u-\psi\right\} = 0 \quad \text{in } \Omega, \tag{43}$$

$$\max\{-D_z u, u-\psi\} = 0 \quad \text{on } \partial_0\Omega, \tag{44}$$

$$u(x,z,T) = \psi(x,z,T) \quad \text{in } \{|x| < z\}. \tag{45}$$

3.2. Application in finance theory: path-dependent options. We show in this section how the results of the previous sections can be used to valuate look-back options. This problem was investigated also in [9]; but the martingale approach used therein, if it yields explicit formulas in the case of some European options with constant diffusion coefficients, does not allow a straightforward generalisation to the American case. The PDE approach we present here provides a very general method to treat this problem (and also the case of variable diffusion coefficients, or dividends, etc).

Let us first set up the model. We consider a stock which price $(P_t)_{t\in[0,T]}$ is given, *in the risk neutral probability*[2] by the Itô equation

$$dP_t = P_t(r(P_t,t)dt + \sigma(P_t,t)dW_t), \tag{46}$$

$$P_0 = p_0 > 0, \tag{47}$$

where $((W_t)_{t\in[0,T]}, (\mathcal{F}_t)_{t\in[0,T]})$ is a Brownian motion and r, σ are Lipschitz, bounded functions on $\mathbb{R}^*_+ \times [0,T]$. We recall that r stands for the risk-free rate available on the market and σ is the volatility parameter. We set, for $t_0 \le t \le T$, $M^{t_0}_t = \sup_{s\in[t_0,t]} P_s$.

In this set-up, an American option is given by an $\mathcal{F}$-adapted process $(\Psi_t)_{t\in[0,T]}$ that represents the cash flow to be paid if the option is exercised at time t. It can be shown that the "fair" price (with respect to finance theory) is given by the process $(U_t)_{t\in[0,T]}$ defined by

$$U_t = \sup_{\theta \text{ such that } t\le\theta\le T} E\left[\Psi_\theta \exp\left(-\int_t^\theta r(P_s,s)ds\right) \middle| \mathcal{F}_t\right], \tag{48}$$

where the supremum runs over all $\mathcal{F}$-stopping times θ such that $t \le \theta \le T$. The valuation problem is thus to determine this supremum.

In the standard case, Ψ_t is actually a deterministic function of P_t and of time (e.g., $\Psi_t = (P_t - E)^+$ for a standard American call with strike E), and we classically prove that $U_t = u(P_t,t)$ also depends only on P_t and t, u being a function that solves a variational inequality.

The so-called path-dependent options, whose financial background is described in the books of Ingersoll [19] and of Cox and Rubinstein [10], are those for which Ψ_t does not depend only on P_t, but more generally of its history $(P_s)_{s\in[0,t]}$. The cases we will consider here are those for Ψ_t depends on the supremum (in order to simplify, but one proceeds analogously for infimum and for L^p means). Let us give some examples, most of them being taken from [9].

1. Call on supremum. It is a call with a specified strike E in which the stock price is replaced by its running maximum

$$\Psi_t = \left(\sup_{s\in[0,t]} P_s - E\right)^+. \tag{49}$$

[2] For more information on finance theory, we refer to [15] or [25].

2. Put with lookback strike. A put where the strike is replaced by the running maximum

$$\Psi_t = \left(\sup_{s\in[0,t]} P_s - P_t \right)^+. \tag{50}$$

3. Lookback limited-risk put. It is like a standard lookback option, except that its strike cannot exceed a specified value M_0

$$\Psi_t = \left(\sup_{s\in[0,t]} P_s \wedge M_0 - P_t \right)^+. \tag{51}$$

We know that the fair price process is given by (48), but contrary to what happens in the standard case, U_t is no longer measurable with respect to P_t, and hence, to get a representation in terms of a deterministic function, we must add the relevant variable $M_t = \sup_{s\in[0,t]} P_s$.

It is then a purely routine job to apply the results of §1 to show that we have

$$U_t = u(P_t, M_t, t),$$

where, letting

$$\Omega = \{(p,m) \in (\mathbb{R}_+^*)^2 \text{ such that } p < m\} \times [0,T],$$

$$\partial_0\Omega = \{(p,m) \in (\mathbb{R}_+^*)^2 \text{ such that } p = m\} \times [0,T],$$

u is the (unique) solution of the variational inequality with oblique Neumann condition

$$\min\left\{ -\frac{\partial u}{\partial t} - \frac{1}{2}\sigma^2 p^2 \frac{\partial^2 u}{\partial p^2} - \beta p \frac{\partial u}{\partial p} + ru, u - \psi \right\} = 0 \quad \text{in } \Omega \tag{52}$$

$$\min\left\{ -\frac{\partial u}{\partial m}, u - \psi \right\} = 0 \quad \text{on } \partial_0\Omega, \tag{53}$$

$$u(p,m,T) = \psi(p,m,T). \tag{54}$$

The obstacle ψ is given, for example, by

$$\psi(p,m,t) = (m-E)^+ \quad \text{in case (49)}, \tag{55}$$

$$\psi(p,m,t) = (m-p)^+ \quad \text{in case (50)}, \tag{56}$$

$$\psi(p,m,t) = (m \wedge M_0 - p)^+ \quad \text{in case (51)}. \tag{57}$$

Remark. We let the reader give the corresponding formulations for European options, which is even simpler (the variational inequality is turned into a linear equation), for the case of *infimum* instead of supremum (the relevant domain is for $p \geq m$) and also the case of L^p *means* instead of extrema (see §1.1).

Let us observe that the value function u yields, as well,

- the price $u(p,m,t)$ of the option;

- the optimal exercise time θ_t after t given by the beginning time

$$\theta_t = \inf\{s \geq t \text{ such that } (P_s, \max\{M_s, m\}, s) \in \mathcal{S}\},$$

where $\mathcal{S}$ is the set

$$\mathcal{S} = \{(p, m, t) \text{ such that } u(p, m, t) = \psi(p, m, t)\};$$

- the heading ratio, that is to say the number of stocks to have in a heading portfolio of the option

$$\delta = -\frac{\partial u}{\partial p}.$$

Appendix. Proof of Theorem 1.2.1(ii). That u^p is a viscosity solution of (13) is merely a consequence of the dynamic programming principle applied to C^2 test functions as proved in [30].

Uniqueness can be treated by the general viscosity method for fully nonlinear second-order equations ([22], [21], [11]). We draw attention to the fact that we must handle an unbounded domain and show that boundedness is indeed a boundary condition for $z = 0$ and $|x|, z \to +\infty$. To take in account only the derivatives appearing in (13), we define simplified superjet and subjet for u at $(x, z, t) \in Q$,

$$\mathcal{J}^+u(x,z,t) = \left\{ \left(\frac{\partial \phi}{\partial t}, D_x\phi, D_z\phi, D^2_{xx}\phi \right) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathcal{S}^n, \right. \tag{58}$$
$$\left. \text{for } \phi \in C^2 \text{ such that } u - \phi \text{ has a local maximum at } (x,z,t) \right\}$$

and

$$\mathcal{J}^-u(x,z,t) = \left\{ \left(\frac{\partial \phi}{\partial t}, D_x\phi, D_z\phi, D^2_{xx}\phi \right) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \times \mathcal{S}^n, \right. \tag{59}$$
$$\left. \text{for } \phi \in C^2 \text{ such that } u - \phi \text{ has a local minimum at } (x,z,t) \right\}.$$

We take the following notation: for $(q_t, q_x, q_z, M) \in \mathbb{R}^{n+2} \times \mathcal{S}^n$ and $u \in \mathbb{R}$, let

$$F(x,z,t,u,q_x,q_z,M) = -q_t + \sup_{\alpha \in A} \left\{ -\frac{1}{2} \operatorname{Tr}(\sigma\sigma^T)(t,x,\alpha)M \right.$$
$$\left. - b(t,x,\alpha)q_x \right\} + u - \frac{z}{p} \left(\frac{|x|}{z} \right)^p q_z$$

(observing that $\mathcal{L}$ can be changed to $\mathcal{L} + I$ by the change of function $u \to e^{-t}u$). We recall that an u.s.c. function u (respectively, a l.s.c. function v) is said to be a viscosity subsolution (respectively, supersolution) of (13) if and only if, for all $(x, z, t) \in \mathbb{R}^n \times \mathbb{R}^*_+ \times (0, T)$,

$$(q_t, q_x, q_z, M) \in \mathcal{J}^+u(x,z,t) \quad (\text{respectively, } \in \mathcal{J}^-u(x,z,t))$$

implies

$$F(x,z,t,u(x,z,t),q_x,q_z,M) \leq 0 \quad (\text{respectively, } \geq 0).$$

Now let u be an u.s.c. bounded subsolution and v a l.s.c. bounded supersolution of (13) such that $u(T) \leq v(T)$. We want to prove that

$$u \leq v.$$

Assume $\max_{\bar{Q}}(u - v) > 0$ and let us show that this leads to a contradiction. As usual in viscosity solution theory, we approximate the above maximum by $\max_{Q\times Q} \psi(x, z, t, y, w, s)$, where

$$\psi(x, z, t, y, w, s) = u(x, z, t) - v(y, w, s) - \Phi(x, z, t, y, w, s)$$

with

$$\Phi(x, z, t, y, w, s) = \frac{(|t - s|^2 + |x - y|^2 + |z - w|^2)}{\varepsilon} - K(x, z, t) - K(y, w, s),$$

and

$$K(x, z, t) = \beta(|x|^2 + |z|^2) + \gamma\frac{1}{|z|^2} + \theta\frac{1}{t},$$

β, γ and θ being positive constants devoted to tend to zero. K is used to penalize the "irrelevant" boundaries (x at ∞, z at 0 and ∞, t at 0). The heart of the proof is the following property (see [11]). If $(\bar{x}, \bar{t}, \bar{z}, \bar{y}, \bar{s}, \bar{w})$ is a maximum point of ψ (such a point can only be in Q), then, for each $\varepsilon > 0$, there are matrices M, N such that

$$\left(\frac{\partial\Phi}{\partial t}, D_x\Phi, M, D_z\Phi\right) \in \bar{\mathcal{J}}^+ u(\bar{x}, \bar{t}, \bar{z}), \tag{60}$$

$$\left(-\frac{\partial\Phi}{\partial s}, -D_y\Phi, -N, -D_w\Phi\right) \in \bar{\mathcal{J}}^- v(\bar{y}, \bar{s}, \bar{w}), \tag{61}$$

$$\begin{pmatrix} M & 0 \\ 0 & N \end{pmatrix} \leq D^2_{xy}\Phi + \varepsilon(D^2_{xy}\Phi)^2, \tag{62}$$

where all derivatives are taken at $(\bar{x}, \bar{t}, \bar{z}, \bar{y}, \bar{s}, \bar{w})$.

By the definition of viscosity solution, this leads to

$$u(\bar{x}, \bar{z}, \bar{t}) - v(\bar{y}, \bar{w}, \bar{s}) \leq (I) + (II),$$

where

$$\begin{aligned}(I) = {} & \frac{\partial\Phi}{\partial t} + \frac{\partial\Phi}{\partial s} \\ & - \sup_{\alpha\in A}\left\{-\frac{1}{2}\operatorname{Tr}(\sigma\sigma^T)(\bar{t}, \bar{x}, \alpha)M - b(\bar{t}, \bar{x}, \alpha)D_x\Phi - f(\bar{t}, \bar{x}, \bar{z}, \alpha)\right\} \\ & + \sup_{\alpha\in A}\left\{-\frac{1}{2}\operatorname{Tr}(\sigma\sigma^T)(\bar{s}, \bar{y}, \alpha)N + b(\bar{s}, \bar{y}, \alpha)D_y\Phi - f(\bar{s}, \bar{y}, \bar{w}, \alpha)\right\}\end{aligned}$$

and

$$(II) = \frac{\bar{z}}{p}\left(\frac{|\bar{x}|}{\bar{z}}\right)^p D_z\Phi - \frac{\bar{w}}{p}\left(\frac{|\bar{y}|}{\bar{w}}\right)^p D_w\Phi.$$

Thus

$$
\begin{aligned}
(I) \leq \frac{\partial \Phi}{\partial t} + \frac{\partial \Phi}{\partial s} & \\
& \sup_{\alpha \in A} \frac{1}{2} |\operatorname{Tr}(\sigma\sigma^T)(\bar{t}, \bar{x}, \alpha) M - \operatorname{Tr}(\sigma\sigma^T)(\bar{s}, \bar{y}, \alpha) N| \\
& + \sup_{\alpha \in A} |b(\bar{t}, \bar{x}, \alpha) D_x \Phi + b(\bar{s}, \bar{y}, \alpha) D_y \Phi| \\
& + \sup_{\alpha \in A} |f(\bar{t}, \bar{x}, \bar{z}, \alpha) - f(\bar{s}, \bar{y}, \bar{w}, \alpha)|.
\end{aligned}
$$

Using (10) and (62), a few computations are then needed to give the following bounds[3]

$$
\begin{aligned}
(I) \leq C^{\beta\gamma\theta}\delta + C^{\beta\gamma\theta} \frac{(|\bar{t} - \bar{s}|^2 + |\bar{x} - \bar{y}|^2 + |\bar{z} - \bar{w}|^2)}{\delta} & \\
+ C^{\gamma\theta}\beta(|\bar{x}|^2 + |\bar{z}|^2 + |\bar{y}|^2 + |\bar{w}|^2) + C\gamma \left(\frac{1}{|\bar{z}|^2} + \frac{1}{|\bar{w}|^2} \right). &
\end{aligned}
$$

The second term brings

$$
(II) \leq C^{\beta\gamma\theta}\delta + \frac{2\beta\bar{z}^2}{p} \left(\frac{|\bar{x}|}{\bar{z}} \right)^p + \frac{2\beta\bar{w}^2}{p} \left(\frac{|\bar{y}|}{\bar{w}} \right)^p.
$$

We let successively δ, β, and γ go to zero. Since the penalization terms also go to zero, the contradiction follows.

Acknowledgments. The authors thank Professor Ivar Ekeland for his constant encouragement and advice and Florence Soule de Lafont for stimulating discussions concerning applications in finance. The authors also thank the referee for pointing out the existence of Barron [6] and Heinricher and Stockbridge [18] and for his (justified) criticisms of the second section. His suggestions led to an improvement of the draft and a simplification of the proof of Theorem 2.4.

REFERENCES

[1] G. BARLES (1990), *An approach of deterministic control problems with unbounded data*, Ann. Inst. H. Poincaré, 7, pp. 235–258.

[2] G. BARLES AND B. PERTHAME (1987), *Discontinuous solutions of deterministic optimal stopping problems*, Math. Model. Numer. Anal., 21, pp. 557–579.

[3] ——— (1988), *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim. 26, pp. 1133–1148.

[4] ——— (1990), *Comparison Principle for Dirichlet type Hamilton-Jacobi equations and Singular Perturbations of degenerated elliptic equations*, Appl. Math. Optim., 21, pp. 21–44.

[5] G. BARLES AND P. E. SOUGANIDIS (1991), *Convergence of approximation schemes for fully non-linear equations*, Asymptotic Anal., 4. pp. 271–283.

[6] E. N. BARRON, *The Bellman equation for the running max of a diffusion and applications to lookback options.* Appl. Anal., to appear.

[7] E. N. BARRON AND H. ISHII (1989), *The Bellman equation for minimizing the maximum cost*, Nonlinear Anal. TMA., 3, pp. 1067–1090.

[8] A. BENSOUSSAN AND J.-L. LIONS (1978), *Applications des inéquations variationnelles en controle stochastique*, Dunod, Paris.

[9] A. CONZE (1990), Ph.D. thesis, Université Paris IX Dauphine.

[10] J. COX AND M. RUBINSTEIN (1985), Options Markets, Prentice Hall, Englewood Cliffs, NJ.

[3] We denote C constants, emphasizing their dependance with respect to β, γ, etc.

[11] M. CRANDALL, H. ISHII, AND P. L. LIONS (1990), *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27, pp. 1–67.
[12] M. CRANDALL AND P. L. LIONS (1983), *Viscosity solutions of Hamilton–Jacobi Equations*, Trans. Amer. Math. Soc., 277, pp. 1–42.
[13] C. DAHER AND M. ROMANO (1990), *Evaluation of assets with a path-dependent cashflow*, preprint.
[14] ——— (1991), *Valuation of options on bonds on a vector and parallel computer*, in High Performance Computing II, M. Durand and F. El Dabaghi, eds. North-Holland, Amsterdam.
[15] D. DUFFIE (1988), Security Markets, Stochastic Models, Academic Press, New York.
[16] P. DUPUIS AND H. ISHII (1991), *On oblique derivative problems for fully nonlinear second-order equations on nonsmooth domains*, Nonlinear Anal. TMA, 12, pp. 1123–1138.
[17] ——— (1991), *On Oblique Derivative Problems for Fully Nonlinear Second-Order Elliptic PDE's on Domains with Corners*, Hokkaido Math. J., 20, pp. 135–164.
[18] A. HEINRICHER AND R. STOCKBRIDGE (1991), *Optimal control of the running max*, SIAM J. Control Optim., 29, pp. 936–953.
[19] J. E. INGERSOLL (1987), *Theory of Financial Decision Making*, Rowman and Littlefield, Totowa, NJ.
[20] H. ISHII (1989), *On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic PDE's*, Comm. Pure Appl. Math., 42, pp. 14–45.
[21] H. ISHII AND P. L. LIONS (1990), *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Differential Equations, 83, pp. 26–78.
[22] R. JENSEN (1988), *The maximum principle for viscosity solutions of fully nonlinear second-order partial differential equations*, Arch. Rational Mech. Anal., 101, pp. 1–27.
[23] ——— (1989), *Uniqueness criteria for viscosity solutions of fully nonlinear elliptic partial differential equations*, Indiana U. Math. J., 38, pp. 629–667.
[24] R. JENSEN, P. L. LIONS, AND P. E. SOUGANIDIS (1988), *A uniqueness result for viscosity solutions of second-order fully nonlinear partial differential equations*, Proc. Amer. Math. Soc., 102, pp. 975–978.
[25] I. KARATZAS (1988), *The pricing of American options*, Appl. Math. Optim., 17, pp. 37–60.
[26] I. KARATZAS AND S. E. SHREVE (1988), *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, Berlin.
[27] N. V. KRYLOV (1980), Controlled Diffusion Processes, Springer-Verlag, New York, Berlin.
[28] P. L. LIONS (1982), Generalized Solutions of Hamilton-Jacobi Equations, Pitman, Boston, MA.
[29] ——— (1983), *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Part I: The dynamic programming principle and applications*, Comm. Partial Differential Equations, 10, pp. 1101–1174.
[30] ——— (1983), *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Part II: Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 11, pp. 1229–1276.
[31] ——— (1985), *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Part III*, in Nonlinear PDE and Appl., College de France Seminar, vol. V, Pitman, Boston, MA.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 635–647, May 1994

003

ON THE GAME RICCATI EQUATIONS ARISING IN H_∞ CONTROL PROBLEMS*

PASCAL GAHINET†

Abstract. In the state-space approach to H_∞ optimal control, feasible closed-loop gains γ are characterized via a pair of game Riccati equations depending on γ. This paper is concerned with the properties of these equations as γ varies. The most general problem is considered ($D_{11} \neq 0$) and the variations of the Riccati solutions are thoroughly analyzed. Insight is gained into the behavior near the optimum and into the dependence on γ of the suboptimality conditions. In addition, concavity is established for a criterion that synthesizes the three conditions $X \geq 0, Y \geq 0$, and $\rho(XY) < \gamma^2$. This suggests a numerically reliable Newton scheme for the computation of the optimal γ.

Most results presented here are extensions of earlier contributions. The main concern is to provide a complete and synthetic overview as well as results and formulas tailored to the development of numerically sound algorithms.

Key words. H_∞ control, algebraic Riccati equation, Newton method

AMS subject classifications. 93C05, 93C35, 93C60, 93C45, 93B40, 49B99

1. Introduction.

Many significant problems in linear system theory can be recast into the abstract framework of H_∞ optimal control. Well-known examples include model matching, disturbance attenuation, mixed sensitivity design, and robust stabilization in the face of uncertainty [5]. The general H_∞ optimal control problem can be stated as follows. Consider a linear, time-invariant plant G, which maps exogenous inputs w and control inputs u to controlled outputs z and measured outputs y. That is,

$$\begin{pmatrix} z \\ y \end{pmatrix} = G(s) \begin{pmatrix} w \\ u \end{pmatrix} = \begin{pmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{pmatrix} \begin{pmatrix} w \\ u \end{pmatrix}.$$

When G is closed by the output feedback law $u = K(s)y$, the closed-loop transfer function from w to z is given by the linear fractional map:

$$\mathcal{F}(G, K) = G_{11} + G_{12}K(I - G_{22}K)^{-1}G_{21}. \tag{1.1}$$

The H_∞ optimal control problem consists of finding some real-rational, proper, and causal controller K_{opt} that internally stabilizes the plant while minimizing the norm $\|\mathcal{F}(G, K)\|_\infty$; that is,

$$\|\mathcal{F}(G, K_{opt})\|_\infty = \inf\{\|\mathcal{F}(G, K)\|_\infty : K \text{ internally stabilizes } G\}. \tag{1.2}$$

The infimum of all achievable gains is denoted γ_{opt}.

Although direct computation of γ_{opt} is a hard problem, the following suboptimal problem is relatively well understood and tractable:

(1.3)

Given $\gamma > 0$, does there exist an internally stabilizing K such that $\|\mathcal{F}(G, K)\|_\infty < \gamma$?

This paper is concerned with Doyle and Glover's state-space approach [4], [9] to solving this suboptimal problem. Over the past decade, this approach has emerged as the most direct and practical solution both on design and numerical grounds. It is now briefly summarized. To begin, we introduce the following minimal realization of the plant G:

$$G(s) = \begin{pmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{pmatrix} = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} (sI - A)^{-1} (B_1 \quad B_2). \tag{1.4}$$

* Received by the editors April 30, 1991; accepted for publication (in revised form) November 2, 1992.

† Institut National de Recherche en Informatique et Automatique, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France (gahinet@colorado.inria.fr.).

Here $A \in \mathbb{R}^{n\times n}$ and z, y, w, and u are vectors of size p_1, p_2, m_1, and m_2, respectively, with the assumption that $m_1 \geq p_2$ and $p_1 \geq m_2$. Accordingly, $D_{11} \in \mathbf{R}^{p_1\times m_1}$, $D_{12} \in \mathbf{R}^{p_1\times m_2}$, $D_{21} \in \mathbf{R}^{p_2\times m_1}$, and $D_{22} \in \mathbf{R}^{p_2\times m_2}$. Associate with this data the following two Hamiltonian matrices:

$$H_\gamma = \begin{pmatrix} A & 0 \\ -C_1^T C_1 & -A^T \end{pmatrix} + \begin{pmatrix} B_1 & B_2 \\ -C_1^T D_{11} & -C_1^T D_{12} \end{pmatrix} \times \begin{pmatrix} \gamma^2 I - D_{11}^T D_{11} & -D_{11}^T D_{12} \\ -D_{12}^T D_{11} & -D_{12}^T D_{12} \end{pmatrix}^{-1} \begin{pmatrix} D_{11}^T C_1 & B_1^T \\ D_{12}^T C_1 & B_2^T \end{pmatrix}; \tag{1.5}$$

$$J_\gamma = \begin{pmatrix} A^T & 0 \\ -B_1 B_1^T & -A \end{pmatrix} + \begin{pmatrix} C_1^T & C_2^T \\ -B_1 D_{11}^T & -B_1 D_{21}^T \end{pmatrix} \times \begin{pmatrix} \gamma^2 I - D_{11} D_{11}^T & -D_{11} D_{21}^T \\ -D_{21} D_{11}^T & -D_{21} D_{21}^T \end{pmatrix}^{-1} \begin{pmatrix} D_{11} B_1^T & C_1 \\ D_{21} B_1^T & C_2 \end{pmatrix}. \tag{1.6}$$

Given a Hamiltonian matrix

$$H = \begin{pmatrix} A & R \\ -Q & -A^T \end{pmatrix}$$

where R, Q are symmetric matrices, we denote by $\mathcal{X}_-(H)$ the stable invariant subspace of H. Associated with H is the algebraic Riccati equation $A^T X + XA + XRX + Q = 0$. Recall that this equation has a (unique) symmetric stabilizing solution if and only if H has no pure imaginary eigenvalue and $\mathcal{X}_-(H)$ is complementary to the subspace Im $\binom{0}{I}$. When existing, such a solution is obtained as $X = QP^{-1}$, where $\binom{P}{Q}$ is any basis of $\mathcal{X}_-(H)$. In the sequel, Ric(H) will refer to the stabilizing solution.

Throughout the paper, we call the *general problem* (GP) the problem (1.3) with the following standing assumptions.

Assumption 1. (A, B_2, C_2) is stabilizable and detectable.

Assumption 2. D_{12} has full column rank and D_{21} has full row rank.

Assumption 3.

$$\text{rank} \begin{pmatrix} j\omega I - A & -B_2 \\ C_1 & D_{12} \end{pmatrix} = n + m_2 \quad \text{and} \quad \text{rank} \begin{pmatrix} j\omega I - A & B_1 \\ -C_2 & D_{21} \end{pmatrix} = n + p_2$$

for all $\omega \in \mathbb{R}$ or equivalently, G_{12} and G_{21} have no transmission zero on the imaginary axis [11].

Assumption 4. $D_{22} = 0$.

While Assumption 1 is necessary and sufficient for solvability of the GP for γ large enough, Assumptions 2 and 3 are restrictive assumptions required for validity of Doyle and Glover's state-space results. Finally, Assumption 4 amounts to a reparametrization of the controller set [9] and hence incurs no loss of generality.

Under Assumptions 1–4, solvability of the GP with attenuation level γ is characterized in terms of the solutions of the two Riccati equations associated with H_γ and J_γ [9].

THEOREM 1.1. *With Assumptions* 1–4, *there exists an internally stabilizing controller* K *such that* $\|\mathcal{F}(G, K)\|_\infty < \gamma$ *if and only if*

$$\gamma > \sigma_d := \max\{\sigma_{\max}((I - D_{12}(D_{12}^T D_{12})^{-1} D_{12}^T) D_{11}), \sigma_{\max}(D_{11}(I - D_{21}^T (D_{21} D_{21}^T)^{-1} D_{21}))\} \tag{1.7}$$

and the following conditions hold:

(C1) *H_γ and J_γ have no eigenvalue on the imaginary axis*;
(C2) *Neither $\mathcal{X}_-(H_\gamma)$ nor $\mathcal{X}_-(J_\gamma)$ intersects* $\operatorname{Im}\binom{0}{I}$;
(C3) $\operatorname{Ric}(H_\gamma) =: X_\gamma \geq 0$ *and* $\operatorname{Ric}(J_\gamma) =: Y_\gamma \geq 0$;
(C4) $\rho(X_\gamma Y_\gamma) < \gamma^2$. □

Conditions (C1)–(C4) are often referred to as DGKF's conditions after the authors of [4]. Note that (1.7) ensures that the inverses in H_γ and J_γ are well defined (see (3.2) together with the definition of $\hat{D}_{11}$ in (3.1)). Note also that the Riccati equations associated with H_γ and J_γ have an indefinite quadratic term and are thus referred to as game Riccati equations (GRE). Finally, Theorem 1.1 characterizes γ_{opt} as the smallest $\gamma > 0$ for which the four conditions (C1)–(C4) are jointly satisfied. In turn, this suggests a straightforward bisection algorithm to compute γ_{opt}.

This paper examines the dependence on γ of conditions (C1)–(C4) and discusses the implications for the computation of the optimal gain γ_{opt}. To this purpose, the properties of the game Riccati equations associated with H_γ and J_γ are analyzed in detail. After simplifying the expressions of H_γ and J_γ by elementary reparametrizations, §3 characterizes those γ for which (C1) holds. In §4, the variations of the pseudoinverses of X_γ and Y_γ are shown to be smooth, monotonic, and even concave wherever (C1) is satisfied. These results are obtained for the GP ($D_{11} \neq 0$) and allow a complete description of the variations of X_γ and Y_γ and of the behavior near γ_{opt} (§5). Finally, the computation of γ_{opt} is reformulated as a convex zero-crossing search problem that can be numerically solved by a Newton method (§6).

We conclude with a justification of our treatment of the GP instead of the simpler *Standard Problem* (SP) considered in [4]. Recall that the SP requires the additional assumptions

Assumption 5. $D_{11} = 0$,

Assumption 6. $D_{12}^T(D_{12}, C_1) = (I, 0)$ and $D_{21}(D_{21}^T, B_1^T) = (I, 0)$.

These assumptions have the advantage of notably simplifying the expressions of H_γ and J_γ. Moreover, such simplifications can be emulated for any GP via the loop-shifting techniques of [14]. Yet, these manipulations destroy the variational and structural properties of X_γ and Y_γ. Indeed, enforcing Assumption 5 requires a γ-dependent transformation that alters the gradient and complicates its computation. In addition, the transformed problem is by no means an SP. In fact, the Riccati equations associated with GPs and SPs are structurally different as illustrated by Theorem 5.6 below. For theoretical and numerical reasons, it is therefore advisable to work in the GP framework.

2. Notation and terminology. Given a square matrix M, a subspace $\mathcal{S}$ is said to be M-invariant if $M\mathcal{S} \subset \mathcal{S}$, and stable (antistable) M-invariant if, moreover, the restriction of M to $\mathcal{S}$ is stable (antistable). The following notation and definitions are used throughout the paper:

$\mathbb{C}_-, \mathbb{C}_0, \mathbb{C}_+$	open left-half plane, imaginary axis, and open right-half plane, respectively;
$\operatorname{Ker} X$, $\operatorname{Im} X$	null and range spaces of a matrix X, respectively;
$\mathcal{X}_-(M)$	stable M-invariant subspace;
$\Lambda(X), \rho(X)$	spectrum and spectral radius of a square matrix X, respectively;
$\sigma_{\max}(X)$	largest singular value of the matrix X;
$\operatorname{In}(X)$	the inertia of $X = X^T$, that is, the triple (π, ν, ζ) where π, ν, ζ denote the number of positive, negative, and zero eigenvalues of X, respectively;
$\mathcal{V}_0(C, A)$	A-invariant subspace associated with the stable, (C, A)-unobservable modes of A.

3. Condition on the Hamiltonian spectrum. This section examines which restriction is imposed on γ by condition (C1). It is shown that the region where H_γ has no eigenvalue on the imaginary axis is a half line $\gamma > \gamma_H$. A computable formula for the threshold γ_H is also derived. These results extend earlier work in [16] to the general case $D_{11} \neq 0$.

To simplify subsequent calculations, the expressions (1.5) and (1.6) of H_γ and J_γ are first condensed in terms of compound parameters. With $D_{12}^+ = (D_{12}^T D_{12})^{-1} D_{12}^T$ denoting the pseudoinverse of D_{12}, introduce the following parameters:

$$\begin{aligned} \hat{A} &:= A - B_2 D_{12}^+ C_1; \qquad \hat{B}_1 := B_1 - B_2 D_{12}^+ D_{11}; \qquad \hat{B}_2 := B_2 (D_{12}^T D_{12})^{-1/2}; \\ &\qquad\quad \hat{C}_1 := (I - D_{12} D_{12}^+) C_1; \qquad \hat{D}_{11} := (I - D_{12} D_{12}^+) D_{11}. \end{aligned} \tag{3.1}$$

By elementary algebra, H_γ can be rewritten as

$$H_\gamma = \begin{pmatrix} \hat{A} & -\hat{B}_2 \hat{B}_2^T \\ -\hat{C}_1^T \hat{C}_1 & -\hat{A}^T \end{pmatrix} + \begin{pmatrix} \hat{B}_1 \\ -\hat{C}_1^T \hat{D}_{11} \end{pmatrix} (\gamma^2 I - \hat{D}_{11}^T \hat{D}_{11})^{-1} (\hat{D}_{11}^T \hat{C}_1, \hat{B}_1^T). \tag{3.2}$$

Similarly, with $D_{21}^+ := D_{21}^T (D_{21} D_{21}^T)^{-1}$ and

$$\begin{aligned} \tilde{A} &:= A - B_1 D_{21}^+ C_2; \qquad \tilde{C}_1 := C_1 - D_{11} D_{21}^+ C_2; \qquad \tilde{C}_2 := (D_{21} D_{21}^T)^{-1/2} C_2; \\ &\qquad\quad \tilde{B}_1 := B_1 (I - D_{21}^+ D_{21}); \qquad \tilde{D}_{11} := D_{11} (I - D_{21}^+ D_{21}), \end{aligned} \tag{3.3}$$

J_γ can be simplified to

$$J_\gamma = \begin{pmatrix} \tilde{A}^T & -\tilde{C}_2^T \tilde{C}_2 \\ -\tilde{B}_1 \tilde{B}_1^T & -\tilde{A} \end{pmatrix} + \begin{pmatrix} \tilde{C}_1^T \\ -\tilde{B}_1 \tilde{D}_{11}^T \end{pmatrix} (\gamma^2 I - \tilde{D}_{11} \tilde{D}_{11}^T)^{-1} (\tilde{D}_{11} \tilde{B}_1^T, \tilde{C}_1). \tag{3.4}$$

Note that the underlying reparametrizations are independent of γ and transparent for analytical purposes since neither H_γ nor J_γ is altered. In addition, the stabilizability of (A, B_2) is equivalent to that of $(\hat{A}, \hat{B}_2)$ and similarly for the detectability of (C_2, A) and $(\tilde{C}_2, \tilde{A})$. Finally, observe that even though $D_{12}^T \hat{C}_1 = 0$ and $D_{21} \tilde{B}_1^T = 0$, and even if $D_{11} = 0$, the reduced expressions (3.2) and (3.4) cannot be associated with any particular SP since $\hat{A}$ and $\tilde{A}$ are distinct in general.

On their domain of existence, the GRE solutions $X_\gamma = \mathrm{Ric}(H_\gamma)$ have the important property of sharing the same null space as γ varies [15]. In other words, the singular part of X_γ is independent of γ and coincides with the stable $(\hat{C}_1, \hat{A})$-unobservable subspace that we denote by $\mathcal{V}_0(\hat{C}_1, \hat{A})$. Of course a similar result holds for Y_γ. This structural property is instrumental to the subsequent variational analysis which involves forming the pseudoinverses of X_γ and Y_γ. Indeed, consider some orthogonal change of coordinates $U = (U_1, U_2)$ such that

$$\begin{aligned} U^T \hat{A} U &= \begin{pmatrix} \hat{A}_{11} & 0 \\ \star & \hat{A}_{22} \end{pmatrix}; \qquad \hat{C}_1 U = (\hat{C}_{11} \quad 0); \\ U^T \hat{B}_1 &= \begin{pmatrix} \hat{B}_{11} \\ \star \end{pmatrix}; \qquad U^T \hat{B}_2 = \begin{pmatrix} \hat{B}_{21} \\ \star \end{pmatrix}; \end{aligned} \tag{3.5}$$

where $\hat{A}_{22}$ is stable and $(\hat{C}_{11}, \hat{A}_{11})$ has no stable unobservable mode—in other words, such that

$$\mathrm{Im}\, U_2 = \mathcal{V}_0(\hat{C}_1, \hat{A}); \qquad \mathrm{Im}\, U_1 = \mathcal{V}_0^\perp(\hat{C}_1, \hat{A}). \tag{3.6}$$

Then Ker $X_\gamma = \text{Im } U_2$, and

$$U^T X_\gamma U = \begin{pmatrix} \bar{X}_\gamma & 0 \\ 0 & 0 \end{pmatrix}$$

or equivalently

$$X_\gamma = U_1 \bar{X}_\gamma U_1^T, \tag{3.7}$$

where $\bar{X}_\gamma$ is nonsingular for all γ's. Hence $X_\gamma^+ = U_1 \bar{X}_\gamma^{-1} U_1^T$ and since U_1 is independent of γ, the variations of X_γ or X_γ^+ are completely described by those of the invertible matrix $\bar{X}_\gamma$. Similarly, Y_γ can be written as

$$Y_\gamma = V_1 \bar{Y}_\gamma V_1^T, \tag{3.8}$$

where V_1 is any orthogonal complement of $\mathcal{V}_0(\tilde{B}_1^T, \tilde{A}^T)$.

Remark 3.1. The (stable) unobservable modes of $(\hat{C}_1, \hat{A})$ are exactly the (stable) invariant zeros of $G_{12}(s)$, i.e., the complex numbers $s \in \mathbb{C}_-$ for which the system matrix

$$P(s) = \begin{pmatrix} sI - A & -B_2 \\ C_1 & D_{12} \end{pmatrix}$$

loses rank. This follows from the identities

$$\begin{pmatrix} sI - \hat{A} & 0 \\ \hat{C}_1 & D_{12} \end{pmatrix} = \begin{pmatrix} I & -B_2 D_{12}^+ \\ 0 & I \end{pmatrix} P(s) \begin{pmatrix} I & 0 \\ -D_{12}^+ C_1 & I \end{pmatrix}$$

and $D_{12}^T \hat{C}_1 = 0$. Indeed, these show that $P(s)$ is (column) rank deficient if and only if $\begin{pmatrix} sI - \hat{A} \\ \hat{C}_1 \end{pmatrix}$ is rank deficient, that is, if and only if s is an unobservable mode of $(\hat{C}_1, \hat{A})$. □

The first requirement (C1) for solvability of the GP is concerned with pure imaginary eigenvalues of H_γ or J_γ. The following theorem characterizes the region where (C1) is satisfied.

THEOREM 3.2. *Assume Assumptions* 1–4 *and consider* H_γ *given by* (1.5). *Then there exists a finite real number* $\gamma_H \geq \sigma_{\max}(\hat{D}_{11})$ *such that*

$$\Lambda(H_\gamma) \cap \mathbb{C}_0 = \emptyset \qquad \textit{if and only if } \gamma > \gamma_H. \tag{3.9}$$

Moreover, γ_H *can be computed as* (*using the notation* (3.5))

$$\gamma_H = \|\hat{D}_{11} + \hat{C}_{11}(sI + \hat{A}_Z)^{-1}(\hat{B}_{11} + Z\hat{C}_{11}^T \hat{D}_{11})\|_\infty, \tag{3.10}$$

where Z *is the unique stabilizing solution of*

$$-\hat{A}_{11} Z - Z\hat{A}_{11}^T - Z\hat{C}_{11}^T \hat{C}_{11} Z + \hat{B}_{21}\hat{B}_{21}^T = 0 \tag{3.11}$$

and $\hat{A}_Z := -\hat{A}_{11} - \hat{C}_{11}^T \hat{C}_{11} Z$ *is the corresponding* (*stable*) *closed-loop matrix.*

Proof. The proof is easily adapted from [16]. See Appendix A for details. □

Hence part of the spectrum of H_γ migrates toward the imaginary axis as γ decreases. The first contact occurs for $\gamma = \gamma_H$ and those eigenvalues which then reach the imaginary axis remain on the axis for all $\sigma_{\max}(\hat{D}_{11}) < \gamma \leq \gamma_H$. If γ_J denotes the counterpart of γ_H for J_γ and γ^* is defined as

$$\gamma^* = \max(\gamma_H, \gamma_J); \tag{3.12}$$

it follows that (C1) holds if and only if $\gamma > \gamma^*$. Note that Assumption 3 is necessary and sufficient to ensure that $\gamma^* < +\infty$ (cf. Remark 3.1). Also observe that $\gamma^* \geq \sigma_d$ of Theorem 1.1 since the matrices involved in the definition of σ_d are exactly $\hat{D}_{11}$ and $\tilde{D}_{11}$, while $\gamma_H \geq \sigma_{\max}(\hat{D}_{11})$ from (3.10) and $\gamma_J \geq \sigma_{\max}(\tilde{D}_{11})$ by duality.

4. Variational properties. We now restrict our attention to the interval $(\gamma^*, +\infty)$ where (C1) is satisfied and examine the regularity and variations of the GRE stabilizing solutions X_γ and Y_γ. Direct characterization of these variations is rendered difficult by the discontinuities arising where the complementarity condition (C2) fails. Fortunately, this problem disappears when considering instead the pseudoinverses of X_γ and Y_γ. Introduced in [16] for the SP, this technique allows a simple and powerful description of the variations with γ. Indeed, the continuous extensions of X_γ^+ and Y_γ^+ turn out to be monotonic and concave functions of the parameter $\alpha = \gamma^{-2}$. These results are now extended to the GP framework ($D_{11} \neq 0$) with no major difficulty except perhaps for the concavity part. On duality grounds, only the case of X_γ is considered here.

Recall from §3 that whenever (C1) and (C2) are satisfied, X_γ can be decomposed independently of γ as $X_\gamma = U_1 \bar{X}_\gamma U_1^T$ where $\bar{X}_\gamma$ is nonsingular. Consequently the pseudoinverse X_γ^+ is obtained as $X_\gamma^+ = U_1 \bar{X}_\gamma^{-1} U_1^T$ and its variations are entirely determined by those of $\bar{X}_\gamma^{-1}$. In the sequel, we therefore restrict our attention to the continuous extension $W_X(\gamma)$ of $\bar{X}_\gamma^{-1}$ on $(\gamma^*, +\infty)$. This extension is easily defined in terms of the decomposition (3.5) and of the stable invariant subspace of the reduced Hamiltonian

$$\bar{H}_\gamma := \begin{pmatrix} \hat{A}_{11} & -\hat{B}_{21}\hat{B}_{21}^T \\ -\hat{C}_{11}^T\hat{C}_{11} & -\hat{A}_{11}^T \end{pmatrix} + \begin{pmatrix} \hat{B}_{11} \\ -\hat{C}_{11}^T\hat{D}_{11} \end{pmatrix} \times (\gamma^2 I - \hat{D}_{11}^T\hat{D}_{11})^{-1}(\hat{D}_{11}^T\hat{C}_{11} \quad \hat{B}_{11}^T). \tag{4.1}$$

Specifically, given any basis $\begin{pmatrix} \bar{P} \\ \bar{Q} \end{pmatrix}$ of $\mathcal{X}_-(\bar{H}_\gamma)$, $\bar{Q}$ is invertible since $(\hat{C}_{11}, -\hat{A}_{11})$ is detectable and $W_X(\gamma) := \bar{P}\bar{Q}^{-1}$ is well defined for all $\gamma > \gamma_H$. With this definition, $W_X(\gamma) = \bar{X}_\gamma^{-1}$ whenever $\bar{X}_\gamma$ exists. Moreover, $W_X(\gamma)$ is the unique stabilizing solution of the GRE

$$\begin{aligned} -(\hat{A}_{11} + \hat{B}_{11}R_\gamma^{-1}\hat{D}_{11}^T\hat{C}_{11})\,W_X(\gamma) - W_X(\gamma)\,(\hat{A}_{11} + \hat{B}_{11}R_\gamma^{-1}\hat{D}_{11}^T\hat{C}_{11})^T \\ - \gamma^2 W_X(\gamma)\hat{C}_{11}^T S_\gamma^{-1}\hat{C}_{11}W_X(\gamma) + \hat{B}_{21}\hat{B}_{21}^T - \hat{B}_{11}R_\gamma^{-1}\hat{B}_{11}^T = 0, \end{aligned} \tag{4.2}$$

where $R_\gamma := \gamma^2 I - \hat{D}_{11}^T\hat{D}_{11} > 0$ and $S_\gamma := \gamma^2 I - \hat{D}_{11}\hat{D}_{11}^T > 0$, and the corresponding (stable) closed-loop matrix is

$$A_W = -(\hat{A}_{11} + \hat{B}_{11}R_\gamma^{-1}\hat{D}_{11}^T\hat{C}_{11})^T - \gamma^2\hat{C}_{11}^T S_\gamma^{-1}\hat{C}_{11}W_X(\gamma). \tag{4.3}$$

Regularity and monotonicity of $W_X(\gamma)$ over the whole interval $(\gamma_H, +\infty)$ are established in the next theorem.

THEOREM 4.1. *With Assumptions* 1–4 *of* §1, *the matrix-valued function* $W_X(\gamma)$ *defined above has the following properties*:

1. $W_X(\gamma)$ *is infinitely differentiable on* $(\gamma_H, +\infty)$;
2. $W_X(\gamma)$ *is monotonically increasing with* γ;
3. $\lim_{\gamma\to+\infty} W_X(\gamma)$ *exists and is positive definite.*

Proof. (1) For $\gamma > \gamma_H$, $W_X(\gamma)$ is the stabilizing solution of the GRE (4.2) whose parameters are infinitely differentiable functions of γ. By the Implicit Function Theorem applied to the stabilizing solution of algebraic Riccati equations (see, e.g., [6], [7], [13]), W_X is therefore infinitely differentiable.

(2) It is equivalent but more convenient to show that W_X is a decreasing function of the parameter $\alpha = \gamma^{-2}$. Introduce $R_\alpha := I - \alpha\hat{D}_{11}^T\hat{D}_{11} > 0$ and $S_\alpha := I - \alpha\hat{D}_{11}\hat{D}_{11}^T > 0$, and observe that $R_\gamma^{-1} = \alpha R_\alpha^{-1}$ and $\gamma^2 S_\gamma^{-1} = S_\alpha^{-1}$. Moreover,

$$\frac{d}{d\alpha}(R_\gamma^{-1}) = R_\alpha^{-2}; \qquad \frac{d}{d\alpha}(\gamma^2 S_\gamma^{-1}) = S_\alpha^{-1}\hat{D}_{11}\hat{D}_{11}^T S_\alpha^{-1} = \hat{D}_{11}R_\alpha^{-2}\hat{D}_{11}^T, \tag{4.4}$$

where the last identity follows from $S_\alpha^{-1}\hat{D}_{11} = \hat{D}_{11}R_\alpha^{-1}$. Differentiating (4.2) with respect to $\alpha = \gamma^{-2}$ then yields

$$A_W^T\frac{dW_X}{d\alpha} + \frac{dW_X}{d\alpha}A_W - \hat{B}_{11}R_\alpha^{-2}\hat{D}_{11}^T\hat{C}_{11}W_X - W_X\hat{C}_{11}^T\hat{D}_{11}R_\alpha^{-2}\hat{B}_{11}^T$$
$$- W_X\hat{C}_{11}^T\hat{D}_{11}R_\alpha^{-2}\hat{D}_{11}^T\hat{C}_{11}W_X - \hat{B}_{11}R_\alpha^{-2}\hat{B}_{11}^T = 0,$$

which can be rewritten more compactly as

$$A_W^T\frac{dW_X}{d\alpha} + \frac{dW_X}{d\alpha}A_W - (\hat{B}_{11} + W_X\hat{C}_{11}^T\hat{D}_{11})R_\alpha^{-2}(\hat{B}_{11} + W_X\hat{C}_{11}^T\hat{D}_{11})^T = 0. \tag{4.5}$$

Since A_W is stable and $R_\alpha > 0$, it follows from the Lyapunov Theorem that $dW_X/d\alpha \le 0$, that is, W_X is a monotonically decreasing function of α and thus a monotonically increasing function of γ.

(3) As $\gamma \to +\infty$, $W_X(\gamma)$ tends to the nonnegative stabilizing solution of the LQG-type Riccati equation

$$-\hat{A}_{11}W - W\hat{A}_{11}^T - W\hat{C}_{11}^T\hat{C}_{11}W + \hat{B}_{21}\hat{B}_{21}^T = 0.$$

Since $(\hat{A}, \hat{B}_2)$ and therefore $(\hat{A}_{11}, \hat{B}_{21})$ are stabilizable, this solution is nonsingular and hence $W_X(\infty) > 0$. □

In addition, W_X turns out to be a concave function of the parameter $\alpha = \gamma^{-2}$. This property is the foundation of the Newton algorithm proposed in §6 for the computation of the optimal gain γ_{opt}.

THEOREM 4.2. *W_X is a monotonically decreasing concave function of the parameter $\alpha = \gamma^{-2}$.*

Proof. The monotonicity was established in Theorem 4.1. To establish concavity, we calculate the second derivative of W_X with respect to α. First observe from (4.3) and (4.4) that

$$\begin{aligned}\frac{dA_W}{d\alpha} &= -\hat{C}_{11}^TS_\alpha^{-1}\hat{C}_{11}\frac{dW_X}{d\alpha} - \hat{C}_{11}^T\hat{D}_{11}R_\alpha^{-2}\hat{B}_{11}^T - \hat{C}_{11}^T\hat{D}_{11}R_\alpha^{-2}\hat{D}_{11}^T\hat{C}_{11}W_X \\ &= -\hat{C}_{11}^TS_\alpha^{-1}\hat{C}_{11}\frac{dW_X}{d\alpha} - \hat{C}_{11}^T\hat{D}_{11}R_\alpha^{-2}F^T\end{aligned} \tag{4.6}$$

with $F := \hat{B}_{11} + W_X\hat{C}_{11}^T\hat{D}_{11}$. Then differentiate (4.5) with respect to α to obtain

$$A_W^T\frac{d^2W_X}{d\alpha^2} + \frac{d^2W_X}{d\alpha^2}A_W + \frac{dA_W^T}{d\alpha}\frac{dW_X}{d\alpha} + \frac{dW_X}{d\alpha}\frac{dA_W}{d\alpha}$$
$$- \frac{dW_X}{d\alpha}\hat{C}_{11}^T\hat{D}_{11}R_\alpha^{-2}F^T - FR_\alpha^{-2}\hat{D}_{11}^T\hat{C}_{11}\frac{dW_X}{d\alpha} - 2FR_\alpha^{-3}\hat{D}_{11}^T\hat{D}_{11}F^T = 0,$$

which combined to (4.6) yields

$$A_W^T\frac{d^2W_X}{d\alpha^2} + \frac{d^2W_X}{d\alpha^2}A_W - 2\left\{\frac{dW_X}{d\alpha}\hat{C}_{11}^TS_\alpha^{-1}\hat{C}_{11}\frac{dW_X}{d\alpha} + \frac{dW_X}{d\alpha}\hat{C}_{11}^T\hat{D}_{11}R_\alpha^{-2}F^T\right.$$
$$\left. + FR_\alpha^{-2}\hat{D}_{11}^T\hat{C}_{11}\frac{dW_X}{d\alpha} + FR_\alpha^{-3}\hat{D}_{11}^T\hat{D}_{11}F^T\right\} = 0. \tag{4.7}$$

A little algebra shows that

$$R_\alpha^{-3}\hat{D}_{11}^T\hat{D}_{11} = R_\alpha^{-2}\hat{D}_{11}^T\hat{D}_{11}R_\alpha^{-1} = R_\alpha^{-2}\{\hat{D}_{11}^T\hat{D}_{11}R_\alpha\}R_\alpha^{-2} = R_\alpha^{-2}\hat{D}_{11}^TS_\alpha\hat{D}_{11}R_\alpha^{-2}$$

and consequently the bracketed term in (4.7) can be factorized to obtain

$$A_W^T \frac{d^2W_X}{d\alpha^2} + \frac{d^2W_X}{d\alpha^2} A_W - 2\left(S_\alpha^{-1/2}\hat{C}_{11}\frac{dW_X}{d\alpha} + S_\alpha^{1/2}\hat{D}_{11}R_\alpha^{-2}F^T\right)^T \cdot \left(S_\alpha^{-1/2}\hat{C}_{11}\frac{dW_X}{d\alpha} + S_\alpha^{1/2}\hat{D}_{11}R_\alpha^{-2}F^T\right) = 0. \tag{4.8}$$

The stability of A_W then clearly ensures $d^2W_X/d\alpha^2 \le 0$, that is, the concavity of W_X as a function of $\alpha = \gamma^{-2}$. □

Once established for $\alpha = \gamma^{-2}$, similar conclusions in terms of the parameters γ^2 and γ^{-1} easily follow from the differentiation formulas for the composition of functions.

COROLLARY 4.3. *W_X is decreasing and concave as a function of γ^{-1} and is increasing and concave as a function of γ^2.*

Proof. Omitted for brevity. □

5. Variations with γ and behavior near the optimum. This section gathers the results accumulated so far to give a complete description of the dependence on γ of conditions (C1)–(C4) of §1. The role played by these conditions in the determination of γ_{opt} is also addressed and a few examples demonstrate that any one of (C1)–(C4) can fail at γ_{opt}.

Firstly, the theorems of §4 are applied to characterizing the behavior of X_γ and Y_γ over the half line $(\gamma^*, +\infty)$. The following technical lemma is a useful preliminary.

LEMMA 5.1. *There are at most n isolated points in $(\gamma_H, +\infty)$ where $W_X(\gamma)$ is singular.*

Proof. See Appendix B. □

We can now proceed with a corollary of Theorem 4.1 which describes the variations of X_γ.

COROLLARY 5.2. *The* GRE *stabilizing solution $X_\gamma = \mathrm{Ric}(H_\gamma)$ has the following properties on the interval $(\gamma_H, +\infty)$:*

1. *X_γ is defined for all $\gamma > \gamma_H$ except at most n points $\gamma_1 < \cdots < \gamma_k$ where $\|X_\gamma\| = +\infty$. These points of discontinuity $\gamma_1, \ldots, \gamma_k$ are exactly the points where $W_X(\gamma)$ is singular;*
2. *With the convention $\gamma_0 := \gamma_H$ and $\gamma_{k+1} := +\infty$, X_γ is monotonically decreasing and of constant inertia on each interval (γ_i, γ_{i+1}), $i = 0 : k$;*
3. *When γ traverses some point of discontinuity γ_i from γ_i^+ to γ_i^-, at least one positive eigenvalue of X_γ escapes to $+\infty$ and reappears at $-\infty$.*

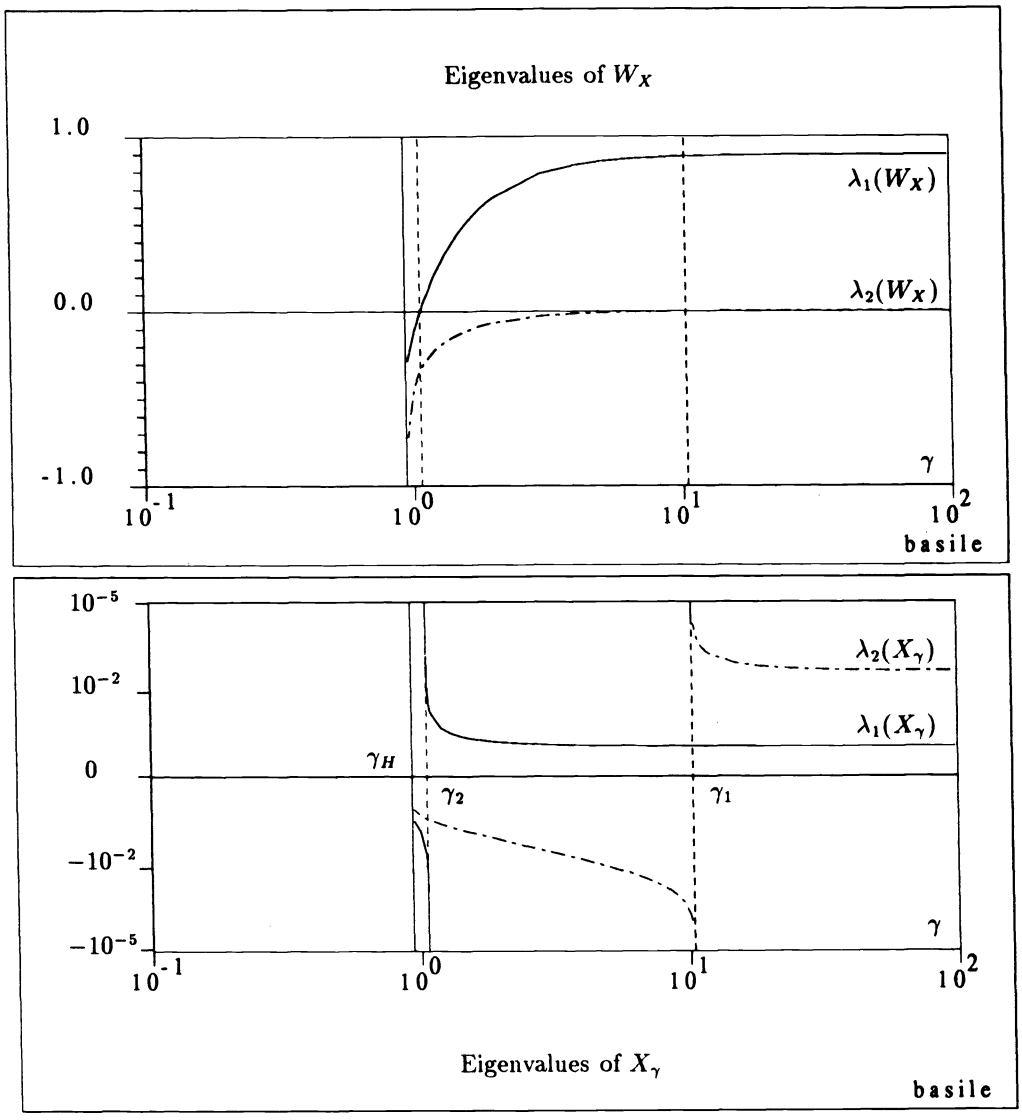
Proof. This corollary is an immediate consequence of Theorem 4.1 and of the fact that $\bar{X}_\gamma = W_X^{-1}(\gamma)$ whenever (C2) holds. To be convinced that the inertia of X_γ is constant on each (γ_i, γ_{i+1}), just observe that $\bar{X}_\gamma$ is nonsingular and continuous on these intervals which prohibits any change of inertia. □

Remarkably, sign changes in the eigenvalues of X_γ never result from zero-crossing. Instead, positive eigenvalues change sign by escaping to $+\infty$ and reappearing at $-\infty$. As γ decreases from $+\infty$ in particular, $\|X_\gamma\|$ necessarily blows up when (C3) is about to fail. Hence (C2) and (C3) first fail simultaneously. These various observations are illustrated on a simple example.

Example 5.3. Consider the GP with matrix data:

$$A = \begin{pmatrix} 10 & 2 \\ -1 & 8 \end{pmatrix}; \qquad B_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}; \qquad B_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix};$$

$$C_1 = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}; \qquad D_{12} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \qquad D_{11} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$



GRAPH 5.1.

We focus on the variations of X_γ and of its inverse $W_X(\gamma)$ (here X_γ is nonsingular whenever defined). The eigenvalues of these two matrix-valued functions are plotted in Graph 5.1 for $\gamma > \gamma_H$ (here, $\gamma_H \approx 0.933$). The largest and smallest eigenvalues of W_X are denoted by $\lambda_1(W_X)$ and $\lambda_2(W_X)$, respectively, while $\lambda_1(X_\gamma)$ and $\lambda_2(X_\gamma)$ are taken as

$$\lambda_1(X_\gamma) = 1/\lambda_1(W_X); \qquad \lambda_2(X_\gamma) = 1/\lambda_2(W_X).$$

Solid lines are used for the plots of $\lambda_1(W_X)$ and $\lambda_1(X_\gamma)$ and dashed lines for the plots of $\lambda_2(W_X)$ and $\lambda_2(X_\lambda)$. Note also the semilogarithmic scale for the eigenvalues of X_γ.

Inspection of Graph 5.1 shows that X_γ has two points of discontinuity γ_1 and γ_2 in $(\gamma_H, +\infty)$ that correspond to the zero-crossings of $\lambda_1(W_X)$ and $\lambda_2(W_X)$. The inertia of X_γ changes when traversing either one of these discontinuities. Indeed, $\mathrm{In}(X_\gamma) = (2,0,0)$ for $\gamma > \gamma_1$, $\mathrm{In}(X_\gamma) = (1,0,1)$ for $\gamma_2 < \gamma < \gamma_1$, and $\mathrm{In}(X_\gamma) = (0,0,2)$ for $\gamma_H < \gamma < \gamma_2$. Finally, X_γ has a finite limit as $\gamma \to \gamma_H^+$. □

In most problems, condition (C4) is the first to fail when approaching the optimum γ_{opt}. Yet, (C1), (C2), or (C3) can sometimes be first to fail as demonstrated by the next two examples.

Example 5.4. This example illustrates that (C1) can fail first at γ_{opt}. Consider the plant $G(s)$ of realization (1.4) with $C_1 = I_2, D_{21} = 1, D_{22} = 0$, and

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}; \qquad B_1 = B_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix};$$
$$C_2^T = D_{11} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \qquad D_{12} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Clearly, Assumption 1 of §1 is satisfied since A is stable and Assumptions 2 and 3 are trivially verified. After calculation of H_γ and J_γ via (1.5) and (1.6), elementary algebra shows that $\Lambda(H_\gamma)$ intersects $\mathbb{C}_0$ for $\gamma \le 1/\sqrt{2}$ while J_γ never has any eigenvalue on $\mathbb{C}_0$. Therefore $\gamma^* = 1/\sqrt{2}$. For all $\gamma \ge \gamma^*$, X_γ is obtained as the stabilizing solution of the GRE associated with H_γ,

$$X_\gamma = \begin{pmatrix} x^2\left(x + \frac{1}{2}\right) & -x^2 \\ -x^2 & x \end{pmatrix} \quad \text{where} \quad x = \frac{1}{1 + \sqrt{2 - \gamma^{-2}}}. \tag{5.1}$$

Meanwhile $Y_\gamma = 0$ for all $\gamma > \gamma^*$. Hence (C2)–(C4) reduce to the sole condition $X_\gamma \ge 0$, which is satisfied for all $\gamma \ge \gamma^*$ as it can be seen from (5.1). Consequently, $\gamma_{opt} = 1/\sqrt{2}$ and (C1) is the first condition to fail as γ decreases.

Example 5.5. This second example shows that (C3) can fail before (C4) in the GP context. That is, $\|X_\gamma\|$ or $\|Y_\gamma\|$ can become unbounded while $\rho(X_\gamma Y_\gamma)$ remains smaller than γ^2. Consider this time the plant $G(s)$ defined by the parameters $B_1 = C_1 = I_2, D_{22} = 0$, and

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \qquad B_2 = C_2^T = \begin{pmatrix} 1 \\ 1 \end{pmatrix};$$
$$D_{11} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}; \qquad D_{12} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \qquad D_{21} = (1 \quad 0). \tag{5.2}$$

It is easily verified that (A, B_2, C_2) is stabilizable and detectable. Via the reparametrization ((3.1) and (3.3)) we find

$$U_1 = V_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad V_1 = U_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{and } \bar{H}_\gamma = \bar{J}_\gamma = \begin{pmatrix} 0 & -1 + \gamma^{-2} \\ -1 & 0 \end{pmatrix}.$$

Hence $\gamma^* = 1$ and for $\gamma > 1, W_X(\gamma) = W_Y(\gamma) = \sqrt{1 - \gamma^{-2}}$ and

$$X_\gamma = U_1 W_X^{-1} U_1^T = \begin{pmatrix} \frac{1}{\sqrt{1-\gamma^{-2}}} & 0 \\ 0 & 0 \end{pmatrix}; \qquad Y_\gamma = V_1 W_X^{-1} V_1^T = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{\sqrt{1-\gamma^{-2}}} \end{pmatrix}.$$

Consequently, $\rho(X_\gamma Y_\gamma) \equiv 0$ on $(\gamma^*, +\infty)$ while $\gamma_{opt} = 1$ since (C3) holds if and only if $\gamma > 1$. Thus, (C3) fails at γ_{opt} while (C4) is trivially satisfied everywhere. □

Incidentally, (C2) and (C3) never fail before (C4) in the restrictive framework of the SP. This fact critically relies on the particular duality relationship between H_γ and J_γ in the SP context.

THEOREM 5.6. *With the Standard Problem assumptions of* §1, $\rho(X_\gamma Y_\gamma)$ *cannot remain bounded when* $\|X_\gamma\|$ *or* $\|Y_\gamma\|$ *become unbounded. Consequently,* (C2) and (C3) *cannot fail before* (C4) *as* γ *decreases from* $+\infty$.

Proof. See Appendix C. □

6. Computation of γ_{opt} by a Newton method. The concavity results of §5 readily suggest a Newton scheme to locate the first failure of condition (C3). To extend such a scheme to the computation of γ_{opt}, (C4) must also be turned into a concave constraint. Such a reformulation is attempted in [16] but the concave criterion proposed therein has two major drawbacks: it applies only in the region where (C3) is satisfied and it mixes X_γ, Y_γ, and their pseudoinverses. In this section, we introduce an alternative criterion that combines (C2)–(C4) into a single concave constraint. First introduced in [10], this criterion applies to the whole region $\gamma > \gamma^*$ and solely involves the pseudoinverses of X_γ and Y_γ.

THEOREM 6.1. *Let* $U = (U_1, U_2)$ *and* $V = (V_1, V_2)$ *be orthogonal transformations of* $\mathbb{R}^{n\times n}$ *satisfying*

$$\text{Im } U_2 = \mathcal{V}_0(\hat{C}_1, \hat{A}); \qquad \text{Im } V_2 = \mathcal{V}_0(\tilde{B}_1^T, \tilde{A}^T). \tag{6.1}$$

With γ^* *given by* (3.12), *define for* $\gamma > \gamma^*$ *the matrix-valued function* Z *by*

$$Z(\gamma^{-1}) := \begin{pmatrix} W_X(\gamma) & \gamma^{-1}U_1^T V_1 \\ \gamma^{-1}V_1^T U_1 & W_Y(\gamma) \end{pmatrix}, \tag{6.2}$$

where W_X *denotes the continuous extension of* $\bar{X}_\gamma^{-1}$ *in* (3.7) *and* W_Y *that of* $\bar{Y}_\gamma^{-1}$ *in* (3.8). *Then* Z *is a concave function of the parameter* γ^{-1} *on the half line* $\gamma > \gamma^*$. *Moreover, there is equivalence between*

(1) *There exists a stabilizing controller* $K(s)$ *such that* $\|\mathcal{F}(K, G)\|_\infty < \gamma$;

(2) H_γ *and* J_γ *have no eigenvalue on* $\mathbb{C}_0$ *and* $Z(\gamma^{-1})$ *is positive definite.*

Proof. Both W_X, W_Y and therefore $Z(\gamma^{-1})$ are well defined for $\gamma > \gamma^*$ and the concavity follows from

$$\frac{d^2 Z}{d(\gamma^{-1})^2} = \begin{pmatrix} \dfrac{d^2 W_X}{d(\gamma^{-1})^2} & 0 \\ 0 & \dfrac{d^2 W_Y}{d(\gamma^{-1})^2} \end{pmatrix}$$

together with Corollary 4.3. As of the equivalence of (1) and (2), it can be found in [10]. □

Hence, the condition $Z(\gamma^{-1}) > 0$ emerges as a more natural and compact characterization of suboptimal γ's. Moreover, it synthesizes (C2)–(C3) into a single concave constraint so that the computation of γ_{opt} reduces to finding the zero-crossing of a concave function. Given some initial guess in the interval (γ^*, γ_{opt}), this problem can be solved by a Newton method with guaranteed quadratic convergence. Details of implementation appear in [8]. The only delicate issue is the initialization of the algorithm. Fortunately, Theorem 3.2 provides explicit formulas for γ^*. To estimate γ^* indeed, it suffices to solve the associated H_2 problem ($\gamma = +\infty$) and to compute the L_∞ norm of two transfer functions, a task for which quadratically convergent algorithms are also available [3].

7. Conclusion. The dependence on γ of the solutions to the H_∞ control problem has been precisely characterized. As a result, insight was gained into the structure, singularities, inertia, and variations of these solutions, and into their behavior near the optimum. This information is valuable for numerically stable testing of the suboptimality conditions and for the computation of the optimal γ. In particular, the concavity properties established in §§4 and 6 allow the design of Newton algorithms that quadratically converge to γ_{opt}.

Appendix A.

Proof of Theorem 3.2. Using (3.5), it is easily verified that $\Lambda(H_\gamma) \cap \mathbb{C}_0 = \Lambda(\bar{H}_\gamma) \cap \mathbb{C}_0$ with $\bar{H}_\gamma$ as in (4.1). Now, $(\hat{A}_{11}, \hat{B}_{21})$ inherits the stabilizability of $(\hat{A}, \hat{B}_2)$ and $(\hat{C}_{11}, -\hat{A}_{11})$ is detectable from (3.5) and Assumption 3 of §1. Consequently [12], (3.11) has a unique symmetric nonnegative definite stabilizing solution Z. Now, it is easily verified that

$$-\begin{pmatrix} I & -Z \\ 0 & I \end{pmatrix} \bar{H}_\gamma \begin{pmatrix} I & Z \\ 0 & I \end{pmatrix}$$
$$= \begin{pmatrix} \hat{A}_Z - F R_\gamma^{-1} \hat{D}_{11}^T \hat{C}_{11} & -F R_\gamma^{-1} F^T \\ \gamma^2 \hat{C}_{11}^T S_\gamma^{-1} \hat{C}_{11} & -\hat{A}_Z^T + \hat{C}_{11}^T \hat{D}_{11} R_\gamma^{-1} F^T \end{pmatrix} := H_Z(\gamma),$$

where $F := \hat{B}_{11} + Z\hat{C}_{11}^T \hat{D}_{11}$. This last identity shows that $-\bar{H}_\gamma$ and $H_Z(\gamma)$ share the same spectrum. Moreover, from the Bounded Real Lemma [1], [2] we know that

$$\Lambda(H_Z(\gamma)) \cap \mathbb{C}_0 = \emptyset \quad \text{iff } \|\hat{D}_{11} + \hat{C}_{11}(sI - \tilde{A}_Z)^{-1} F\|_\infty < \gamma.$$

Consequently, $\bar{H}_\gamma$ or equivalently H_γ has no eigenvalue on $\mathbb{C}_0$ if and only if $\gamma > \gamma_H := \|\hat{D}_{11} + \hat{C}_{11}(sI - \tilde{A}_Z)^{-1} F\|_\infty$. Note that $\gamma_H < +\infty$ since $\tilde{A}_Z$ is stable. □

Appendix B.

Proof of Lemma 5.1. From (2) of Theorem 4.1, the eigenvalues of W_X are monotonically increasing functions of γ. Hence it suffices to show that the singularities of W_X are isolated. To this purpose, suppose that $W_{X,0} := W_X(\gamma_0)$ is singular and let $\dot{W}_{X,0} := (dW_X/d\alpha)(\gamma_0)$ denote its derivative. To prove that γ_0 is isolated, it is sufficient to show that Ker $W_{X,0} \cap$ Ker $\dot{W}_{X,0} = \{\vec{0}\}$. This is established by contradiction.

Assume this intersection is nontrivial and spanned by the columns of a full-rank matrix L. Then $W_{X,0}L = 0$ together with (4.2) provides $L^T(\hat{B}_{21}\hat{B}_{21}^T - \hat{B}_{11}R_\gamma^{-1}\hat{B}_{11}^T)L = 0$, and $\dot{W}_{X,0}L = 0$ together with (4.5) provides $L^T \hat{B}_{11} R_\alpha^{-2} \hat{B}_{11}^T L = 0$. In turn, these two identities yield $\hat{B}_{11}^T L = \hat{B}_{21}^T L = 0$ and postmultiplying (4.2) and (4.5) by L then provides $\dot{W}_{X,0} A_W L = -\dot{W}_{X,0}\hat{A}_{11}^T L = 0$ and $W_{X,0}\hat{A}_{11}^T L = 0$, respectively. Consequently, Im $L =$ Ker $W_{X,0} \cap$ Ker $\dot{W}_{X,0}$ is $\hat{A}_{11}^T$-invariant. Now, $\binom{W_{X,0}}{I}$ spans $\mathcal{X}_-(\bar{H}_\gamma)$, and hence $\binom{W_{X,0}L}{L} = \binom{0}{L}$ is a subspace of $\mathcal{X}_-(\bar{H}_\gamma)$. Observing that

$$\bar{H}_\gamma \begin{pmatrix} 0 \\ L \end{pmatrix} = \begin{pmatrix} 0 \\ -\hat{A}_{11}^T L \end{pmatrix},$$

it follows that Im L is antistable $\hat{A}_{11}^T$-invariant. Together with $\hat{B}_{21}^T L = 0$, this contradicts the stabilizability of $(\hat{A}_{11}, \hat{B}_{21})$, which is inherited from that of $(\hat{A}, \hat{B}_2)$ or equivalently (A, B_2). □

Appendix C.

Proof of Theorem 5.6. The proof is by contradiction. Assume for instance that X_γ is unbounded at γ_{opt}, that is, $W_{X,0} = W_X(\gamma_{opt})$ is singular. Introduce a matrix L whose

columns form a basis of Ker $W_{X,0}$ and let $Z(.)$ be the function defined in Theorem 6.1. Since $Z(\gamma_{opt}^{-1}) \geq 0$, we have, for all matrix M of compatible dimensions,

$$0 \leq (L^T \quad M^T) Z(\gamma_{opt}^{-1}) \begin{pmatrix} L \\ M \end{pmatrix} = M^T W_{Y,0} M + \gamma_{opt}^{-1}(L^T U_1^T V_1 M + M^T V_1^T U_1 L).$$

This requires $V_1^T U_1 L = 0$, or equivalently Im $U_1 L \subset$ Im $V_2 = \mathcal{V}_0(\tilde{B}_1^T, \tilde{A}^T)$. Observing that $(\tilde{A}^T, \tilde{B}_1^T) = (A^T, B_1^T)$ in the SP context, it follows that Im $U_1 L$ is A^T-invariant and that $B_1^T U_1 L = 0$. Now, pre- and post-multiply (4.2) by L^T and L, respectively, and use the fact that $W_{X,0} L = 0$. This gives $L^T(\hat{B}_{21}\hat{B}_{21}^T - \gamma^{-2}\hat{B}_{11}\hat{B}_{11}^T)L = 0$, or equivalently (using (3.5)),

$$\text{(C.1)} \quad 0 = (U_1 L)^T(\hat{B}_2\hat{B}_2^T - \gamma^{-2}\hat{B}_1\hat{B}_1^T)(U_1 L) = (U_1 L)^T(B_2 B_2^T - \gamma^{-2} B_1 B_1^T)(U_1 L).$$

This together with $B_1^T U_1 L = 0$ imposes $B_2^T U_1 L = 0$. A contradiction to the stabilizability of (A, B_2) can then be derived by an argument similar to that in the proof of Lemma 5.1. □

REFERENCES

[1] B. D. O. ANDERSON AND S. VONPANITLERD, *Network Analysis and Synthesis*, Prentice Hall, Englewood Cliffs, NJ, 1973.

[2] S. BOYD, V. BALAKRISHNAN, AND P. KABAMBA, *A bisection method for computing the H_∞ norm of a transfer matrix and related problems*, Math. Control Signal Systems, 2 (1989), pp. 207–219.

[3] S. BOYD AND V. BALAKRISHNAN, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞ norm*, in Proc. 28th IEEE CDC, Tampa, Florida, 1989, pp. 954–955.

[4] J. C. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.

[5] B. FRANCIS, *A course in H_∞ control theory*, Lecture notes in Control and Information Sciences 88, Springer-Verlag, New York, 1987.

[6] D. F. DELCHAMPS, *A note on the analyticity of the Riccati metric*, in Algebraic and Geometric Methods in Linear Systems Theory, Lecture notes in Applied Mathematics 18, Amer. Math. Soc., Providence, RI, 1980, pp. 37–41.

[7] P. GAHINET AND A. J. LAUB, *Computable bounds for the sensitivity of the algebraic Riccati equation*, SIAM J. Control Optim., 28 (1990), pp. 1461–1480.

[8] P. GAHINET AND P. PANDEY, *A fast and numerically robust algorithm for computing the H_∞ optimum*, in Proc. CDC, 1991, Brighton, UK, pp. 200–205.

[9] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[10] K. GLOVER, D. J. N. LIMEBEER, J. C. DOYLE, E. M. KASENALLY, AND M. G. SAFONOV, *A characterization of all solutions to the four-block general distance problem*, SIAM J. Control Optim., 29 (1991), pp. 283–324.

[11] M. GREEN, K. GLOVER, D. LIMEBEER, AND J. DOYLE, *A J-spectral factorization approach to H_∞ optimization*, SIAM J. Control Optim., 28 (1990), pp. 1350–1371.

[12] V. KUCERA, *Contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344–347.

[13] P. PANDEY, C. KENNEY, A. J. LAUB, AND A. PACKARD, *A gradient method for computing the optimal H_∞ norm*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 887–890.

[14] M. G. SAFONOV, D. J. LIMEBEER, AND R. Y. CHIANG, *Simplifying the H_∞ theory via loop-shifting, matrix-pencil and descriptor concepts*, Internat. J. Control, 50 (1989), pp. 2467–2488.

[15] J. SEFTON AND K. GLOVER, *Pole/zero cancellations in the general H_∞ problem with reference to a two block design*, Systems Control Lett., 14 (1990), pp. 295–306.

[16] C. SCHERER, *H_∞-control by state-feedback and fast algorithms for the computation of optimal H_∞-norms*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 1090–1099.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 648–672, May 1994

004

ADAPTIVE BOUNDARY AND POINT CONTROL OF LINEAR STOCHASTIC DISTRIBUTED PARAMETER SYSTEMS*

T. E. DUNCAN†, B. MASLOWSKI‡, AND B. PASIK-DUNCAN†

Abstract. An adaptive control problem for the boundary or the point control of a linear stochastic distributed parameter system is formulated and solved in this paper. The distributed parameter system is modeled by an evolution equation with an infinitesimal generator for an analytic semigroup. Since there is boundary or point control, the linear transformation for the control in the state equation is also an unbounded operator. The unknown parameters in the model appear affinely in both the infinitesimal generator of the semigroup and the linear transformation of the control. Strong consistency is verified for a family of least squares estimates of the unknown parameters. An Itô formula is established for smooth functions of the solution of this linear stochastic distributed parameter system with boundary or point control. The certainty equivalence adaptive control is shown to be self-tuning by using the continuity of the solution of a stationary Riccati equation as a function of parameters in a uniform operator topology. For a quadratic cost functional of the state and the control, the certainty equivalence control is shown to be self-optimizing; that is, the family of average costs converges to the optimal ergodic cost. Some examples of stochastic parabolic problems with boundary control and a structurally damped plate with random loading and point control are described that satisfy the assumptions for the adaptive control problem solved in this paper.

Key words. stochastic adaptive control, linear stochastic distributed parameter systems, boundary control problems, identification

AMS subject classifications. 93C40, 93C20, 93E12, 60H15

1. Introduction.

An important family of controlled linear, distributed parameter control systems are those with boundary or point control. Perturbations or inaccuracies in the mathematical model can often be effectively modeled by white noise. Since in many control situations there are unknown parameters in these linear, stochastic distributed parameter systems, it is necessary to solve a stochastic adaptive control problem. We now give a brief summary of each of the sections in this paper. In §2 the unknown linear stochastic distributed parameter system is described by an evolution equation where the unknown parameters appear in the infinitesimal generator of an analytic semigroup and the unbounded linear transformation for the boundary control. The noise process is a cylindrical, white noise. Some properties of the optimal control for the infinite-time quadratic cost functional for the associated deterministic system are reviewed, especially the stationary Riccati equation. These results are given in [8], [11], [12], [18]. In §3 an Itô formula is obtained for smooth functions of the solution of a linear or semilinear stochastic distributed parameter system with an analytic semigroup. This result is verified using the Yosida approximation of the infinitesimal generator of the semigroup. While some other Itô formulas in infinite dimensions are available (e.g., [6], [15]), none seems to be appropriate for our applications. In §4 a family of least squares estimates are constructed from the observations of the unknown stochastic system. This family of estimates is shown to be strongly consistent under verifiable conditions. A stochastic differential equation is given for the family of estimates. This verification of the strong consistency of a family of least squares estimates is a generalization of the results in [9], [10]. In §5 the self-tuning and the self-optimizing properties of an adaptive control law are investigated. If an adaptive control is self-tuning, then it is shown that the system satisfies some stability properties and the adaptive control is self-optimizing. The certainty equivalence adaptive control, that is, using the optimal stationary control with the estimates of the

* Received by the editors April 2, 1992; accepted for publication (in revised form) November 23, 1992. This research was partially supported by National Science Foundation grants ECS-8718026, ECS-9102714, and ECS-9113029.

† Department of Mathematics, University of Kansas, Lawrence, Kansas 66045.

‡ Institute of Mathematics, Czech Academy of Sciences, Prague, Czech Republic.

parameters, is shown to be self-optimizing; that is, the optimal ergodic cost is achieved. In §6 some examples are given that satisfy the various assumptions used in this paper.

2. A boundary control model. The unknown linear stochastic distributed parameter system with boundary or point control is formally described by the following stochastic differential equation:

$$\begin{aligned} dX(t;\alpha) &= (A(\alpha)X(t;\alpha) + B(\alpha)U(t))dt + \Phi\, dW(t), \\ X(0;\alpha) &= X_0, \end{aligned} \tag{2.1}$$

where $X(t;\alpha) \in H$; H is a real, separable, infinite-dimensional Hilbert space; $(W(t), t \geq 0)$ is a cylindrical Wiener process on H; $\Phi \in \mathcal{L}(H), \alpha = (\alpha^1, \ldots, \alpha^q)$; and $t \geq 0$.

The probability space is denoted $(\Omega, \mathcal{F}, P)$, where P is a probability measure that is induced from the cylindrical Wiener measure and $\mathcal{F}$ is the P-completion of the Borel σ-algebra on Ω. Let $(\mathcal{F}_t, t \geq 0)$ be an increasing P-complete family of sub-σ-algebras of $\mathcal{F}$ such that X_t is $\mathcal{F}_t$-measurable for $t \geq 0$ and $(\langle \ell, W(t)\rangle, \mathcal{F}_t, t \geq 0)$ is a martingale for each $\ell \in H$. $A(\alpha)$ is the infinitesimal generator of an analytic semigroup on H. For some $\beta \geq 0$, the operator $-A(\alpha) + \beta I$ is strictly positive, so that the fractional powers $(-A(\alpha) + \beta I)^\gamma$ and $(-A(\alpha)^* + \beta I)^\gamma$ and the spaces $D^\gamma_{A(\alpha)} = \mathcal{D}((-A(\alpha) + \beta I)^\gamma)$ and $D^\gamma_{A^*(\alpha)} = \mathcal{D}((-A^*(\alpha) + \beta I)^\gamma)$ with the graph norm topology for $\gamma \in \mathbb{R}$ can be defined. It is assumed that $B(\alpha) \in \mathcal{L}\big(H_1, D^{\varepsilon-1}_{A(\alpha)}\big)$, where H_1 is a real, separable Hilbert space and $\varepsilon \in (0,1)$ (cf. assumption (A4) below). For the solution of (2.1) on $[0,T]$, the control $(U(t), t \in [0,T])$ is an element of $M^p_W(0,T,H_1)$, where $M^p_W(0,T,H_1) = \big\{u : [0,T] \times \Omega \to H_1, u$ is $(\mathcal{F}_t)$-nonanticipative and $E\int_0^T |u(t)|^p\, dt < \infty\big\}$ and $p > \max(2, 1/\varepsilon)$ is fixed.

A selection of the following assumptions are used subsequently.

(A1) The family of unknown parameters are the elements of a compact set $\mathcal{K}$.

(A2) For $\alpha \in \mathcal{K}$, the operator $\Phi^*(-A^*(\alpha) + \beta I)^{-1/2+\delta}$ is Hilbert–Schmidt for some $\delta \in \big(0, \frac{1}{2}\big)$.

(A3) There are real numbers $M > 0$ and $\omega > 0$ such that, for $t > 0$ and $\alpha \in \mathcal{K}$,

$$|S(t;\alpha)|_{\mathcal{L}(H)} \leq Me^{-\omega t}$$

and

$$|A(\alpha)S(t;\alpha)|_{\mathcal{L}(H)} \leq Mt^{-1}e^{-\omega t},$$

where $(S(t;\alpha), t \geq 0)$ is the analytic semigroup generated by $A(\alpha)$.

(A4) For all $\alpha_1, \alpha_2 \in \mathcal{K}$, $\mathcal{D}(A(\alpha_1)) = \mathcal{D}(A(\alpha_2))$, $D^\delta_{A(\alpha_1)} = D^\delta_{A(\alpha_2)}$ and $D^\delta_{A^*(\alpha_1)} = D^\delta_{A^*(\alpha_2)}$ for $\delta \in \mathbb{R}$.

(A5) For each $\alpha \in \mathcal{K}$ and $x \in H$, there is a control $u_{\alpha,x} \in L^2(\mathbb{R}_+, H_1)$ such that

$$y(\cdot) = S(\cdot\,;\alpha)x + \int_0 S(\cdot - t;\alpha)B(\alpha)u_{\alpha,x}(t)\, dt \in L^2(\mathbb{R}_+, H).$$

(A6) The operator $A(\alpha)$ has the form

$$A(\alpha) = F_0 + \sum_{i=1}^{q} \alpha^i F_i,$$

where F_i is a linear, densely defined operator on H for $i = 0, 1, \ldots, q$ such that $\cap_{i=0}^q \mathcal{D}(F_i^*)$ is dense in H.

It is well known that the strong solution of (2.1) may not exist, so usually the mild solution of (2.1) is used, that is,

$$X(t;\alpha) = S(t;\alpha)X_0 + \int_0^t S(t-r;\alpha)B(\alpha)U(r)\,dr + \int_0^t S(t-r;\alpha)\Phi\,dW(r), \tag{2.2}$$

where $S(t;\alpha) = e^{tA(\alpha)}$. The mild solution is equivalent to the following inner product equation: For each $y \in \mathcal{D}(A^*(\alpha))$,

$$\begin{aligned}\langle y, X(t;\alpha)\rangle &= \langle y, X(0)\rangle + \int_0^t \langle A^*(\alpha)y, X(s;\alpha)\rangle\,ds \\ &\quad + \int_0^t \langle \Psi(\alpha)y, U(s)\rangle\,ds + \langle \Phi^* y, W(t)\rangle,\end{aligned} \tag{2.3}$$

where $\Psi(\alpha) = B^*(\alpha) \in \mathcal{L}\big(D^{1-\varepsilon}_{A^*(\alpha)}, H_1\big)$. The following lemma verifies that $(X(t;\alpha), t \in [0,T])$ is a well-defined process in $M^p_W(0,T,H)$.

LEMMA 2.1. *Assume that* (A2) *is satisfied. For* $T > 0$ *and* $\alpha \in \mathcal{K}$, *the processes* $(Z(t;\alpha), t \in [0,T])$ *and* $\hat{Z}(t;\alpha), t \in [0,T])$ *given by the equations*

$$Z(t;\alpha) = \int_0^t S(t-r;\alpha)\Phi\,dW(r), \tag{2.4}$$

$$\hat{Z}(t;\alpha) = \int_0^t S(t-r;\alpha)B(\alpha)U(r)\,dr \tag{2.5}$$

for $U \in M^p_W(0,T,H_1)$ *are elements of* $M^p_W(0,T,H)$, *with versions that have continuous sample paths.*

Proof. Let $|\cdot|_{HS}$ be the Hilbert–Schmidt norm on $\mathcal{L}(H)$. If (e_n) is an orthonormal basis of H, by (A2) we have that

$$\begin{aligned}\int_0^T |S(t;\alpha)\Phi|^2_{HS}\,dt &\le \int_0^T \frac{c}{t^{1-2\delta}} \sum_n |(-A(\alpha)+\beta I)^{-1/2+\delta}\Phi e_n|^2\,dt \\ &= |(-A(\alpha)+\beta I)^{-1/2+\delta}\Phi|^2_{HS} \int_0^T \frac{c}{t^{1-2\delta}}\,dt < \infty,\end{aligned}$$

where c is a constant. Thus $(Z(t;\alpha), t \in [0,T])$ is a well-defined H-valued process. To verify the existence of a continuous modification of $(Z(t;\alpha), t \in [0,T])$, the following processes are introduced:

$$\tilde{W}(t;\alpha) = W(t)[(-A(\alpha)+\beta I)^{-1/2+\delta}\Phi]^*$$

and

$$\tilde{Z}(t;\alpha) = \int_0^t S(t-r;\alpha)\,d\tilde{W}(r;\alpha) \quad \text{for } t \in [0,T].$$

There is a $D^{1/2-\delta}_A$-continuous modification of $(\tilde{Z}(t;\alpha), t \in [0,T])$ [7, Thm. 4]. Thus the process

$$Z(t;\alpha) = \int_0^t S(t-r;\alpha)(-A+\beta I)^{1/2-\delta} d\tilde{W}(t;\alpha)$$

has an H-continuous modification.

Since the inequality

$$|S(t-r;\alpha)B(\alpha)|_{\mathcal{L}(H_1,H)} \leq \frac{c}{(t-r)^{1-\varepsilon}}$$

is satisfied for $0 < r < t < T$, we can apply the Hölder inequality with the exponents p and $q = p/(p-1)$ to the integral (2.5) to verify that $(\hat{Z}(t;\alpha), t \in [0,T])$ is a well-defined H-valued process in $M_W^p(0,T,H)$ with a continuous modification. □

If $A(\alpha) = A^*(\alpha)$ and if $(A(\alpha) - \beta I)^{-1}$ is compact, then assumption (A2) is equivalent to the assumption that

$$\int_0^T t^{-2\delta}|S(t;\alpha)\Phi|_{HS}^2\,dt < \infty$$

for $T > 0$. For notational convenience, the dependence on α is suppressed. By the compactness of the resolvent of A, there is a sequence (λ_k), where $\lambda_k > \lambda_0 > 0$ and $\lambda_k \uparrow \infty$, and an orthonormal basis of (e_k) of H such that $(A - \beta I)e_k = \lambda_k e_k$ for $k \in \mathbb{N}$ and

$$S(t)\cdot = e^{\beta t}\sum_k e^{-\lambda_\kappa t} < \cdot, e_k > e_k.$$

If $\tilde{S}(t) = e^{-\beta t}S(t)$, then

$$\begin{aligned}\int_0^T t^{-2\delta}|\tilde{S}(t)\Phi|_{HS}^2\,dt &= \int_0^T t^{-2\delta}|\Phi^*\tilde{S}(t)|_{HS}^2\,dt \\ &= \sum_k |\Phi^* e_k|^2 \lambda_k^{2\delta-1}\int_0^{\lambda_k T} s^{-2\delta}e^{-2s}\,ds \\ &= \sum_k \left|\Phi^*\lambda_k^{-1/2+\delta}e_k\right|^2 b_k,\end{aligned}$$

where $0 < b_1 \leq b_k \leq b_{k+1}, b_k \to b_\infty < \infty$. Since

$$\sum_k \left|\Phi^*\lambda_k^{-1/2+\delta}e_k\right|^2 = |\Phi^*(-A+\beta I)^{-1/2+\delta}|_{HS}^2,$$

our assertion follows.

Consider the quadratic cost functional

$$J(X_0,U,\alpha,T) = \int_0^T [\langle QX(s),X(s)\rangle + \langle PU(s),U(s)\rangle]\,ds, \tag{2.6}$$

where $T \in (0,\infty], X(0) = X_0, Q \in \mathcal{L}(H), P \in \mathcal{L}(H_1)$ are selfadjoint operators satisfying

$$\langle Qx,x\rangle \geq r_1|x|^2, \tag{2.7}$$

$$\langle Py,y\rangle \geq r_2|y|^2 \tag{2.8}$$

for $x \in H, y \in H_1$ and constants $r_1 > 0$ and $r_2 > 0$. For the deterministic control problem for (2.1) with $\Phi \equiv 0$ and the cost functional (2.6) with $T = +\infty$ assuming (A5), the optimal cost is $\langle V(\alpha)X_0, X_0\rangle$ [8], [12], [18], where V satisfies the formal stationary Riccati equation

$$A^*(\alpha)V(\alpha) + V(\alpha)A(\alpha) - V(\alpha)B(\alpha)P^{-1}\Psi(\alpha)V(\alpha) + Q = 0 \tag{2.9}$$

and $\Psi(\alpha) = B^*(\alpha)$.

Equation (2.9) can be modified to a meaningful inner product equation as

$$\langle A(\alpha)x, Vy\rangle + \langle Vx, A(\alpha)y\rangle - \langle P^{-1}\Psi(\alpha)Vx, \Psi(\alpha)Vy\rangle + \langle Qx, y\rangle = 0 \tag{2.10}$$

for $x, y \in \mathcal{D}(A(\alpha))$. It has been shown [8], [12], [18] that, if (A5) is satisfied, then V is the unique, nonnegative, selfadjoint solution of (2.10) and $V \in \mathcal{L}(H, D_{A^*}^{1-\varepsilon})$. The solution of (2.9) is understood to be the solution of (2.10).

For adaptive control, the control policies $(U(t), t \geq 0)$ that are considered are linear feedback controls, that is,

$$U(t) = K(t)X(t), \tag{2.11}$$

where $(K(t), t \geq 0)$ is an $\mathcal{L}(H, H_1)$-valued process that is uniformly bounded almost surely by a constant $R > 0$. Let $\Delta > 0$ be fixed. It is assumed that the $\mathcal{L}(H, H_1)$-valued process $(K(t), t \geq 0)$ has the property that $K(t)$ is adapted to $\sigma(X(u), u \leq t - \Delta)$ for each $t \geq \Delta$. It is also assumed that $(K(t), t \in [0, \Delta])$ is a deterministic, operator-valued function. For such an admissible adaptive control, there is a unique solution of (2.1) with $K(t) = \tilde{K}(X(s), 0 \leq s \leq t - \Delta)$. If $\Delta = 0$, then (2.1) may not have a unique solution. Furthermore, the delay $\Delta > 0$ accounts for some time that is required to compute the adaptive feedback control law from the observation of the solution of (2.1).

Two more assumptions, (A7) and (A8), are now given that are used for the verification of the strong consistency of a family of least squares estimates of the unknown parameter vector α. Define $\mathbb{K} \subset \mathcal{L}(H, H_1)$ as

$$\mathbb{K} = \{K \in \mathcal{L}(H, H_1) : |K|_{\mathcal{L}(H,H_1)} \leq R\},$$

where R is given above.

Assume that $B(\alpha)$ is either independent of $\alpha \in \mathcal{K}$ or has the form

$$B(\alpha) = \Psi^*(\alpha), \tag{2.12}$$

where $\Psi(\alpha) = \hat{B}^* A^*(\alpha) \in \mathcal{L}\big(D_{A^*(\alpha)}^{1-\varepsilon}, H_1\big)$ and the operator $\hat{B} \in \mathcal{L}\big(H_1, D_{A(\alpha)}^{\varepsilon}\big)$ is given.

(A7) There is a finite-dimensional projection $\tilde{P}$ on H with range in $\cap_{i=1}^q \mathcal{D}(F_i^*)$ such that $i_{\tilde{P}}\Phi\Phi^* i_{\tilde{P}}^* > 0$, where $i_{\tilde{P}} : H \to \tilde{P}(H)$ is the projection map and $B(\alpha)$ is either independent of α or has the form (2.12). In the latter case, there is a finite-dimensional projection $\hat{P}$ on H and a constant $c > 0$ such that

$$|\hat{P}(I + K^*\hat{B}^*)F^*\tilde{P}|_{\mathcal{L}(H)} > c$$

is satisfied for all $F \in \{F_1, \ldots, F_q\}$ and $K \in \mathbb{K}$.

It is easy to verify that, if H is infinite-dimensional, if $\hat{B} \in \mathcal{L}(H_1, H)$ is compact, and if $(F_i^*)^{-1} \in \mathcal{L}(H)$ for $i = 1, 2, \ldots, q$, then (A7) is satisfied.

Let $(U(t), t \geq 0)$ be an admissible control, denoted generically as $U(t) = K(t)X(t)$, where $(X(t), t \geq 0)$ is the (unique) mild solution of (2.1) using the above admissible control. Let

$$\mathcal{A}(t) = (a_{ij}(t)) \tag{2.13}$$

and

$$\tilde{\mathcal{A}}(t) = (\tilde{a}_{ij}(t)), \tag{2.14}$$

where

$$a_{ij}(t) = \int_0^t \langle \tilde{P}F_iX(s), \tilde{P}F_jX(s)\rangle\, ds \tag{2.15a}$$

if B does not depend on α or

$$a_{ij}(t) = \int_0^t \langle \tilde{P}(F_i + F_i\hat{B}K(s))X(s), \tilde{P}(F_j + F_j\hat{B}K(s))X(s)\rangle\, ds \tag{2.15b}$$

if $B(\alpha)$ has the form (2.12) and

$$\tilde{a}_{ij}(t) = \frac{a_{ij}(t)}{a_{ii}(t)}. \tag{2.16}$$

It is easy to verify that the integrations in (2.15a) and (2.15b) are well defined.

For the verification of the strong consistency of a family of least squares estimates of the unknown parameter vector, the following assumption is used.

(A8) For each admissible adaptive control law, $(\tilde{\mathcal{A}}(t), t \geq 0)$ satisfies

$$\lim_{t\to\infty} \inf |\det \tilde{\mathcal{A}}(t)| > 0 \quad \text{a.s.}$$

3. An Itô formula. In this section, an Itô formula is verified for a smooth function of the solution of (2.1). While some Itô formulas are available for evolution equations (e.g., [6], [15]), apparently no result is available for an equation of the form (2.1).

Since the parameter vector α is fixed in this section, the dependence of (2.1) on α is suppressed throughout this section. The Itô equation obtained here is verified by an approximation of (2.1) using the resolvent. For $\lambda > \beta$, let $R(\lambda)$ be defined by

$$R(\lambda) = \lambda R(\lambda, A), \tag{3.1}$$

where $R(\lambda, A) = (\lambda I - A)^{-1}$ is the resolvent of A. By assumptions (A2), $R(\lambda)^{1/2-\delta}\Phi$ is Hilbert–Schmidt and $R(\lambda)\Phi$ is Hilbert–Schmidt, so there is an H-valued Wiener process $(W_\lambda(t), t \geq 0)$ defined by

$$W_\lambda(t) = W(t)\Phi^* R^*(\lambda), \tag{3.2}$$

where $W_\lambda(1)$ has the nuclear covariance $R(\lambda)\Phi\Phi^* R^*(\lambda)$. Consider the stochastic differential equation

$$\begin{aligned} dX_\lambda(t) &= AX_\lambda(t)\, dt + R(\lambda)BU(t)\, dt + R(\lambda)\, dW_\lambda(t), \\ X_\lambda(0) &= R(\lambda)X_0, \end{aligned} \tag{3.3}$$

where $\lambda > \beta$. It is shown that (3.3) has a strong solution.

LEMMA 3.1. *For $\lambda > \beta$, the stochastic equation* (3.3) *has a unique strong solution on* $[0, T]$, *that is,*

$$\int_0^T |AX_\lambda(t)|\, dt < \infty \quad \text{a.s.} \tag{3.4}$$

and

$$\begin{aligned} X_\lambda(t) = R(\lambda)X_0 &+ \int_0^t AX_\lambda(s)\, ds + \int_0^t R(\lambda)BU(s)\, ds \\ &+ R(\lambda)W_\lambda(t) \quad \text{a.s.} \end{aligned} \tag{3.5}$$

Proof. To verify that the mild solution of (3.3) satisfies (3.5), it is necessary to show that

$$\int_0^T \int_0^t |AS(t-s)R(\lambda)BU(s)|\,ds\,dt < \infty \quad \text{a.s.} \tag{3.6}$$

and

$$\int_0^T \int_0^t |AS(t-s)R(\lambda)|^2_{\mathcal{L}(H)}\,ds\,dt < \infty \quad \text{a.s.} \tag{3.7}$$

We have

$$\begin{aligned} &|AS(t-s)R(\lambda)B|_{\mathcal{L}(H_1,H)} \\ &\le |AS(t-s)|_{\mathcal{L}(D_A^{\varepsilon},H)}|R(\lambda)|_{\mathcal{L}(D_A^{\varepsilon-1},D_A^{\varepsilon})}|B|_{\mathcal{L}(H_1,D_A^{\varepsilon-1})} \\ &\le \frac{c}{(t-s)^{1-\varepsilon}}, \end{aligned}$$

so that

$$\begin{aligned} &\int_0^T \int_0^t |AS(t-s)R(\lambda)BU(s)|\,ds\,dt \\ &\le T\left(\int_0^T \frac{c}{t^{q(1-\varepsilon)}}\,dt\right)^{1/q}\left(\int_0^T |U(s)|^p\,ds\right)^{1/p}, \end{aligned}$$

which verifies (3.6). Since $|AS(t-s)R(\lambda)|_{\mathcal{L}(H)}$ is bounded for $0 \le s \le t \le T$, (3.7) is satisfied. Use the Fubini theorem to compute $\int_0^t AX_\lambda(s)\,ds$ as in [6], [14] to verify (3.5). □

Now it is shown that a sequence of processes can be obtained from solutions of (3.5) as $\lambda \to \infty$, which converges to (2.2).

LEMMA 3.2. *There is a sequence* (λ_n) *such that* $\lambda_n \uparrow +\infty$, *and, for* $t \in [0,T]$,

$$\lim_{n\to\infty} X_{\lambda_n}(t) = X(t) \quad \text{a.s.} \tag{3.8}$$

and

$$\sup\{|X_{\lambda_n}(t)| : \lambda_n > 0, t \in [0,T]\} < \infty \quad \text{a.s.}, \tag{3.9}$$

where $(X_{\lambda_n}(t), t \in [0,T])$ *satisfies* (3.5) *and* $(X(t), t \in [0,T])$ *satisfies* (2.1).

Proof. The Yosida approximation implies that $R(\lambda)X_0 \to X_0$ as $\lambda \to \infty$ for all $X_0 \in H$ and $|R(\lambda)|_{\mathcal{L}(H)} \le c$ for some $c > 0$. Since

$$|S(t-s)(R(\lambda)-I)BU(s)| \le \frac{K}{(t-s)^{1-\varepsilon}}|U(s)|, \tag{3.10}$$

where $K > 0$ is independent of $\lambda > \beta$, the Hölder inequality implies that

$$\sup_{t\in[0,T]}\left|\int_0^t S(t-s)R(\lambda)BU(s)\,ds\right| \le C_T\left(\int_0^T |U(s)|^p\,ds\right)^{1/p} \tag{3.11}$$

for some $C_T > 0$ and the dominated convergence theorem implies that

$$\lim_{\lambda\to\infty}\left|\int_0^t S(t-s)(R(\lambda)-I)BU(s)\,ds\right| = 0 \tag{3.12}$$

for $t \in [0,T]$. Let $\bar{W}_\lambda(t) = W(t)\Phi^*(I - R^2(\lambda))^*$, and, recalling (2.4), we have

$$\begin{aligned} &E \sup_{t\in[0,T]} \left| Z(t) - \int_0^t S(t-r)R(\lambda)\, dW_\lambda(r) \right|^2 \\ &= E \sup_{t\in[0,T]} \left| \int_0^t S(t-r)\, d\bar{W}_\lambda(r) \right|^2. \end{aligned} \tag{3.13}$$

By Theorem 2.1 of [16], the right-hand side of (3.13) tends to zero as $\lambda \to \infty$ if the trace of the covariance operator of $(-A+\beta I)^{-(1/2)+\delta}W_\lambda(1)$ tends to zero as $\lambda \to \infty$. If (e_n) is an orthonormal basis of H, then

$$\begin{aligned} &\mathrm{Tr}[(-A+\beta I)^{-(1/2)+\delta}(I - R^2(\lambda))\Phi\Phi^*(I - R^2(\lambda))^*(-A^* + \beta I)^{-(1/2)+\delta}] \\ &= |(I - R^2(\lambda))(-A+\beta I)^{-(1/2)+\delta}\Phi|_{HS}^2 \\ &= \sum_n |(I - R^2(\lambda))(-A+\beta I)^{-(1/2)+\delta}\Phi e_n|^2. \end{aligned} \tag{3.14}$$

This infinite series converges to zero as $\lambda \to \infty$ because

$$|(I - R^2(\lambda))(-A+\beta I)^{-(1/2)+\delta}\Phi e_n|^2 \to 0$$

for $n \in \mathbb{N}$, and the series is dominated by

$$\sum_n 2(c^4+1)|(-A+\beta I)^{-(1/2)+\delta}\Phi e_n| = 2(c^4+1)|(-A+\beta I)^{-(1/2)+\delta}\Phi|_{HS}^2 < \infty.$$

Thus the right-hand side of (3.13) converges to zero, and there is a sequence (λ_n) such that

$$\sup_{t\in[0,T]} \left| \int_0^t S(t-r)R(\lambda_n)\, dW_{\lambda_n}(t) - Z(t) \right|^2 \to 0 \quad \text{a.s.}$$

as $n \to \infty$. Therefore, for $t \in [0,T]$,

$$\lim_{n\to\infty} X_{\lambda_n}(t) = X(t) \quad \text{a.s.}$$

by (3.12) and (3.9) is satisfied because

$$\sup_{\lambda>\beta, t\in[0,T]} \left| \int_0^t S(t-s)R(\lambda)BU(s)\, ds \right| < \infty \quad \text{a.s.}$$

by (3.11). □

An Itô formula is now verified for smooth functions of a solution of (2.1).

LEMMA 3.3. *Let $V \in C^{1,2}([0,T]\times H, \mathbb{R})$ be such that $V_x(t,x) \in D_{A^*}^{1-\varepsilon}$ for all $t \in (0,T)$ and $V_x(t,\cdot) : H \to D_{A^*}^{1-\varepsilon}$ is continuous. Assume that the function $\langle Ax, V_x(t,x)\rangle$ for $x \in \mathcal{D}(A)$ can be extended to a continuous function $h : [0,T]\times H \to \mathbb{R}$, the following limit exists*:

$$\lim_{\lambda\to\infty} \mathrm{Tr}\, V_{xx}(t,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 = \pi(t,x) < \infty \tag{3.15}$$

and the map

$$x \mapsto \mathrm{Tr}\, V_{xx}(t,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 \tag{3.16}$$

is continuous on H uniformly with respect to $\lambda \geq \beta$ and

$$
\begin{aligned}
&|h(t,x)| + |\operatorname{Tr} V_{xx}(t,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2| + |V(t,x)| \\
&\quad + |V_x(t,x)|_{D^{1-\varepsilon}_{A^*}} + |V_{xx}(t,x)|_{\mathcal{L}(H)} + |V_t(t,x)| \leq k(1+|x|^p)
\end{aligned} \tag{3.17}
$$

for $(t,x) \in (0,T) \times H, \lambda \geq \beta$ where $k > 0$ and $p > 0$. Then

$$
\begin{aligned}
&V(t,X(t)) - V(\tau, X(\tau)) \\
&\quad = \int_\tau^t \Big[h(s,X(s)) + V_s(s,X(s)) + \langle U(s), \Psi V_x(s,X(s))\rangle \\
&\qquad + \frac{1}{2}\pi(s,X(s))\Big]\, ds + \int_\tau^t \langle \Phi^* V_x(s,X(s)), dW(s)\rangle \quad \text{a.s.},
\end{aligned} \tag{3.18}
$$

where $0 \leq \tau \leq t \leq T$, $\Psi = B^*$ and $(X(t), t \in [0,T])$ satisfies (2.1).

Proof. The verification of (3.18) is accomplished by using a sequence of processes that satisfy (3.5). Since $(X_\lambda(t), t \in [0,T])$ is a strong solution of (3.5), the Itô formula [6] can be applied to $(V(X_\lambda(t)), t \in [0,T])$ to obtain

$$
\begin{aligned}
V(t,X_\lambda(t)) - V(\tau, X_\lambda(\tau)) = \int_\tau^t \Big[& h(s,X_\lambda(s)) + V_s(s,X_\lambda(s)) \\
& + \langle U(s), \Psi R^*(\lambda) V_x(s,X_\lambda(s))\rangle \\
& + \frac{1}{2} \operatorname{Tr} R^*(\lambda) V_{xx}(s,X_\lambda(s)) R^2(\lambda)\Phi\Phi^* R^*(\lambda)\Big]\, ds \\
& + \int_\tau^t \langle R^*(\lambda) V_x(s,X_\lambda(s)), dW_\lambda(s)\rangle \quad \text{a.s.}
\end{aligned} \tag{3.19}
$$

It suffices to assume that $(U(t), t \in [0,T])$ is uniformly bounded, almost surely. Lemma 3.2 verifies that

$$
\lim_{n\to\infty} V(t, X_{\lambda_n}(t)) = V(t,X(t)) \quad \text{a.s.}, \tag{3.20}
$$

$$
\lim_{n\to\infty} V_s(s, X_{\lambda_n}(s)) = V_s(s,X(s)) \quad \text{a.s.}, \tag{3.21}
$$

$$
\lim_{n\to\infty} V(\tau, X_{\lambda_n}(\tau)) = V(\tau,X(\tau)) \quad \text{a.s.}, \tag{3.22}
$$

$$
\lim_{n\to\infty} h(s, X_{\lambda_n}(s)) = h(s,X(s)) \quad \text{a.s.} \tag{3.23}
$$

for $0 \leq \tau \leq s \leq t \leq T$, where $\lambda_n \to \infty$. Let $\mathcal{L}_\lambda V$ and $\mathcal{L}V$ be defined as

$$
\begin{aligned}
\mathcal{L}_\lambda V(s,x) = & \langle U(s), \Psi R^*(\lambda) V_x(s,x)\rangle \\
& + \tfrac{1}{2} \operatorname{Tr} R^*(\lambda) V_{xx}(s,x) R^2(\lambda)\Phi\Phi^* R^*(\lambda)
\end{aligned}
$$

and

$$
\mathcal{L}V(s,x) = \langle U(s), \Psi V_x(s,x)\rangle + \tfrac{1}{2}\pi(s,x)
$$

for $(s,x) \in [0,T] \times H$. We have

$$
\begin{aligned}
&|\mathcal{L}_\lambda V(s,X_\lambda(s)) - \mathcal{L}V(s,X(s))| \\
&\leq |\mathcal{L}_\lambda V(s,X(s)) - \mathcal{L}V(s,X(s))| + |\mathcal{L}_\lambda V(s,X(s)) - \mathcal{L}_\lambda V(s,X_\lambda(s))|.
\end{aligned} \tag{3.24}
$$

From (3.15), we have that

$$\lim_{\lambda\to\infty} \mathcal{L}_\lambda V(s,X(s)) = \mathcal{L}V(s,X(s)) \quad \text{a.s.} \tag{3.25}$$

Furthermore,

$$\begin{aligned} &|\mathcal{L}_\lambda V(s,x) - \mathcal{L}_\lambda V(s,y)| \\ &\le |U(s)|\ |\Psi|_{\mathcal{L}(D^{1-\varepsilon}_{A^*},H_1)} |R^*(\lambda)|_{\mathcal{L}(H)} |V_x(s,x) - V_x(s,y)|_{D^{1-\varepsilon}_{A^*}} \\ &\quad + \tfrac{1}{2}|\operatorname{Tr} V_{xx}(s,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 \\ &\quad - \operatorname{Tr} V_{xx}(s,y)R^2(\lambda)\phi\phi^*(R^*(\lambda))^2|. \end{aligned} \tag{3.26}$$

The right-hand side of (3.26) tends to zero as $x \to y$ in H uniformly with respect to $\lambda \ge \beta$ by (3.15). Thus the second term on the right-hand side of (3.24) converges to zero almost surely. Choosing a sequence (λ_n) such that $\lambda_n \to \infty$, from Lemma 3.2 we obtain by (3.9), (3.17) and the dominated convergence theorem

$$\begin{aligned} &\lim_{n\to\infty} \int_\tau^t [h(s,X_{\lambda_n}(s)) + V_s(s,X_{\lambda_n}(s)) + \mathcal{L}_{\lambda_n} V(s,X_{\lambda_n}(s))\,ds \\ &= \int_\tau^t [h(s,X(s)) + V_s(s,X(s)) + \mathcal{L}V(s,X(s))]\,ds \quad \text{a.s.} \end{aligned} \tag{3.27}$$

Furthermore,

$$\begin{aligned} &E\left|\int_\tau^t \langle R^*(\lambda_n)V_x(s,X_{\lambda_n}(s)), dW_{\lambda_n}(s)\rangle - \int_\tau^t \langle \Phi^* V_x(s,X(s)), dW(s)\rangle\right|^2 \\ &= E\int_\tau^t |\Phi^*(R^*(\lambda_n))^2 V_x(s,X_{\lambda_n}(s)) - \Phi^* V_x(s,X(s))|^2\,ds \\ &\le 2E\int_\tau^t [|\Phi^*(R^*(\lambda_n))^2(V_x(s,X_{\lambda_n}(s)) - V_x(s,X(s))|^2 \\ &\quad + |[\Phi^*(R^*(\lambda))^2 - \Phi^*]V_x(s,X(s))|^2]\,ds. \end{aligned} \tag{3.28}$$

The right-hand side of (3.28) tends to zero as $n \to \infty$ by Lemma 3.2 and (3.17). Thus there is subsequence (λ_{n_j}) such that

$$\lim_{j\to\infty} \int_\tau^t \langle R^*(\lambda_{n_j})V_x(s,X_{\lambda_{n_j}}(s)), dW_{\lambda_{n_j}}(s)\rangle = \int_\tau^t \langle \Phi^* V_x(s,X(s)), dW(s)\rangle \quad \text{a.s.}$$

Thus (3.20)–(3.23), (3.27) verifies (3.18). □

Now some of the hypotheses of Lemma 3.3 are replaced by ones that are more easily verified while still obtaining the same conclusion.

PROPOSITION 3.4. *Assume that* (A2) *is satisfied. Let* $V \in C^{1,2}([0,T]\times H)$ *be such that* $V_x(t,x) \in D^{1-\varepsilon}_{A^*}$, $V_x(t,\cdot) : H \to D^{1-\varepsilon}_{A^*}$ *is continuous,* $\langle Ax, V_x(t,x)\rangle$ *for* $x \in \mathcal{D}(A)$ *can be extended to a continuous function* $h : [0,T]\times H \to \mathbb{R}$ *and*

$$|h(t,x)| + |V(t,x)| + |V_x(t,x)|_{D^{1-\varepsilon}_{A^*}} + |V_{xx}(t,x)|_{\mathcal{L}(H)} + |V_t(t,x)| \le k(1+|x|^p) \tag{3.29}$$

for $(t,x) \in [0,T)\times H$ *and* $p > 0, k > 0$. *Assume that one of the following three conditions is satisfied*:

(i) Φ *is Hilbert–Schmidt*;

(ii) $V_{xx}(t,x)$ *is nuclear,* $V_{xx}(t,\cdot)$ *is continuous in the norm* $|\cdot|_1$ *of nuclear operators and* $|V_{xx}(t,x)|_1 \le k(1+|x|^p)$ *for* $(t,x) \in (0,T) \times H$, *where* $k > 0$ *and* $p > 0$;

(iii) $V_{xx}(t,x) \in \mathcal{L}\big(D_A^{\delta-(1/2)}, D_{A^*}^{(1/2)-\delta}\big)$ *for* $(t,x) \in [0,T] \times H$, *the function*

$$L(\cdot) = (R^*(\beta))^{-(1/2)+\delta} V_{xx}(t,\cdot)(R(\beta))^{-(1/2)+\delta} : H \to \mathcal{L}(H)$$

is continuous and $|L(x)|_{\mathcal{L}(H)} \le k(1+|x|^p)$ *is satisfied for* $t > 0$ *and* $x \in H$.

Then (3.18) *is satisfied, where, for* (i) *and* (ii), $\pi(t,x) = \operatorname{Tr} V_{xx}(t,x)\Phi\Phi^*$ *and, for* (iii), $\pi(t,x) = \operatorname{Tr}(R^*(\beta))^{\delta-(1/2)} V_{xx}(t,x)\Phi\Phi^*(R^*(\beta))^{(1/2)-\delta}$.

Proof. By Lemma 2.3, it suffices to show that (i), (ii), or (iii) implies (3.15)–(3.17). Assume that (i) is satisfied. Then

$$\begin{aligned} &|\operatorname{Tr}[V_{xx}(s,x)R^2(\lambda)\Phi\Phi^*(R(\lambda))^2 - V_{xx}(s,x)\Phi\Phi^*]| \\ &= |\operatorname{Tr}[(R^*(\lambda))^2 V_{xx}(s,x)R^2(\lambda)\Phi\Phi^* - V_{xx}(s,x)\Phi\Phi^*]| \\ &\le \sum_j |(R^*(\lambda))^2 V_{xx}(s,x)R^2(\lambda) - V_{xx}(s,x))\Phi\Phi^* e_j|, \end{aligned} \tag{3.30}$$

where (e_j) is an orthonormal basis in H that includes the eigenvectors of $\Phi\Phi^*$. The series on the right-hand side of (3.30) is dominated by

$$\text{const} \sum_j |\Phi\Phi^* e_j| < \infty,$$

so the dominated convergence theorem implies that the series in (3.30) converges to zero as $\lambda \to \infty$. This verifies (3.15). Since

$$|\operatorname{Tr}(V_{xx}(s,x) - V_{xx}(s,y))R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2| \le c|V_{xx}(s,x) - V_{xx}(s,y)|_{\mathcal{L}(H)} \operatorname{Tr} \Phi\Phi^*,$$

where $c > 0$ does not depend on $\lambda \ge \beta$, (3.16) is verified. Equation (3.17) follows from (3.29).

Assume that (ii) is satisfied. Then

$$\begin{aligned} &|\operatorname{Tr} V_{xx}(s,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 - \operatorname{Tr} V_{xx}\Phi\Phi^*| \\ &= |\operatorname{Tr}[R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 - \Phi\Phi^*]V_{xx}(s,x)| \\ &\le \sum_j |(R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 - \Phi\Phi^*)V_{xx}(s,x)\hat{e}_j|, \end{aligned}$$

where $(\hat{e}_j)$ is an orthonormal basis in H including the eigenvectors of $V_{xx}(s,x)$. Proceeding as for (i), it follows that (3.15) is satisfied. Since

$$\begin{aligned} &|\operatorname{Tr}(V_{xx}(s,x) - V_{xx}(s,y))R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2| \\ &\le |R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2|_{\mathcal{L}(H)} |V_{xx}(s,x) - V_{xx}(s,y)|_1, \end{aligned}$$

$$\begin{aligned} &|\operatorname{Tr} V_{xx}(s,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2| \\ &\le k|R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2|_{\mathcal{L}(H)}(1+|x|^p), \end{aligned}$$

so (3.16), (3.17) are satisfied.

Assume that (iii) is satisfied. Then

$$\begin{aligned} &|\operatorname{Tr} V_{xx}(s,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 - \pi(s,x)| \\ &= |\operatorname{Tr}[(R^*(\lambda))^2 L(x)R^2(\lambda)R^{(1/2)-\delta}(\beta)\Phi\Phi^*(R^*(\beta))^{(1/2)-\delta} \\ &\quad - L(x)R^{(1/2)-\delta}(\beta)\Phi\Phi^*(R(\beta))^{(1/2)-\delta}]|. \end{aligned}$$

Since the operator $R^{(1/2)-\delta}(\beta)\Phi\Phi^*(R^*(\beta))^{(1/2)-\delta}$ is nuclear by (A2), we can choose an orthonormal basis in H that includes the eigenvectors of this operator and proceed as in (i) to verify (3.15). Since

$$|\operatorname{Tr} V_{xx}(s,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2 - \operatorname{Tr} V_{xx}(s,y)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2|$$
$$\leq |(R^*(\lambda))^2(L(x)-L(y))R^2(\lambda)|_{\mathcal{L}(H)}|R^{(1/2)-\delta}(\beta)\Phi\Phi^*(R^*(\beta))^{(1/2)-\delta}|_1$$

and

$$|\operatorname{Tr} V_{xx}(s,x)R^2(\lambda)\Phi\Phi^*(R^*(\lambda))^2|$$
$$\leq |(R^*(\lambda))^2L(x)R^2(\lambda)|_{\mathcal{L}(H)} \cdot |R^{(1/2)-\delta}(\beta)\Phi\Phi^*(R^*(\beta))^{(1/2)-\delta}|_1$$
$$\leq \text{const}(1+|x|^p),$$

these inequalities verify (3.16), (3.17). □

For use of this Itô formula in the adaptive control problem, it is useful to state explicitly the case where $v(x) = \langle Vx, x\rangle$, where $V \in \mathcal{L}(H)$ is a selfadjoint operator.

COROLLARY 3.5. *Let $V \in \mathcal{L}(H)$ be selfadjoint such that $V \in \mathcal{L}(H, D_{A^*}^{1-\varepsilon})$ and $|\langle Vx, Ax\rangle| \leq k|x|^2$ for $x \in \mathcal{D}(A)$, where $k > 0$. Assume that one of the following conditions is satisfied:*

(i) *Φ is Hilbert–Schmidt,*

(ii) *V is nuclear,*

(iii) *$V \in \mathcal{L}(D_A^{\delta-(1/2)}, D_{A^*}^{(1/2)-\delta})$.*

Then, for all $0 \leq \tau \leq t \leq T$,

$$\begin{aligned}&\langle VX(t), X(t)\rangle - \langle VX(\tau), X(\tau)\rangle\\ &= \int_\tau^t [h(X(s)) + 2\langle U(s), \Psi VX(s)\rangle + \Pi(V)]\,ds\\ &\quad + 2\int_\tau^t \langle \Phi^*VX(s), dW(s)\rangle \quad \text{a.s.},\end{aligned} \tag{3.31}$$

where h is the continuous extension of $2\langle Vx, Ax\rangle$ on H, and, for (i) *and* (ii), $\Pi(V) = \operatorname{Tr} V\Phi\Phi^*$ *and, for* (iii), $\Pi(V) = \operatorname{Tr}(R^*(\beta))^{\delta-(1/2)}V\Phi\Phi^*(R^*(\beta))^{(1/2)-\delta}$.

4. Parameter identification. For the identification of the unknown parameters in the linear stochastic distributed parameter system (2.1), a family of least squares estimates are formed. In this section, it is assumed that $\beta = 0$, that is, $-A(\alpha)$ is strictly positive. Let $\tilde{P}$ be the projection given in (A7). The estimate of the unknown parameter vector at time t, $\hat{\alpha}(t)$ is the minimizer of the quadratic functional of α, $L(t;\alpha)$, given by

$$\begin{aligned}L(t;\alpha) = &-\int_0^t \langle \tilde{P}(A(\alpha)+B(\alpha)K(s))X(s), d\tilde{P}X(s)\rangle\\ &+\frac{1}{2}\int_0^t |\tilde{P}(A(\alpha)+B(\alpha)K(s))X(s)|^2\,ds,\end{aligned} \tag{4.1}$$

where $U(s) = K(s)X(s)$ is an admissible adaptive control.

THEOREM 4.1. *Let $(K(t), t \geq 0)$ be an admissible feedback control law. Assume that* (A2), (A6)–(A8) *are satisfied and $\alpha_0 \in \mathcal{K}^0$. Then the family of least squares estimates $(\hat{\alpha}(t), t > 0)$, where $\hat{\alpha}(t)$ is the minimizer of* (4.1), *is strongly consistent, that is,*

$$P_{\alpha_0}(\lim_{t\to\infty} \hat{\alpha}(t) = \alpha_0) = 1, \tag{4.2}$$

where α_0 is the true parameter vector.

Proof. If B does not depend on $\alpha \in \mathcal{K}$, then the proof of Theorem 4.1 follows from the proof of Theorem 1 in [10]. Therefore we may assume that $B(\alpha)$ has the form (2.12). Since the strong law of large numbers for Brownian motion is used to verify strong consistency, it is shown initially that

$$\lim_{t\to\infty} \int_0^t |\langle \ell, X(s)\rangle|^2 \, ds = +\infty \quad \text{a.s.} \tag{4.3}$$

for suitable $\ell \in H$.

If (P_n) is a sequence of increasing finite-dimensional projections with range in $\cap_{i=0}^p \mathcal{D}(F_i^*)$ that converges strongly to the identity I and $F \in \{F_1, \ldots, F_q\}$, then

$$\begin{aligned} &\lim_{n\to\infty} \int_0^t \langle \tilde{P}F(I + \hat{B}K(s))P_nX(s), \tilde{P}F(I + \hat{B}K(s))P_nX(s)\rangle \, ds \\ &= \int_0^t \langle \tilde{P}F(I + \hat{B}K(s))X(s), \tilde{P}F(I + \hat{B}K(s))X(s)\rangle \, ds \end{aligned} \tag{4.4}$$

in $L^1(P)$ almost surely because the sequence of integrals is monotone increasing.

Fix $n \in \mathbb{N}$. For the process $(P_nX(t), t \geq 0)$ with nonzero values in a finite-dimensional space, the verification of (4.3) is accomplished by using some of the methods in [9] for finite-dimensional systems. Since the expectation of the Lebesgue measure of the amount of time that a scalar Brownian motion is strictly away from zero is infinite, the 0–1 law for Brownian motion implies that, for almost all sample paths, the Lebesgue measure of the amount of time that a sample path is strictly away from zero is infinite.

If $\ell \in H$ and $\Phi^*\ell \neq 0$, then a well-known property of a scalar Brownian motion implies that

$$\liminf_{t\to\infty} \langle \Phi^*\ell, W(t)\rangle = -\infty \quad \text{a.s.} \tag{4.5}$$

and

$$\limsup_{t\to\infty} \langle \Phi^*\ell, W(t)\rangle = +\infty \quad \text{a.s.} \tag{4.6}$$

Fix $n \in \mathbb{N}$ and $\ell \in \cap_{i=0}^q \mathcal{D}(F_i^*)$ such that $\Phi^*\ell \neq 0$ and $P_n\ell \neq 0$. Assume that $P_n\tilde{P} = \tilde{P}$. Let (T_n) be a sequence of stopping times such that $\{\lim_{n\to\infty}\langle \Phi^*\ell, W(T_n)\rangle = +\infty\}$. Let $\Lambda_+ = \{\limsup_{n\to\infty} \int_0^{T_n} \langle M^*(s)\ell, X(s)\rangle \, ds = +\infty\}$, where $M^*(s) = (A^* + K^*(s)\hat{B}^*A^*)P_n$. For each $\omega \in \Lambda_+$, there is a subsequence $(T_{n_j}(\omega))$ such that

$$\lim_{j\to\infty} \int_0^{T_{n_j}(\omega)} \langle M^*(s,\omega)\ell, X(s,\omega)\rangle \, ds = +\infty. \tag{4.7}$$

Since

$$\begin{aligned} \langle \ell, X(t)\rangle^2 \geq &\left(\int_0^t \langle M^*(s)\ell, X(s)\rangle \, ds\right)^2 \\ &+ 2\int_0^t \langle M^*(s)\ell, X(s)\rangle \, ds \, \langle \Phi^*\ell, W(t)\rangle \\ &+ \langle \Phi^*\ell, W(t)\rangle^2, \end{aligned}$$

it follows directly that, for almost all $\omega \in \Lambda_+$,

$$\limsup_{t\to\infty} \langle \ell, X(t,\omega)\rangle^2 = +\infty.$$

Let $\Lambda_- = \{\limsup_{n\to\infty} \int_0^{T_n} \langle M^*(s)\ell, X(s)\rangle\, ds = -\infty\}$. For $\omega \in \Lambda_-$,

$$\lim_{n\to\infty} \int_0^{T_n} \langle M^*(s,\omega)\ell, X(s,\omega)\rangle\, ds = -\infty.$$

Since $-W$ has the same probability law as W, we have that, for almost $\omega \in \Lambda_-$,

$$\limsup_{t\to\infty} \langle \ell, X(t,\omega)\rangle^2 = +\infty.$$

Let $\Lambda_0 = \{\limsup_{n\to\infty} \mid \int_0^{T_n} \langle M^*(s)\ell, X(s)\rangle\, ds \mid < \infty\}$. It follows immediately that, for almost all $\omega \in \Lambda_0$,

$$\limsup_{t\to\infty} \langle \ell, X(t,\omega)\rangle^2 = +\infty.$$

Combining the results of the above three cases, we have

$$\limsup_{t\to\infty} \langle \ell, X(t)\rangle^2 = +\infty \quad \text{a.s.}$$

Since the Lebesgue measure of the amount of time that $(\langle \ell, P_n X(t)\rangle, t \geq 0)$ is strictly away from zero is infinite for almost all sample paths, it follows that

$$\lim_{t\to\infty} \int_0^t |\langle \ell, X(s)\rangle|^2\, ds = +\infty \quad \text{a.s.}$$

By (A7), it follows that there is a $\bar{c} > 0$ such that

$$\text{Tr}[(\hat{P}(I + K^*(s)\hat{B}^*)F^*\tilde{P})^*(\hat{P}(I + K^*(s)\hat{B}^*)F^*\tilde{P})] > c, \tag{4.8}$$

for all $s \in \mathbb{R}_+$, and $F \in \{F_1, \ldots, F_q\}$, which implies that

$$\lim_{t\to\infty} \int_0^t \langle \hat{P}(I + K^*(s)\hat{B}^*)F^*\tilde{P}X(s), \hat{P}(I + K^*(s)\hat{B}^*)F^*\tilde{P}X(s)\rangle\, ds = +\infty \quad \text{a.s.}$$

and, consequently,

$$\lim_{t\to\infty} \int_0^t \langle \tilde{P}F(I + \hat{B}K(s))X(s), \tilde{P}F(I + \hat{B}K(s))X(s)\rangle\, ds = +\infty \quad \text{a.s.}$$

To minimize (4.1) with respect to α, it is necessary and sufficient that $D_\alpha L(t;\alpha) = 0$. Computing the family of partial derivatives and using (2.1), we obtain the family of linear equations

$$\mathcal{A}(t)\hat{\alpha}(t) = \mathcal{A}(t)\alpha_0 + b(t) \tag{4.9}$$

or

$$\tilde{\mathcal{A}}(t)\hat{\alpha}(t) = \tilde{\mathcal{A}}(t)\alpha_0 + \tilde{b}(t), \tag{4.10}$$

where $\mathcal{A}(t)$ and $\tilde{\mathcal{A}}(t)$ are given by (2.13) and (2.14), respectively, and

$$b_j(t) = \int_0^t \langle \tilde{P}(F_j + \hat{B}K(s))X(s), d\tilde{P}\Phi W(s)\rangle,$$

$$\tilde{b}_j(t) = \frac{b_j(t)}{a_{jj}(t)},$$

$$b(t) = (b_1(t), \dots, b_q(t))',$$

$$\tilde{b}(t) = (\tilde{b}_1(t), \dots, \tilde{b}_q(t))'.$$

Let (c_n) be a sequence of positive, real numbers such that $c_n \downarrow 0$. Let $\Lambda_n = \{\liminf_{t\to\infty} |\det \tilde{\mathcal{A}}(t)| > c_n > 0\}$. The sequence (Λ_n) is increasing. By (A8), we have that $P(\Lambda_n) \uparrow 1$ as $n \to \infty$. Given $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that $P(\Lambda_N) > 1 - \varepsilon$. There is a random time such that $|\det \tilde{\mathcal{A}}(t,\omega)| > c_n$ for $\omega \in \Lambda_N$ and $t \geq T(\omega)$. Since $\tilde{b}(t) \to 0$ almost surely as $t \to \infty$ by the strong law of large numbers for Brownian motion, since $\tilde{\mathcal{A}}^*(t)\tilde{\mathcal{A}}(t)$ is uniformly bounded almost surely, and since $\varepsilon > 0$ is arbitrary, it follows that $\hat{\alpha}(t) \to \alpha_0$ almost surely as $t \to \infty$. □

For the applications of identification and adaptive control, it is important to have recursive estimators of the unknown parameters. Let $\langle \tilde{F}(s)x, y\rangle$ be the vector whose ith component is $\langle \tilde{P}F_i(I + \hat{B}K(s))x, y\rangle$. Using (2.1), (4.9), we have

$$\hat{\alpha}(t) = \mathcal{A}^{-1}(t) \int_0^t \langle \tilde{F}(s)X(s), d\tilde{P}X(s) - \tilde{P}F_0X(s)\,ds\rangle. \tag{4.11}$$

Since $\mathcal{A}^{-1}(t)$ satisfies the differential equation

$$d\mathcal{A}^{-1}(t) = -\mathcal{A}^{-1}(t)\,d\mathcal{A}(t)\,\mathcal{A}^{-1}(t),$$

the differential of (4.11) satisfies

$$d\hat{\alpha}(t) = \mathcal{A}^{-1}(t)\langle \tilde{F}(t)X(t), d\tilde{P}X(s) - \tilde{P}A(\hat{\alpha}(t))(I + \hat{B}K(t))X(t)\,dt\rangle. \tag{4.12}$$

5. Optimality for an adaptive control. In this section, the certainty equivalence, optimal ergodic control law is shown to be self-tuning and self-optimizing. The self-tuning property is obtained by using the continuity of the solution of a stationary Riccati equation with respect to parameters in the topology induced by a suitable operator norm. Since the unbounded operator $B(\alpha)$ appears in the linear transformation of the control in (2.1), this operator topology is more restrictive than for bounded linear transformations on the Hilbert space. This continuity property is also used to show that the certainty equivalence control stabilizes the unknown system in a suitable sense. The self-optimizing property is verified for this adaptive control.

The solution V of the stationary Riccati equation (2.9) satisfies the assumptions of Corollary 3.5 if one of the following three conditions is satisfied: (i) Φ is Hilbert–Schmidt, (ii) V is nuclear, or (iii) A is strictly negative. By (A5), $V \in \mathcal{L}(H, D_{A^*}^{1-\varepsilon})$ (see [12], [18]), and (2.10) implies that

$$|\langle Ax, Vx\rangle| = |\langle Rx, x\rangle| \leq k|x|^2$$

for some $R \in \mathcal{L}(H)$. If A is strictly negative, then it easily follows that

$$V \in \mathcal{L}\big(D_A^{\delta-(1/2)}, D_{A^*}^{(1/2)-\delta}\big).$$

Moreover, if (A2) is satisfied with $\Phi = I$, then V is nuclear because, from Theorems 1 and 2 of [12], it follows that $P_a = (-A^* + \beta I)^a V \in \mathcal{L}(H)$ for each $a \in (0,1)$. Thus

$V = P_a^*(-A + \beta I)^{-a}$ is nuclear because $(-A + \beta I)^{-a}$ is nuclear for $a = 1 - 2\delta$ by (A2), (A5).

If an adaptive control is self-tuning and some stability properties are satisfied for the solution of (2.1), then this adaptive control is self-optimizing.

PROPOSITION 5.1. *Assume that* (A2), (A5) *are satisfied, that the solution* V *of* (2.10) *satisfies the assumptions of Corollary* 3.5, *and that*

$$\lim_{t\to\infty} \frac{1}{t}\langle VX(t), X(t)\rangle = 0 \quad \text{a.s.}, \tag{5.1}$$

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t |X(s)|^2\, ds < \infty \quad \text{a.s.}, \tag{5.2}$$

where $(X(t), t \geq 0)$ *is the solution of* (2.1) *with* $\alpha_0 \in \mathcal{K}$ *and the control* $U \in \cap_{T>0} M_W^p(0, T, H_1)$. *Then*

$$\liminf_{T\to\infty} \frac{1}{T} J(X_0, U, \alpha_0, T) \geq \Pi(V) \quad \text{a.s.}, \tag{5.3}$$

where V *is the solution of* (2.10) *with* $\alpha = \alpha_0$. *Furthermore, if* U *is an admissible control* $U(t) = K(t)X(t)$ *such that*

$$\lim_{t\to\infty} K(t) = k_0 \quad \text{a.s.} \tag{5.4}$$

in the uniform $\mathcal{L}(H, H_1)$ *topology where* $k_0 = -P^{-1}\Psi V$, *then*

$$\lim_{T\to\infty} \frac{1}{T} J(X_0, U, \alpha_0, T) = \Pi(V) \quad \text{a.s.} \tag{5.5}$$

Proof. For $U \in M_W^p(0, T, H_1)$, we have

$$\begin{aligned}\langle VX(t), X(t)\rangle - \langle Vx, x\rangle = &\int_0^t h(X(s)) + 2\langle U(s), \Psi VX(s)\rangle\, ds \\ &+ t\Pi(V) + 2\int_0^t \langle \Phi^* VX(s), dW(s)\rangle \quad \text{a.s.},\end{aligned} \tag{5.6}$$

where $h(x)$ is the continuous extension of $2\langle Ax, Vx\rangle, x \in \mathcal{D}(A)$. Using the stationary Riccati equation (2.10), we obtain

$$\begin{aligned}&\langle VX(t), X(t)\rangle - \langle VX_0, X_0\rangle \\ &= \int_0^t [2\langle U(s), \Psi VX(s)\rangle + \langle P^{-1}\Psi VX(s), \Psi VX(s)\rangle - \langle QX(s), X(s)\rangle]\, ds \\ &\quad + 2\int_0^t \langle \Phi^* VX(s), dW(s)\rangle + t\Pi(V) \quad \text{a.s.}\end{aligned}$$

By a similar method as in Proposition 2 of [10], we obtain from (5.2) that

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \langle \Phi^* VX(s), dW(s)\rangle = 0 \quad \text{a.s.} \tag{5.7}$$

Thus

$$\lim_{T\to\infty}\left(\frac{1}{T} J(X_0, U, \alpha_0, T) - \frac{1}{T}\int_0^T |P^{1/2}U(s) + P^{-1}\Psi VX(s)|^2\, ds\right) = \Pi(V) \quad \text{a.s.},$$

and (5.3) is verified. If $U(t) = K(t)X(t)$ and if (5.4) is satisfied, then

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T |P^{1/2}(U(s) + P^{-1}\Psi VX(s))|^2 \, ds = 0 \quad \text{a.s.}$$

by (5.3). Thus (5.5) is verified. □

Now it is shown that the stability conditions (5.1), (5.2) are satisfied for an admissible, self-tuning adaptive control.

PROPOSITION 5.2. *Assume that* (A2), (A5) *are satisfied. Let the solution* V *of* (2.10) *satisfy the assumptions of Corollary* 3.5. *If* $(X(t), t \geq 0)$ *is the solution of* (2.1) *with* $\alpha_0 \in \mathcal{K}$ *and an adaptive control law* $(K(t), t \geq 0)$ *that satisfies* (5.4), *then* (5.1), (5.2) *are satisfied.*

Proof. Apply the Itô formula (3.31) of Corollary 3.5 to $\langle VX(t), X(t)\rangle$ again to obtain (5.6). Let $(P(t), t \geq 0)$ satisfy

$$P(t) = \int_0^t \Big[h(X(s)) + 2\langle K(s)X(s), \Psi VX(s)\rangle + \frac{1}{2}\langle QX(s), X(s)\rangle \Big] \, ds + 2\int_0^t \langle \Phi^* VX(s), dW(s)\rangle$$

and use the stationary Riccati equation (2.10) to obtain

$$P(t) = \int_0^t \Big[2\langle K(s)X(s), \Psi VX(s)\rangle + 2\langle P^{-1}\Psi VX(s), \Psi VX(s)\rangle - \langle P^{-1}\Psi VX(s), \Psi VX(s)\rangle - \frac{1}{2}\langle QX(s), X(s)\rangle \Big] \, ds + 2\int_0^t \langle \Phi^* VX(s), dW(s)\rangle.$$

By (2.7), (2.8) and the boundedness of $\Phi^* V$, there are constants c_0, c_1, c_2, and c_3 such that

$$\begin{aligned} P(t) &\leq \int_0^t [2\langle (P^{-1}\Psi V + K(s))X(s), \Psi VX(s)\rangle - c_0|X(s)|^2] \, ds + 2\int_0^t \langle \Phi^* VX(s), dW(s)\rangle \\ &\leq \int_0^t (-c_1 + c_2|K(s) - k_0|_{\mathcal{L}(H,H_1)})|X(s)|^2 \, ds \\ &\quad + \int_0^t |\Phi^* VX(s)|^2 \, ds \left[-c_3 + \frac{2}{\int_0^t |\Phi^* VX(s)|^2 \, ds} \int_0^t \langle \Phi^* VX(s), dW(s)\rangle \right] \\ &= P_1(t) + P_2(t). \end{aligned}$$

Since $K(s) \to k_0$ almost surely as $s \to \infty$, $\limsup_{t\to\infty}(1/t)P_1(t) \leq 0$ almost surely, and by the strong law of large numbers for Brownian motion, we have $\limsup_{t\to\infty}(1/t)P_2(t) \leq 0$ almost surely. Thus

$$\Pi(V) \geq \limsup_{t\to\infty} \left[\frac{1}{t}\langle VX(t), X(t)\rangle + \frac{1}{2t}\int_0^t \langle QX(s), X(s)\rangle \, ds \right] \quad \text{a.s.,}$$

and (5.2) is verified.

To verify (5.1), again use the Itô formula (Corollary 3.5) for $\langle VX(t), X(t)\rangle$ as

$$\begin{aligned}
&\langle VX(t), X(t)\rangle - \langle VX(\tau), X(\tau)\rangle \\
&= \int_\tau^t [2\langle K(s)X(s), \Psi VX(s)\rangle - 2\langle k_0 X(s), \Psi VX(s)\rangle \\
&\quad - \langle QX(s), X(s)\rangle + \Pi(V)]\, ds + 2\int_\tau^t \langle \Phi^* VX(s), dW(s)\rangle \\
&\le \int_\tau^t (-c_1 + c_2|K(s) - k_0|)|X(s)|^2\, ds + (t-\tau)\Pi(V) \\
&\quad + 2\int_\tau^t \langle \Phi^* VX(s), dW(s)\rangle,
\end{aligned}$$

where $c_1 > 0, c_2 > 0$ and $\tau \in [0, t]$. Thus

$$\begin{aligned}
&\langle VX(t), X(t)\rangle - \langle VX(\tau), X(\tau)\rangle \\
&\le \int_\tau^t - c_3\langle VX(s), X(s)\rangle\, ds + c_4(t-\tau) + M(\tau, t)
\end{aligned} \tag{5.8}$$

for some $c_3 > 0, c_4 > 0$, and $t \ge \tau > T_0$, where T_0 is a random time and where

$$M(\tau, t) = 2\int_\tau^t \langle \Phi^* VX(s), dW(s)\rangle.$$

Let $\psi(t) = \langle VX(t), X(t)\rangle$ and let $(y(t), t \ge T_0)$ satisfy

$$y(t) = \psi(T_0) - c_3\int_{T_0}^t y(s)\, ds + c_4(t - T_0) + M(T_0, t) \tag{5.9}$$

for $t \ge T_0$. Taking the difference of (5.8) and (5.9), it is clear that $\psi(t) \le y(t)$ almost surely for $t \ge T_0$. Solving the integral equation (5.9), we have

$$\begin{aligned}
y(t) &= \psi(T_0)e^{-c_3(t-T_0)} + \frac{c_4}{c_3}\left(1 - e^{-c_3(t-T_0)}\right) \\
&\quad - c_3\int_{T_0}^t e^{-c_3(t-s)} M(T_0, s)\, ds + M(T_0, t) \quad \text{a.s.}
\end{aligned}$$

From (5.2), which has been verified above, and from the strong law of large numbers for Brownian motion, we have that

$$\lim_{t\to\infty} \frac{1}{t}\, y(t) = 0 \quad \text{a.s.},$$

which verifies (5.1). □

To verify the self-tuning property for the certainty equivalence adaptive control, $K(t) = -P^{-1}\Psi(\hat\alpha(t-\Delta))V(\hat\alpha(t-\Delta))$, where $(\hat\alpha(t), t \ge 0)$ is a family of strongly consistent estimators of the true parameter vector α_0. It is important to show a suitable continuous dependence of the solution $V(\alpha)$ of the stationary Riccati equation on $\alpha \in \mathcal{K}$. For B bounded, some results are given in [5], [10]. For B unbounded, as in (2.1), we can use a continuity result from [17, Thms. 1.1 and 5.3], which is reformulated below. It is assumed that $A(\alpha)$ is strictly negative for each $\alpha \in \mathcal{K}$. For notational convenience, let $A_0 = A(\alpha_0)$, where $\alpha_0 = \mathcal{K}$ is the true parameter value.

LEMMA 5.3. *Assume that* (A1), (A3), (A4) *are satisfied and that*

$$\lim_{\alpha\to\alpha_0} |\Psi(\alpha)-\Psi(\alpha_0)|_{\mathcal{L}(D^{1-\varepsilon}_{A_0^*},H_1)}=0, \tag{5.10}$$

$$\lim_{\alpha\to\alpha_0} |S(t;\alpha)-S(t;\alpha_0)|_{\mathcal{L}(D^{\varepsilon-1}_{A_0},H)}=0 \tag{5.11}$$

for each $t\geq 0$, *where* $\Psi(\alpha)=B^*(\alpha)$. *Then*

$$\lim_{\alpha\to\alpha_0} |V(\alpha)-V(\alpha_0)|_{\mathcal{L}(H,D^{1-\varepsilon}_{A_0^*})}=0. \tag{5.12}$$

Note that (A3) and (5.11) imply that

$$\lim_{\alpha\to\alpha_0} |A^{-1}(\alpha)-A^{-1}(\alpha_0)|_{\mathcal{L}(H)}=0,$$

and, from (A1), (5.10) and (5.11), we have that

$$\lim_{\alpha\to\alpha_0} |\Psi(\alpha)S^*(t;\alpha)-\Psi(\alpha_0)S^*(t;\alpha_0)|_{\mathcal{L}(H,H_1)}=0.$$

Thus we can follow the proof of Theorem 5.3 in [17] to obtain (5.12).

The self-optimizing property is now verified for a self-tuning adaptive control.

THEOREM 5.4. *Assume that* (A1)–(A4), (A6)–(A8) *are satisfied. Let* $(\hat{\alpha}(t),t\geq 0)$ *be the family of least squares estimates, where* $\hat{\alpha}(t)$ *is the minimizer of* (4.1). *Let* $(K(t),t\geq 0)$ *be an admissible adaptive control law such that*

$$K(t)=-P^{-1}\Psi(\hat{\alpha}(t-\Delta))V(\hat{\alpha}(t-\Delta)), \tag{5.13}$$

where $\Psi(\alpha)=B^*(\alpha)$ *and* $V(\alpha)$ *is the solution of* (2.10) *for* $\alpha\in\mathcal{K}$. *Then the family of estimates* $(\hat{\alpha}(t),t\geq 0)$ *is strongly consistent,*

$$\lim_{t\to\infty} K(t)=k_0 \quad \text{a.s.} \tag{5.14}$$

in $\mathcal{L}(H,H_1)$ *where* $k_0=-P^{-1}\Psi(\alpha_0)V(\alpha_0)$, *and*

$$\lim_{T\to\infty}\frac{1}{T}J(X_0,U,\alpha_0,T)=\operatorname{Tr}\Pi(V(\alpha_0)) \quad \text{a.s.}, \tag{5.15}$$

where $U(t)=K(t)X(t)$ *and* $\Pi(V)$ *is given in Corollary* 3.5.

The proof follows directly from Theorem 4.1, Lemma 5.3, and Propositions 5.1 and 5.2 with $A=A(\alpha_0)$. The solution $V=V(\alpha_0)$ of the Riccati equation satisfies the assumptions of Corollary 3.5 because $A(\alpha_0)$ is strictly negative.

6. Some examples.

Example 6.1. This is a family of examples from elliptic differential operators. Let G be a bounded, open domain in $\mathbb{R}^n$ with C^∞-boundary ∂G with G locally on one side of ∂G and let $L(x,D)$ be an elliptic differential operator of the form

$$L(x,D)f=\sum_{i,j=1}^{n} D_i a_{ij}(x)D_j f+\sum_{i=1}^{n}[b_i(x)D_i f+D_i(d_i(x)f)]+c(x)f, \tag{6.1}$$

where the coefficients a_{ij},b_i,d_i,c are elements of $C^\infty(\bar{G})$,

$$\sum_{i,j} a_{ij}(x)\xi_i\xi_j\geq\hat{\nu}|\xi|^2, \tag{6.2}$$

where $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n, x \in G, \hat{\nu} > 0$ is a constant, and $\{a_{ij}(x)\}$ is symmetric. Consider a stochastic parabolic control problem formally described by the equations

$$\frac{\partial y}{\partial t}(t,x) = L(x,D)y(t,x) + \eta(t,x) \tag{6.3}$$

for $(t,x) \in \mathbb{R}_+ \times G$ and

$$\frac{\partial y}{\partial \nu}(t,x) + h(x)y(t,x) = u(t,x) \tag{6.4}$$

for $(t,x) \in \mathbb{R}_+ \times \partial G$ and $y(0,x) = y_0(x)$, where $\partial/\partial\nu = \sum_{i,j=1}^n a_{ij}\nu_j D_i$ is the normal derivative, $\nu = (\nu_1, \ldots, \nu_n)$ is the outward normal to ∂G, the process $(\eta(t,x); (t,x) \in \mathbb{R}_+ \times G)$ formally denotes a space dependent white noise, $u \in L^2(0,T,L^2(\partial G))$ for any $T > 0, h \in C^\infty(\partial G)$, and $h \geq 0$.

To give a precise meaning to (6.3), (6.4), the semigroup approach is used. An intuitive justification of the semigroup model (2.1) is given. Let $H = L^2(G), H_1 = L^2(\partial G)$ and define $Af = L(x,D)f$. $A : H \to H$ is densely defined, and $\mathcal{D}(A) = \{f \in H^2(G) : \partial f/\partial\nu + hf = 0 \text{ on } \partial G\}$. It is well known that A generates an analytic semigroup, the linear operator $(A - \beta I)$ is strictly negative for some $\beta \geq 0$.

To introduce the control operator, consider the elliptic problem

$$(L(x,D) - \beta)z = 0 \quad \text{on } G, \tag{6.5}$$

$$\frac{\partial z}{\partial \nu} + hz = -g \quad \text{on } \partial G. \tag{6.6}$$

For $g \in L^2(\partial G)$, there is a unique solution $z \in H^{3/2}(G)$ [19]. Define $\hat{B} \in \mathcal{L}(H_1, H^{3/2}(G))$ by the equation $\hat{B}g = -z$. For $\varepsilon < 3/4$, we have $\hat{B} \in \mathcal{L}(H_1, D_A^\varepsilon)$ because $D_A^{(3/4)-\gamma} = H^{(3/2)-2\gamma}$ for any sufficiently small $\gamma > 0$ [13]. Let $y_\beta(t,x) = e^{-\beta t}y(t,x)$ and $\eta(t,x)\,dt = \Phi\,dW(t)$ for some $\Phi \in \mathcal{L}(H)$ and a cylindrical Wiener process $(W(t), t \geq 0)$ in H. From (6.5), (6.6), we have

$$dy_\beta = (L(x,D) - \beta)y_\beta\,dt + e^{-\beta t}\Phi\,dW(t), \tag{6.7}$$

$$\begin{aligned} \frac{\partial y_\beta}{\partial \nu} + hy_\beta &= e^{-\beta t}u = u_\beta(t) \quad \text{on } \partial G, \\ y_\beta(0) &= y(0). \end{aligned} \tag{6.8}$$

Formally performing the differentiation $(\partial/\partial t)\hat{B}u_\beta(t)$, we obtain

$$d\omega_\beta(t) = ((L(x,D) - \beta)y_\beta(t) - \hat{B}\dot{u}_\beta(t))\,dt + e^{-\beta t}\Phi\,dW(t),$$

$$\frac{\partial \omega_\beta}{\partial \nu} + h\omega_\beta = 0 \quad \text{on } \mathbb{R}_+ \times \partial G,$$

where $\omega_\beta(t) = y_\beta(t) - \hat{B}u_\beta(t)$. For (6.7), the formula for the mild solution is

$$\begin{aligned} \omega_\beta(t) = S_\beta(t)(y(0) + \hat{B}u(0)) &+ \int_0^t S_\beta(t-r)\Phi e^{\beta r}\,dW(r) \\ &- \int_0^t S_\beta(t-r)\hat{B}\dot{u}_\beta(r)\,dr, \end{aligned} \tag{6.9}$$

where $S_\beta(t) = e^{t(A-\beta I)}$. Formally integrating by parts the last integral in (6.9) yields

$$y_\beta(t) = \omega_\beta(t) - \hat{B}u_\beta(t) = S_\beta(t)y(0) + \int_0^t (A-\beta I)S_\beta(t-r)\hat{B}u_\beta(r)\,dr + \int_0^t e^{-\beta r}S_\beta(t-r)\Phi\,dW(r).$$

Thus, cancelling $e^{-\beta t}$, we have

$$y(t) = S(t)y(0) + \int_0^t S(t-r)Bu(r)\,dr + \int_0^t S(t-r)\Phi\,dW(r),$$

which is a mild solution of the form (2.3), where $B = \Psi^*$ and $\Psi \in \mathcal{L}(D_{A^*}^{1-\varepsilon}, H_1)$ extends the operator $\hat{B}^*(A^* - \beta I)$.

The assumptions that are used in this paper are now verified for this example. Assumption (A2) may not be satisfied so that it can be considered as a condition on the noise term (specifically on Φ). If Φ is Hilbert–Schmidt or if $\Phi W(t)$ evolves in H, then (A2) is satisfied. If $n = 1$ and $\Phi \in \mathcal{L}(H)$, then (A2) is satisfied. In this case,

$$Af = \frac{\partial}{\partial x}a(x)\frac{\partial}{\partial x}f + \left(b(x)\frac{\partial}{\partial x}f + \frac{\partial}{\partial x}d(x)f + c(x)f\right).$$

Let $A_1 = (\partial/\partial x)a(x)(\partial/\partial x)$. By Corollary 2.6.11 of [20],

$$|(-A_1^* + \beta I)^\gamma x|^2 \leq \text{ const}|(-A^* + \beta I)^\gamma x|^2$$

for $x \in \mathcal{D}((\beta I - A^*)^\gamma)$, where $\gamma \in \left(0, \frac{1}{2}\right)$. It follows that $(-A+\beta I)^{-\gamma}$ is Hilbert–Schmidt if $(-A_1 + \beta I)^{-\gamma}$ is Hilbert–Schmidt. Since $A_1 = A_1^*$ and $(A_1 - \beta I)^{-1}$ is compact, we can use the comments following Lemma 2.1 to conclude that $(-A_1 + \beta I)^{-(1/2)+\delta}$ is Hilbert–Schmidt if and only if

$$\int_0^T t^{-2\delta}|\tilde{S}(t)|_{HS}^2\,dt < \infty \tag{6.10}$$

for $T > 0$, where $(\tilde{S}(t), t \geq 0)$ is the semigroup generated by $A_1 - \beta I$. We have

$$|\tilde{S}(t)|_{HS}^2 = \int_G \int_G |G(t,\theta,r)|^2\,dr\,d\theta,$$

where $G(t,\theta,r)$ is the Green function for the problem

$$\frac{\partial \omega}{\partial t} = [a(x)\omega']' - \beta\omega, \qquad \frac{\partial \omega}{\partial \nu} + h\omega = 0,$$

since

$$|G(t,\theta,r)| \leq \frac{k_1}{\sqrt{t}}\exp\left[k_2\frac{|\theta - r|^2}{t}\right]$$

for $t > 0$ and $\theta, r \in G$, where k_1 and k_2 are positive constants [1], [2]. Condition (6.10) is satisfied for any $\delta \in \left(0, \frac{1}{4}\right)$. Thus (A2) is satisfied for any $\Phi \in \mathcal{L}(H)$. Assumption (A5) can be shown to be satisfied. For example, this is trivially satisfied if the operator A is strictly negative. In the above example, if $A(= A(\alpha_0))$ is strictly negative and (A2) is satisfied, then,

for the control system (2.1) with $\alpha = \alpha_0$ and the cost functional (2.6) where $Q \in \mathcal{L}(L^2(G))$ and $P \in \mathcal{L}(L^2(\partial G))$ are uniformly positive, the self-optimizing property (5.5) of Proposition 5.1 is satisfied.

Now consider a parameter dependent version of (6.3), (6.4)

$$\frac{\partial y}{\partial t}(t,x) = \alpha L(x,D)y(t;x) + \eta(t,x), \qquad (t,x) \in \mathbb{R}_+ \times G, \tag{6.11}$$

$$\begin{aligned} \frac{\partial y}{\partial \nu} &= u(t,x), \qquad (t,x) \in \mathbb{R}_+ \times \partial G, \\ y(0,x) &= y_0(x), \end{aligned} \tag{6.12}$$

where $\alpha \in \mathcal{K} = [\alpha_1, \alpha_2]$ is scalar parameter for $0 < \alpha_1 < \alpha_2$. Assume that the operator A corresponding to $L(x,D)$ is strictly negative and that (A2) is satisfied. Using the same semigroup model as above, we have that

$$\begin{aligned} y(t;\alpha) = S(t;\alpha)y_0 + \alpha \int_0^t S(t-r;\alpha)BU(r)\,dr \\ + \int_0^t S(t-r;\alpha)\Phi\,dW(r), \end{aligned} \tag{6.13}$$

where $S(t;\alpha) = e^{t\alpha A}$, $B = [\hat{B}^* A^*]^* \in \mathcal{L}(D_{A^*}^{1-\varepsilon}, H_1)$ and $\hat{B} \in \mathcal{L}(H_1, D_A^{\varepsilon})$ solves the elliptic problem

$$\begin{aligned} L(x,D)(\hat{B}g) &= 0 \quad \text{on } G, \\ \frac{\partial}{\partial \nu}(\hat{B}g) &= -g \quad \text{on } \partial G. \end{aligned}$$

Assumptions (A1), (A3)–(A6), and (5.10) are now trivially satisfied because $A(\alpha) = \alpha A, \alpha \in [\alpha_1, \alpha_2], \alpha_1 > 0$, and A is strictly negative. Condition (5.11) is satisfied because $S(\cdot) \in C((0,\infty), D_A^{1-\varepsilon})$. Furthermore, we have that $(A^*)^{-1} \in \mathcal{L}(H), \hat{B} \in \mathcal{L}(H_1, D_A^{\varepsilon})$ and that the embedding $D_A^{\varepsilon} \to H$ is compact, so (A7) is satisfied. Since the parameter is scalar, (A8) is trivially satisfied. Thus, by Theorem 4.1, the family of least squares estimates given in the statement there is strongly consistent for $\alpha_0 \in (\alpha_1, \alpha_2)$. For any strongly consistent family of estimators $(\hat{\alpha}(t), t \geq 0)$, the cost functional (2.6) with a uniformly positive $Q \in \mathcal{L}(L^2(G))$ and $P \in \mathcal{L}(L^2(\partial G))$, system (2.1) with $A(\alpha_0) = \alpha_0 A, B$ as above, $\beta = 0$, the adaptive control

$$U(t) = -P^{-1}\Psi(\hat{\alpha}(t-\Delta))V(\hat{\alpha}(t-\Delta))X(t)$$

has the self-optimizing property (5.15) by Theorem 5.4.

An elementary example of a boundary control problem with a vector parameter α is described that satisfies (A8). It is a specialization of (6.5). Let $H = L^2([0,1], \mathbb{R})$, let F_1 and F_2 be the linear operators

$$F_1 = \frac{d^2}{dx^2}, \qquad F_2 = \frac{d}{dx},$$

and let $A(\alpha)$ be

$$A(\alpha) = \alpha_1 F_1 + \alpha_2 F_2,$$

where $\alpha = (\alpha_1, \alpha_2), \alpha_i \in [\bar{a}_i, \bar{b}_i]$ and $\bar{a}_i > 0, \bar{b}_i < \infty$ for $i = 1, 2$. The domain of $A(\alpha)$ is $\mathcal{D}(A(\alpha)) = \{f \in H^2(0,1): \partial f/\partial \nu = 0 \text{ on } \{0,1\}\}$. Let $(k_n, \ell_n; n = 0, 1, \ldots)$ be the basis of H, defined as

$$k_n(x) = \sqrt{2} \sin 2n\pi x,$$

$$\ell_n(x) = \sqrt{2} \cos 2n\pi x.$$

Fix a positive integer N and let $\tilde{P}$ be the projection determined by the family $(k_n, \ell_n, n = 1, 2, \ldots, N)$. Since the adaptive control law $(BK(t), t \geq 0)$ is a family of compact operators, it is the limit of a family of finite rank operators. Thus, to evaluate $\mathcal{A}(t)$, it suffices to apply F_1 and F_2 to the finite sum $\sum_{j=1}^{\tilde{N}} (a_j k_j + b_j \ell_j)$. It is elementary to verify that

$$\langle F_1 k_n, F_2 k_m \rangle = \langle F_1 \ell_n, F_2 \ell_m \rangle = 0$$

for all m and n and that

$$\langle F_1 k_n, F_2 \ell_m \rangle = \langle F_1 \ell_n, F_2 k_m \rangle = 0$$

for $n \neq m$. Thus

$$\left\langle F_1 \sum_{j=1}^{\tilde{N}} (a_j k_j + b_j \ell_j), F_2 \sum_{j=1}^{\tilde{N}} (a_j k_j + b_j \ell_j) \right\rangle$$
$$= \sum_{j=1}^{\tilde{N}} a_j b_j (\langle F_1 k_j, F_2 \ell_j \rangle + \langle F_1 \ell_j, F_2 k_j \rangle) \quad \text{a.s.}$$

Since we have

$$\langle F_1 k_j, F_2 \ell_j \rangle = (2j\pi)^3, \qquad \langle F_1 \ell_j, F_2 k_j \rangle = -(2j\pi)^3,$$

by passage to the limit and integration, it follows that

$$a_{12}(t) = 0 \quad \text{a.s.}$$

for $t > 0$, and the matrix $\mathcal{A}(t)$ is diagonal. Thus, for $t > 0, \tilde{\mathcal{A}}(t) = I$ almost surely and $\det \tilde{\mathcal{A}}(t) = 1$ almost surely; so (A8) is trivially satisfied. This example can be generalized to many space dimensions. For example, consider the dimension-2 case. Let $H = L^2([0,1] \times [0,1], \mathbb{R})$ and let F_1, F_2, and F_3 be the linear operators

$$F_1 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \quad F_2 = \frac{\partial}{\partial x}, \quad F_3 = \frac{\partial}{\partial y}.$$

Let $A(\alpha)$ be

$$A(\alpha) = \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3,$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3), \alpha_i \in [\tilde{a}_i, \tilde{b}_i], \tilde{a}_i > 0$, and $\tilde{b}_i < \infty$ for $i = 1, 2, 3$. It easily follows by computations that are similar to the above that the matrix $\mathcal{A}(t)$ is diagonal, so that, for $t > 0, \tilde{\mathcal{A}}(t) = I$ almost surely and $\det \tilde{\mathcal{A}}(t) = 1$ almost surely. Thus (A8) is satisfied.

Example 6.2. This example is a structurally damped plate with random loading and point control. Consider the following model of a plate in the deflection w:

$$w_{tt}(t,x) + \Delta^2 w(t,x) - \alpha \Delta w(t,x) = \delta(x - x_0)u(t) + \eta(t,x) \tag{6.14}$$

for $(t,x) \in \mathbb{R}_+ \times G$,

$$w(0,\cdot) = w_0, \qquad w_t(0,\cdot) = w_1, \tag{6.15}$$

$$w|_{\mathbb{R}_+ \times \partial G} = \Delta w|_{\mathbb{R}_+ \times \partial G} = 0, \tag{6.16}$$

where $\alpha > 0$ is an unknown constant, $\eta(t,x)$ formally represents a space-dependent Gaussian white noise on the open, bounded, smooth domain $G \subset \mathbb{R}^n$ for $n \leq 3$, and $\delta(x - x_0)$ is the Dirac distribution at $x_0 \in G$. The cost functional is

$$J(w_0, w_1, u, \alpha, T) = \int_0^T (|w(t)|^2_{H^2(G)} + |w_t(t)|^2_{L^2(G)} + |u(t)|^2)\, dt. \tag{6.17}$$

For a mathematical treatment of the deterministic problem (6.14)–(6.17) where $\eta \equiv 0$, refer to [3], [4], [17] and references therein. Define the linear operator $\mathcal{A}$ by the equation $\mathcal{A}h = \Delta^2 h$, where $\mathcal{D}(\mathcal{A}) = \{h \in H^4(G) : h|_{\partial G} = \Delta h|_{\partial G} = 0\}$. Following [4], [17], (6.14)–(6.17) are rewritten in the form (2.1), (2.6), where $H = \mathcal{D}(\mathcal{A}^{1/2}) \times L^2(G) = (H^2(G) \cap H_0^1(G)) \times L^2(G), H_1 = \mathbb{R}$,

$$A(\alpha) = \begin{bmatrix} 0 & I \\ -\mathcal{A} & -\alpha \mathcal{A}^{1/2} \end{bmatrix},$$

$$Bu = \begin{bmatrix} 0 \\ \delta(x - x_0)u \end{bmatrix},$$

$$\Phi = \begin{bmatrix} 0 & 0 \\ 0 & \Phi_1 \end{bmatrix},$$

where $\Phi_1 \in \mathcal{L}(L^2(G))$ is a Hilbert–Schmidt operator and where $\Phi_1 \Phi_1^* > 0, Q = I, P = I$, and $(W(t), t \geq 0)$ in (2.1) is a cylindrical Wiener process on H. It is known [4] that $A(\alpha)$ generates a stable analytic semigroup, $(S(t;\alpha), t \geq 0)$, and that $B \in \mathcal{L}(H_1, D^{\varepsilon-1}_{A(\alpha)})$ for $\varepsilon \in (0, 1 - n/4)$, which is possible for $n \leq 3$ (cf. [17]). Suppose that the unknown parameter $\alpha \in \mathcal{K} = [a_0, a_1]$, where $0 < a_0 < a_1$. Assumptions (A1), (A2), (A4)–(A6) are clearly satisfied. Since B does not depend on $\alpha \in \mathcal{K}$, assumption (A7) is satisfied with a finite-dimensional projection $\tilde{P} : H \to \tilde{P}(H)$ of the form

$$\tilde{P} = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{P}_1 \end{bmatrix},$$

where $\tilde{P}_1 : L^2(G) \to H^2(G)$ and $\tilde{P}_1 \neq 0$. Assumption (A8) is trivially satisfied because the parameter α is scalar. The assumptions of the uniform analyticity and the exponential stability of the semigroup $(S(t;\alpha), t \geq 0)$ and the continuous dependence of this semigroup on α, (5.11), can be verified by the explicit spectral expansions of $A(\alpha)$ and $(S(t;\alpha), t \geq 0)$ [4, Thm. A3]. Therefore, by Theorem 4.1, the family of least squares estimates given in the statement there is strongly consistent for $\alpha_0 \in (a_0, a_1)$. For any strongly consistent family of estimators $(\hat{\alpha}(t), t \geq 0)$, system (2.1) with $A(\alpha_0), B, \Phi$ as above, $\beta = 0$, and the adaptive control $(U(t), t \geq 0)$ given by

$$U(t) = -\Psi(\hat{\alpha}(t - \Delta))V(\hat{\alpha}(t - \Delta))X(t)$$

has the self-optimizing property (5.15) by Theorem 5.4.

Acknowledgments. We thank A. Bensoussan for suggesting the study of adaptive boundary control problems, and we dedicate this paper to him on his fiftieth birthday, when this work was initiated, following his suggestion. We also thank I. Vrkoc and J. Seidler for reading the manuscript and suggesting improvements, and I. Lasiecka for providing important information on deterministic, distributed parameter systems that broadened the scope of this paper.

REFERENCES

[1] H. AMANN, *On abstract parabolic fundamental solutions*, J. Math. Soc. Japan, 39 (1987), pp. 93–116.
[2] R. ARIMA, *On general boundary value problem for parabolic equations*, J. Math. Kyoto Univ., 4 (1964), pp. 207–243.
[3] G. CHEN AND D. L. RUSSELL, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 39 (1982), pp. 433–454.
[4] S. CHEN AND R. TRIGGIANI, *Proof of extensions of two conjectures on structural damping for elastic systems*, Pacific J. Math., 136 (1989), pp. 15–55.
[5] A. CHOJNOWSKA-MICHALIK, T. E. DUNCAN, AND B. PASIK-DUNCAN, *Uniform operator continuity of the stationary Riccati equation in Hilbert space*, Appl. Math. Optim., 25 (1992), pp. 171–187.
[6] R. CURTAIN AND P. FALB, *Itô's lemma in infinite dimensions*, J. Math. Anal. Appl., 31 (1970), pp. 434–448.
[7] G. DAPRATO, S. KWAPIEN, AND J. ZABCZYK, *Regularity of solutions of linear stochastic equations in Hilbert spaces*, Stochastics, 23 (1987), pp. 1–23.
[8] G. DAPRATO AND A. ICHIKAWA, *Riccati equations with unbounded coefficients*, Ann. Mat. Pura Appl., 140 (1985), pp. 209–221.
[9] T. E. DUNCAN AND B. PASIK-DUNCAN, *Adaptive control of continuous time linear stochastic systems*, Math. Control Signals Systems, 3 (1990), pp. 45–60.
[10] T. E. DUNCAN, B. GOLDYS, AND B. PASIK-DUNCAN, *Adaptive control of linear stochastic evolution systems*, Stochastics and Stochastic Reports, 35 (1991), pp. 129–142.
[11] F. FLANDOLI, *Direct solution of a Riccati equation arising in a stochastic control problem with control and observations on the boundary*, J. Appl. Math. Optim., 14 (1986), pp. 107–129.
[12] ———, *Algebraic Riccati equation arising in boundary control problems*, SIAM J. Control Optim., 25 (1987), pp. 612–636.
[13] D. FUJIWARA, *Concrete characterizations of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad. Ser. A. Math. Sci., 43 (1967), pp. 82–86.
[14] A. ICHIKAWA, *Stability of semilinear stochastic evolution equations*, J. Math. Anal. Appl., 190 (1982), pp. 12–44.
[15] ———, *Semilinear stochastic evolution equations: Boundedness, stability and invariant measures*, Stochastics, 12 (1984), pp. 1–39.
[16] P. KOTELENEZ, *A maximal inequality of stochastic convolution integrals on Hilbert spaces and space-time regularity of linear stochastic partial differential equations*, Stochastics, 21 (1987), pp. 345–358.
[17] I. LASIECKA AND R. TRIGGIANI, *Numerical approximations of algebraic Riccati equations modelled by analytic semigroups and applications*, Math. Comput., 57 (1991), pp. 639–662, 513–537.
[18] ———, *The regulator problem for parabolic equations with Dirichlet boundary control* I, Appl. Math. Optim., 16 (1987), pp. 147–168.
[19] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, 1972.
[20] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 673–689, May 1994

005

SOLUTION OF SOME TRANSPORTATION PROBLEMS WITH RELAXED OR ADDITIONAL CONSTRAINTS*

S. T. RACHEV† AND L. RÜSCHENDORF‡

Abstract. The authors consider some modifications of the usual transportation problem by allowing bounds for the admissible supply—respectively, demand—distributions. In particular, the case that the marginal distribution function of the supply is bounded below by a *df* F_1, while the marginal *df* of the demand is bounded above by a *df* is considered. For the case that the difference of the marginals is fixed—this is an extension of the well-known Kantorovich–Rubinstein problem—the authors obtain new and general explicit results and bounds, even without the assumption that the cost function is of Monge type. The multivariate case is also treated. In the last section, the authors study Monge–Kantorovich problems with constraints of a local type, that is, on the densities of the marginals. In particular, the classical Dobrushin theorem on optimal couplings is extended with respect to total variation.

Key words. marginal problem, Monge function, marginal constraint, transportation problem

AMS subject classifications. 60E15, 49A36

1. Introduction.

For distribution functions F_1, F_2 let $\mathcal{F}(F_1, F_2)$ denote the set of all *df*'s on $\mathbb{R}^2$ with marginals F_1, F_2 (i.e., $F(x,\infty) = F_1(x), F(\infty, y) = F_2(y)$). Then the transportation problem with cost function $c \geq 0$ is to

$$\text{minimize} \int_{\mathbb{R}^2} c(x,y)dF(x,y) \quad \text{over all } F \in \mathcal{F}(F_1,F_2). \tag{1.1}$$

F_1 may be viewed as the supply distribution and F_2 as the demand distribution. Clearly, (1.1) is an infinite dimensional analogue of the discrete transportation problem: given $a_i \geq 0, b_j \geq 0, \sum_{i=1}^m a_i = \sum_{j=1}^n b_j$,

$$\text{minimize} \sum_{i=1}^m \sum_{j=1}^n c_{ij}x_{ij}, \quad \text{subject to the conditions:}$$
$$\sum_{j=1}^n x_{ij} = a_i, \quad 1 \leq i \leq m, \qquad \sum_{i=1}^m x_{ij} = b_j, \quad j = 1,\dots,n, \quad x_{ij} \geq 0, \quad \forall i,j. \tag{1.2}$$

If $c(x,y)$ (respectively (c_{ij})) satisfies the "Monge" conditions, i.e., c is right continuous and

$$c(x',y') - c(x,y') - c(x',y) + c(x,y) \leq 0 \quad \text{for all } x' \geq x, y' \geq y, \tag{1.3}$$

respectively

$$c_{ij} + c_{i+1,j+1} - c_{i,j+1} - c_{i+1,j} \leq 0, \quad \forall\, 1 \leq i < m, 1 \leq j < n, \tag{1.4}$$

then the solution of (1.1), (1.2) is well known and based on the "North-West corner rule," which leads to a greedy algorithm. For (1.1) the solution is given by the *df* F^*

$$F^*(x,y) = \min\{F_1(x), F_2(y)\}. \tag{1.5}$$

* Received by the editors October 15, 1990; accepted (in revised form) December 1, 1992.

† Statistics Department, University of California, Santa Barbara, California 93106. Supported in part by North Atlantic Treaty Organization grant CRG 900798 and by the Ministere de la Recherche et de la Technologie, France.

‡ Institut für Mathematik Statistik, Universität Münster, Einsteinstr. 62, D-4400 Münster, Germany. This author was supported in part by a Deutsche Forschunsgemeinschaft grant.

F^* is the upper Fréchet-bound. The Fréchet-bounds provide the following characterization of $\mathcal{F}(F_1, F_2)$:

(1.6)
$$F \in \mathcal{F}(F_1, F_2) \quad \text{if and only if}$$
$$F_*(x,y) := (F_1(x) + F_2(y) - 1)_+ \le F(x,y) \le F^*(x,y) \quad (\text{here } (\cdot)_+ = \max(0, \cdot)).$$

The lower Fréchet bound yields to a solution of the maximization problem corresponding to (1.1) (cf. [4], [5], [11]–[13]).

In terms of random variables an equivalent formulation of the transportation problem is the following:

(1.7) $$\text{minimize } Ec(X,Y), \quad \text{subject to } F_X = F_1, F_Y = F_2,$$

where X, Y are random variables on a rich enough (e.g., atomless) probability space $(\Omega, \mathcal{U}, \mathcal{P})$. The solutions (1.5) respectively (1.6) then can be represented as distributions of rv's X^*, Y^*:

(1.8) $$X^* = F_1^{-1}(U), \qquad Y^* = F_2^{-1}(U) \quad (\text{for } (1.1), (1.5)),$$

respectively

(1.9) $$X^* = F_1^{-1}(U), \qquad Y^* = F_2^{-1}(1-U) \quad (\text{for } F_*),$$

where U is uniformly distributed on (0,1), and $F_1^{-1}(u) = \inf\{y : F_1(y) \ge u\}$ is the generalized inverse of F_1 (cf. [4], [11]–[13]). (Throughout the paper we assume that *df*'s are right continuous.) For a general review on the Monge-Kantorovich transportation problem we refer to [8] and [1].

In this paper we study modifications of the transportation problem (1.1), where we relax or add new constraints. One type of additional side conditions has been studied by Barnes and Hoffman [2], in the discrete transportation problem (1.2); namely, additional capacity constraints $\sum_{r=1}^{i} \sum_{s=1}^{j} x_{rs} \le \gamma_{ij}, i \le m-1, j \le n-1$, and a solution was obtained by a greedy algorithm.

In the first part of this paper we make use of the assumption that the cost function is of Monge type. These conditions seem to be necessary, since already in the simpler discrete case there are no general explicit solutions without conditions of this type. In the second part, under the restrictions of given difference of the marginals, we obtain explicit results without the Monge condition. We study extensions to the multivariate case for cost functions of the type $c_p(x,y) = \|x-y\|_p$, $\| \quad \|_p$ the p-norm on $\mathbb{R}^n$ (c_p is not a Monge function for $n \ge 2$, and this problem is unsolved also in the discrete case). In the final section we consider local constraints on the marginals. In particular, we extend the classical Dobrushin result providing a construction of optimal couplings.

As for the proof of our results we use different methods from marginal problems, stochastic ordering, and duality theory. It seems that it is not possible to derive them all in a unified way; e.g., in §2, we construct in Theorems 1 and 2 solutions of the transportation problem with upper and lower bounds on the marginals under different assumptions on the cost functions. The proof of Theorem 1—for symmetric cost functions—is based on marginal problems, while the proof of Theorem 2—for unimodal cost functions—is based on stochastic ordering arguments.

2. Relaxation of the marginal constraints. Consider for *df*'s F_1, F_2 the set

(2.1) $$\mathcal{H}(F_1, F_2) = \{F : F \text{ is a } df \text{ on } \mathbb{R}^2 \text{ with marginal } df\text{'s } \tilde{F}_1 \le F_1, \tilde{F}_2 \ge F_2\}$$

of all df's F with $\tilde{F}_1(x) = F(x,\infty) \le F_1(x)$, always $x \in \mathbb{R}^1$, and $\tilde{F}_2(y) = F(\infty,y) \ge F_2(y), \forall\, y \in \mathbb{R}^1$. We study the transportation problem:

$$\text{(2.2)} \qquad \text{minimize} \int_{\mathbb{R}^2} c(x,y)dF(x,y), \quad \text{subject to } F \in \mathcal{H}(F_1,F_2)$$

or, equivalently,

$$\text{(2.3)} \qquad \text{minimize } Ec(X,Y), \quad \text{subject to } F_X \le F_1, F_Y \ge F_2.$$

In the discrete case the problem is to minimize $\sum c_{ij}x_{ij}$ where for some "supplies" $s_1,\dots,s_n$, $a_1 \le s_1, a_1 + a_2 \le s_1 + s_2, \dots$, and for some demands $d_1,\dots,d_n, b_1 \ge d_1, b_1 + b_2 \ge d_1 + d_2, \dots, (a_i, b_i$ as in (1.2)). This describes production and consumption processes based on priorities (e.g., by time) with capacities $s_1,\dots,s_n$, such that what is remained in stage i of the production (respectively consumption) process can be transferred to some of the next stages $i+1,\dots,n$.

THEOREM 1. *Suppose the cost function $c(x,y)$ is symmetric, $c(x,y)$ satisfies the Monge-condition* (1.3), *and let $c(x,x) = 0, \forall\, x$. Define*

$$\text{(2.3)} \qquad H^*(x,y) = \min\{F_1(x), \max\{F_1(y), F_2(y)\}\}, \qquad x,y \in \mathbb{R}.$$

Then

$$\text{(2.4)} \quad \begin{array}{ll} \text{(a)} & H^* \in \mathcal{H}(F_1,F_2), \\ \text{(b)} & H^* \textit{ solves the relaxed transportation problem } (2.2), \\ \text{(c)} & \displaystyle\int_{\mathbb{R}^2} c(x,y)dH^*(x,y) = \int_0^1 c(F_1^{-1}(u), \min(F_1^{-1}(u), F_2^{-1}(u)))du. \end{array}$$

Remark 1. Setting the df $G_1(y) = \max\{F_1(y), F_2(y)\}$, we see from Theorem 1 that the relaxed transportation problem (2.2) is equivalent to the transportation problem (1.1) with marginals F_1, G_1. In terms of random variables a solution is given by

$$\text{(2.5)} \quad X^* = F_1^{-1}(U), \qquad Y^* = G_1^{-1}(U) = \min(F_1^{-1}(U), F_2^{-1}(U)) \qquad \text{(cf. (1.8))}.$$

Proof. From the Monge condition the function $-c(x,y)$ may be viewed as a "distribution function" corresponding to a nonnegative measure μ_c on $\mathbb{R}^2$. Let X, Y be any real rv's and for $x,y \in \mathbb{R}^1$ denote $x \vee y = \max\{x,y\}, x \wedge y = \min\{x,y\}$. Theorem 1 is a consequence of the following two claims.

CLAIM 1 (Cambanis, Simons, and Stout [4], Dall'Aglio [5]). *For $c(x,y) = |x-y|^p$*

$$\text{(2.6)} \qquad \begin{aligned} 2Ec(X,Y) = \int_{\mathbb{R}^2} &(P(X < x \wedge y, Y \ge x \vee y) \\ &+ P(X \ge x \vee y, Y < x \wedge y))\mu_c(dx,dy). \end{aligned}$$

For the proof of Claim 1 define the function $f(x,y,w) : \mathbb{R}^2 \times \Omega \to \mathbb{R}$ by

$$f(x,y,w) = \begin{cases} 1 & \text{if } (X(w) < x, y \le Y(w)) \text{ or } (Y(w) < x, y \le X(w)) \\ 0 & \text{otherwise} \end{cases}$$

By Fubini's theorem,

$$\text{(2.7)} \qquad E_w \int_{\mathbb{R}^2} f(x,y,w)\mu_c(dx,dy) = \int_{\mathbb{R}^2} (E_w f(x,y,w))\mu_c(dx,dy).$$

Next the symmetry of $c(x,y)$ and $c(x,x)=0$ yields

$$\int_{\mathbb{R}^2} f(x,y,w)d\mu_c = -[c(Y(w),Y(w)) + c(X(w),X(w)) \\ -c(X(w),Y(w)) - c(Y(w),X(w))] = 2c(X(w),Y(w)). \tag{2.8}$$

Clearly,

$$E_w f(x,y,w) = P(X < x\wedge y, Y \geq x\vee y) + P(X \geq x\vee y, Y < x\wedge y). \tag{2.9}$$

Combining (2.7), (2.8), (2.9), we obtain (2.6).

CLAIM 2. *Define* $X^* = F_1^{-1}(U), Y^* = \min(F_1^{-1}(U), F_2^{-1}(U))$; *then*

$$Ec(X^*,Y^*) = \min\{Ec(X,Y); F_X \leq F_1, F_Y \geq F_2\} \tag{2.10}$$

and the value of the expectation in (2.10) *is given by*

$$\begin{aligned} Ec(X^*,Y^*) &= \frac{1}{2}\int_{\mathbb{R}^2} \max\{0, F_2((x\wedge y)-) - F_1((x\vee y)-)\}\mu_c(dx,dy) \\ &= \int_0^1 c(F_1^{-1}(t), \min\{F_1^{-1}(t), F_2^{-1}(t)\})dt. \end{aligned} \tag{2.11}$$

For the proof of Claim 2 let X, Y be any rv's with df's $F_X \leq F_1, F_Y \geq F_2$. Using Claim 1 we obtain

$$\begin{aligned} 2Ec(X,Y) &\geq \int_{\mathbb{R}^2} P(X \geq x\vee y, Y < x\wedge y)\mu_c(dx,dy) \\ &= \int_{\mathbb{R}^2} \{P(Y < x\wedge y) - P(X < x\vee y, Y < x\wedge y)\}\mu_c(dx,dy) \\ &\geq \int_{\mathbb{R}^2} \{P(Y < x\wedge y) - \min\{P(X < x\vee y), P(Y < x\wedge y)\}\mu_c(dx,dy) \\ &= \int_{\mathbb{R}^2} (P(Y < x\wedge y) - P(X < x\vee y))_+\mu_c(dx,dy) \\ &\geq \int_{\mathbb{R}^2} (F_2((x\wedge y)-) - F_1((x\vee y)-))_+\mu_c(dx,dy). \end{aligned} \tag{2.12}$$

Next we check that the lower bound we get in (2.12) is attained for $X^* = F_1^{-1}(U), Y^* = \min(F_1^{-1}(U), F_2^{-1}(U))$. In fact, by Claim 1 using $X^* \geq Y^*$ and $\{U < F_2(z)\} = \{F_2^{-1}(U) < z\}$ almost surely we obtain

$$\begin{aligned} &2Ec(X^*,Y^*) \\ &= \int_{\mathbb{R}^2} \{P(X^* \geq x\vee y, Y^* < x\wedge y) + P(X^* < x\wedge y, Y^* \geq x\vee y)\}\mu_c(dx,dy) \\ &= \int_{\mathbb{R}^2} P(X^* \geq x\vee y, Y^* < x\wedge y)\mu_c(dx,dy) \\ &= \int_{\mathbb{R}^2} P(F_1^{-1}(U) \geq x\vee y, \min(F_1^{-1}(U), F_2^{-1}(U)) < x\wedge y)\mu_c(dx,dy) \\ &= \int_{\mathbb{R}^2} P(F_1^{-1}(U) \geq x\vee y, F_2^{-1}(U) < x\wedge y)\mu_c(dx,dy) \\ &= \int_{\mathbb{R}^2} P(U \geq F_1(x\vee y), U < F_2(x\wedge y))_+\mu_c(dx,dy) \\ &= \int_{\mathbb{R}^2} (F_2((x\wedge y)-) - F_1((x\vee y)-))_+\mu_c(dx,dy). \end{aligned} \tag{2.13}$$

Obviously, $F_{(X^*,Y^*)} = H^* \in \mathcal{H}(F_1, F_2)$ and the proof of Theorem 1 is completed. □

Remark 2. Equation (2.5) suggests the following "greedy" algorithm for solving the finite discrete transportation problem with relaxed side conditions:

$$
\begin{aligned}
&\text{minimize} && \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} \\
&\text{subject to:} && x_{ij} \geq 0 \\
&&& \sum_{s=1}^{j}\sum_{r=1}^{n} x_{rs} \geq \sum_{s=1}^{j} b_s =: G_j, \qquad 1 \leq j \leq n \\
&&& \sum_{r=1}^{i}\sum_{s=1}^{n} x_{rs} \leq \sum_{r=1}^{i} a_r =: F_i, \qquad 1 \leq i \leq n,
\end{aligned}
\tag{2.14}
$$

where the sum of the "demands" $\sum_{s=1}^{n} b_s$ equals the sum of the "supplies" $\sum_{r=1}^{n} a_r$, assuming that (c_{ij}) are symmetric, $c_{ii} = 0$ and c satisfying the Monge condition (1.4). Denote

$$
\begin{aligned}
&H_i = \max(F_i, G_i), \qquad 1 \leq i \leq n, \quad \text{and} \\
&\delta_1 = H_1, \delta_{i+1} = H_{i+1} - H_i, \qquad 1 \leq i \leq n-1;
\end{aligned}
\tag{2.15}
$$

(2.14) is equivalent to the standard transportation problem (1.2) with side conditions $(a_i), (\delta_i)$. In the following example we compare the solution of problem (2.14) with inequality constraints with the "greedy" solution of the standard transportation problem with equality constraints (1.2). For the problem with inequality constraints we first calculate the new artificial demands δ_j as in (2.15) and then apply the North-West corner rule.

Example.

							supply a_i	$F_i = \sum_{r=1}^{i} a_r$
$\boxed{y_{ij}}$ x_{ij}	$\boxed{20}$ 10	10					20	20
							0	20
		$\boxed{20}$ 20	$\boxed{20}$ 10	10			40	60
				$\boxed{20}$ 20			20	80
				$\boxed{10}$ 10			10	90
						$\boxed{10}$ 10	10	100
demand b_j	10	30	10	40	0	10		
$G_j = \sum_{s=1}^{j} b_s$	10	40	50	90	90	100		
$H_j = F_j \vee G_j$	20	40	60	90	90	100		
$\delta_1 = H_1, \delta_{j+1} = H_{j+1} - H_j$	20	20	20	30	0	10	"artificial" demands	

x_{ij} = solution of the standard transportation problem (1.2), using the classical North-West corner

y_{ij} = solution of the transportation problem with relaxed side conditions.

We next extend the solution to the non-symmetric case. We assume instead of symmetry the following unimodality condition, saying that for any x, y the functions $c(x,\cdot), c(\cdot,y)$ are unimodal; more precisely,

$$\begin{aligned} &c(x,y_1) \le c(x,y_2) \quad \text{if } x \le y_1 \le y_2 \text{ or } y_2 \le y_1 \le x, \quad \text{and} \\ &c(x_1,y) \le c(x_2,y) \quad \text{if } x_2 \le x_1 \le y \text{ or } y \le x_1 \le x_2. \end{aligned} \tag{2.16}$$

For the proof of this unimodal case we basically make use of stochastic ordering arguments.

THEOREM 2. *If $c(x,x) = 0$ for all x, and c satisfies the Monge condition* (1.3) *and the unimodality condition* (2.16), *then the relaxed transportation problem,*

$$\text{minimize } Ec(X,Y) \text{ subject to: } F_X \ge F_1, F_Y \le F_2, \tag{2.17}$$

has the solution

$$\begin{aligned} &X^* = F_1^{-1}(U), \qquad Y^* = \max\,(F_1^{-1}(U), F_2^{-1}(U)), \quad \text{so} \\ &F_{X^*,Y^*}(x,y) = \min\,(F_1(x), \min\,(F_1(y), F_2(y)) \quad \text{and} \\ &Ec(X^*,Y^*) = \int_0^1 c(F_1^{-1}(u), \max\,(F_1^{-1}(u), F_2^{-1}(u))du. \end{aligned} \tag{2.18}$$

Proof. Let X, Y be rv's with $F_X \ge F_1, F_Y \le F_2$; then by (1.8)

$$Ec(X,Y) \ge Ec(F_X^{-1}(U), F_Y^{-1}(U)). \tag{2.19}$$

Let $G(y) = \min\,(F_X(y), F_Y(y))$; then $F_X^{-1} \le F_1^{-1}, F_Y^{-1} \ge F_2^{-1}$ and $G^{-1} = \max\,(F_X^{-1}, F_Y^{-1})$. We now state the following.

CLAIM 1.

$$\int_0^1 c(F_X^{-1}(u), F_Y^{-1}(u))du \ge \int_0^1 c(F_X^{-1}(u), G^{-1}(u))du. \tag{2.20}$$

To show Claim 1 let for fixed $u \in (0,1), x = F_X^{-1}(u), y_1 = F_X^{-1}(u) \vee F_Y^{-1}(u) = G^{-1}(u)$, $y_2 = F_Y^{-1}(u)$.

Case 1. $x < y_2$. In this case, $x \le y_1 \le y_2$, and, therefore, the unimodality condition (2.18) implies $c(x,y_2) \ge c(x,y_1)$.

Case 2. $y_2 \le x$. In this case, $y_1 = x$ and therefore, $y_2 \le y_1 = x$. Again by the unimodality condition $c(x,y_2) \ge c(x,y_1)$. So Claim 1 holds.

CLAIM 2.

$$\int_0^1 c(F_X^{-1}(u), F_Y^{-1}(u) \vee F_X^{-1}(u))du \ge \int_0^1 c(F_1^{-1}(u), F_2^{-1}(u) \vee F_1^{-1}(u))du. \tag{2.21}$$

For the proof, define $\tilde{x}_1 = F_X^{-1}(u), \tilde{x}_2 = F_Y^{-1}(u), x_1 = F_1^{-1}(u), x_2 = F_2^{-1}(u)$ for fixed u. Then $\tilde{x}_1 \le x_1, x_2 \le \tilde{x}_2$.

$$\begin{aligned} &\text{If} \quad \tilde{x}_1 < \tilde{x}_2, \quad \text{then } \tilde{x}_1 \le \tilde{x}_2 \vee x_2 \le \tilde{x}_2, \\ &\text{if} \quad \tilde{x}_1 \ge \tilde{x}_2, \quad \text{then } \tilde{x}_1 = \tilde{x}_1 \vee x_2 \ge \tilde{x}_2. \end{aligned} \tag{2.22}$$

From (2.22) we obtain the following claim.

CLAIM 3.

$$c(\tilde{x}_1, \tilde{x}_1 \vee x_2) \ge c(x_1, x_1 \vee x_2). \tag{2.23}$$

For the proof of Claim 3 we use the relation $x_1 \geq \tilde{x}_1$. By (2.22) we have two cases.

Case 1. $x_2 > x_1 > \tilde{x}_1$. Then $c(\tilde{x}_1, x_2) = c(\tilde{x}_1, \tilde{x}_1 \vee x_2) \geq c(x_1, x_2) = c(x_1, x_1 \vee x_2)$ by the unimodality condition.

Case 2. (a) $x_1 \geq x_2 \geq \tilde{x}_1$. Then, trivially, $c(\tilde{x}_1, x_2) = c(\tilde{x}_1, x_2 \vee \tilde{x}_1) \geq c(x_1, x_1 \vee x_2) = c(x_1, x_1) = 0$.

(b) $x_1 \geq \tilde{x}_1 \geq x_2$. Then again $c(\tilde{x}_1, \tilde{x}_1) = c(\tilde{x}_1, \tilde{x}_1 \vee x_2) \geq c(x_1, x_1 \vee x_2) = c(x_1, x_1) = 0$, trivially.

Claims 1, 2, and 3 imply (2.18).

Remark 3.

(a) The unimodality assumption (2.16) is natural from the application point of view. Note that the transportation problem in Theorem 2 is the same as in Theorem 1 (where only the indices 1 and 2 have been changed). We used this change to demonstrate that the optimal solution F^* is not unique, but there is a large range of solutions. As a consequence observe that in order to achieve an optimal solution for the transportation problem with side conditions, either the demands can be adjusted by transports on or below the diagonal, or alternatively, the supplies can be adjusted in a similar way. Without the symmetry, respectively the unimodality condition, the solution may change extremely. Consider for any right continuous function $f = f(y) \geq 0$ the cost function $c(x, y) = f(y)$. Then c satisfies the Monge-condition, and so (2.17) is equivalent to the problem,

$$\text{minimize} \int f(y) dF_Y(y) \quad \text{subject to } F_Y \leq F_2, \tag{2.24}$$

i.e., we are looking for a $df \tilde{F}_2 \leq F_2$, such that the distribution of f with respect to $\tilde{F}_2$ has a minimal first moment. Obviously, the solution (2.20) of Theorem 2 is not a solution of (2.24).

(b) For the proof of Theorem 2 the assumption $c(x, x) = 0$ can be replaced by the weaker one,

$$c(x, x) \leq c(x, y) \wedge c(y, x), \quad \forall x, y. \tag{2.25}$$

3. Given sum of the marginals. Consider a flow in a network with n-nodes $i = 1, \ldots, n$, and let x_{ij} be the flow from node i to node j. Assume that for all nodes k the value of $\sum_i x_{ik} + \sum_j x_{kj}$ is fixed to be h_k. For a motivation of this problem let $a_i = \sum_{k=1}^n x_{ik}$, $b_i = \sum_{k=1}^n x_{ki}$ be the amount of labor corresponding to the outflow respectively to the inflow in node i. Assume that the total labor capacity in node i is given by h_i (in a certain time unit); then an admissible flow (x_{ij}) should satisfy the condition

$$h_i = a_i + b_i, \qquad 1 \leq i \leq n. \tag{3.1}$$

Let $F_1(k) = \sum_{i=1}^k a_i$, $F_2(k) = \sum_{i=1}^k b_i$, $H(k) = \sum_{i=1}^k h_k$; then $h_k = F_1(k) + F_2(k) - (F_1(k-1) + F_2(k-1))$ and (3.1) is equivalent to

$$H(k) = F_1(k) + F_2(k), \qquad 1 \leq k \leq n. \tag{3.2}$$

Let c_{ij} denote the cost of transporting a unit from node i to node j; then the problem is to minimize the total cost $\sum c_{ij} x_{ij}$ subject to condition (3.2) and $x_{ij} \geq 0$.

The general formulation of this problem is the following. For two df's F_1, F_2 define $G(x) := \frac{1}{2}(F_1(x) + F_2(x))$. For a cost function $c(x, y)$ consider the problem,

$$\text{minimize} \int_{\mathbb{R}^2} c(x, y) dF(x, y) \quad \text{subject to } F \in \mathcal{F}_G, \tag{3.3}$$

where $\mathcal{F}_G$ is the set of all df's $F(x,y)$ with marginal df's $\tilde{F}_1, \tilde{F}_2$ satisfying $\tilde{F}_1(x) + \tilde{F}_2(x) = 2G$.

In the special case $c(x,y) = |x-y|$, let X, Y be real rv's. Then by the triangle inequality

$$E|X-Y| \leq \inf_{a \in \mathbb{R}^1} (E|X-a| + E|Y-a|), \tag{3.4}$$

(3.4) is the optimal bound if one knows only $E|X-a|, a \in \mathbb{R}^1$. Note that $E|X-a| + E|Y-a| = \int |x-a| d(F_X + F_Y)(x)$ only depends on the sum of the marginals. Equation (3.3) is the best possible improvement of (3.4) provided $F_X + F_Y$ is known.

It was shown in [9] that

$$\sup\{E|X-Y|^p; F_X + F_Y = 2G\} = \int_0^1 |G^{-1}(t) - G^{-1}(1-t)|^p dt, \qquad p \geq 1. \tag{3.5}$$

PROPOSITION 3. *If $c \geq 0$ is symmetric and satisfies the Monge condition* (1.3), *then*

$$\inf\left\{\int c(x,y) dF(x,y); F \in \mathcal{F}_G\right\} = \int_0^1 c(G^{-1}(u), G^{-1}(u)) du, \tag{3.6}$$

$$\sup\left\{\int c(x,y) dF(x,y); F \in \mathcal{F}_G\right\} = \int_0^1 c(G^{-1}(u), G^{-1}(1-u) du. \tag{3.7}$$

Optimal pairs of rv's are given by $(G^{-1}(U), G^{-1}(U))$ respectively $(G^{-1}(U), G^{-1}(1-U))$.

Proof. Since c is symmetric, we obtain for any $F \in \mathcal{F}_G$, $\int c(x,y) dF(x,y) = \int \frac{1}{2}(c(x,y) + c(y,x)) dF(x,y) = \int c(x,y) d\{[F(x,y) + F(y,x)]/2\}$. But $F_s(x,y) = [F(x,y) + F(y,x)]/2 \in \mathcal{F}(G,G)$, so we obtain (3.6), (3.7) by application of (1.8), (1.9). □

For non-symmetric cost functions we have the following.

PROPOSITION 4. *If $c(x,y)$ satisfies the Monge condition and furthermore $x_1 \leq y \leq x_2$ implies that $c(x_1, x_2) \geq c(y,y)$, then*

$$\inf\left\{\int c(x,y) dF(x,y); F \in \mathcal{F}_G\right\} = \int_0^1 c(G^{-1}(u), G^{-1}(u)) du. \tag{3.8}$$

Proof. For rv's X, Y with $F_{X,Y} \in \mathcal{F}_{A+B}$, by the Monge condition $Ec(X,Y) \geq Ec(F_X^{-1}(U), F_Y^{-1}(U))$. Since $F_X(x) + F_Y(x) = 2G(x)$, it follows that $F_X \wedge F_Y \leq G \leq F_X \vee F_Y$, and therefore, $F_X^{-1} \wedge F_Y^{-1} \leq G^{-1} \leq F_X^{-1} \vee F_Y^{-1}$. It follows that $c(F_X^{-1}(U), F_Y^{-1}(U)) \geq c(G^{-1}(U), G^{-1}(U))$ implying (3.8). □

Remark 4. The set of marginals in the class $\mathcal{F}_G$ has a smallest and a largest element, namely

$$F_1^*(x) = \begin{cases} 2G(x), & x < x_0 \\ 1 & x \geq x_0 \end{cases} \quad \text{and} \quad F_2^*(x) = \begin{cases} 2G(x) - 1, & x \geq x_0 \\ 0 & x < x_0 \end{cases},$$

where $x_0 = \inf\{y; 2G(y) \geq 1\}$. There is no smallest df in $\mathcal{F}_G$. For the proof let $H_1(x), H_2(x)$ be the marginal df's of a smallest element $H \in \mathcal{F}_G$ and let G_1, G_2 be df's such that $G_1(x) + G_2(x) = 2G(x)$. If the lower Fréchet bounds satisfy $(H_1(x) + H_2(y) - 1)_+ \leq (G_1(x) + G_2(y) - 1)_+$, then $H_1 \leq G_1$ and $H_2 \leq G_2$, which amounts to $H_1 = G_1, H_2 = G_2$. In particular, this implies that $(G^{-1}(U), G^{-1}(1-U))$ is in the general non-symmetric case no longer a solution to the problem to maximize $\int c(x,y) dF(x,y)$ in the class $\mathcal{F}_G$. Let e.g., G be the df of $\frac{1}{4}\sum_{i=1}^4 \varepsilon_{\{i\}}$; then $P_1 = P^{(G^{-1}(U), G^{-1}(1-U))} = \frac{1}{4}(\varepsilon_{(1,4)} + \varepsilon_{(2,3)} + \varepsilon_{(3,2)} + \varepsilon_{(4,1)})$, while $P_2 = P^{((F_1^*)^{-1}(U), (F_2^*)^{-1}(1-U))} = \frac{1}{2}(\varepsilon_{(1,4)} + \varepsilon_{(2,3)})$. For $c_1(x,y) = 1_{(-\infty,(3,2)]}(x,y)$, we have $E_{P_1} c_1 = \frac{1}{4}, E_{P_2} c_1 = 0$, while for $c_2 = 1_{[(2,3),\infty)}, E_{P_1} c_2 = \frac{1}{4}, E_{P_2} c_2 = \frac{1}{2}$. Note that both functions, $-c_1, -c_2$, are Monge functions (but are not unimodal).

4. Given difference of the marginals. We next consider the case where in the network example we fix the total outflow minus the inflow of each node. This problem is known in the literature as minimal network flow problem (cf. e.g., [3, §9], or [1]). Similarly to §3 the outflow minus the inflow of each node is fixed; i.e., the following Kirchhoff equations hold: $\sum_k x_{ik} - \sum_k x_{ki} = a_i - b_i = h_i$ for all i, or, equivalently, with $F_1(k) = \sum_{j=1}^k a_j, F_2(k) = \sum_{j=1}^k b_j, H(k) = \sum_{j=1}^k h_j, H(k) = F_1(k) - F_2(k), 1 \le k \le n$. Let more generally F_1, F_2 be distribution functions and let $\mathcal{F}_H$ be the set of all "df's" of finite measures on $\mathbb{R}^2$ with marginals $\tilde{F}_1, \tilde{F}_2$ satisfying $\tilde{F}_1 - \tilde{F}_2 = F_1 - F_2 =: H$. We consider the following transportation problem:

$$\text{(4.1)} \qquad \text{minimize } \int c(x,y)dF(x,y) \quad \text{subject to } F \in \mathcal{F}_H.$$

$c(x,y)$ is symmetric, nonnegative and continuous, but does not need to satisfy the Monge conditions. For the solution we shall make use of the following dual representation (cf. Rachev and Shortt [10]):

$$\text{(4.2)} \qquad \begin{aligned} &\inf\left\{\int c(x,y)dF(x,y); F \in \mathcal{F}_H\right\} \\ &= \sup\left\{\int f dH(x); f(x) - f(y) \le c(x,y), \forall x, y\right\}. \end{aligned}$$

We first consider a special type of cost functions.

PROPOSITION 5. *Let* $c(x,y) = |x-y| \max(1, h(|x-a|), h(|y-a|))$, *where* h *is monotonically nondecreasing. Then*

$$\text{(4.3)} \qquad \inf\left\{\int c(x,y)dF(x,y); F \in \mathcal{F}_H\right\} = \int \max(1, h(|x-a|))|H|(x)dx,$$

provided $h(|x-a|)$ *is locally integrable.*

Proof. For the cost function c we observe that $f(x) - f(y) \le c(x,y)$, for all x, y, if and only if f is absolutely continuous with $|f'(x)| \le \max(1, h(|x-a|))$ almost surely. By the dual representation (4.2) and partial integration we obtain

$$\begin{aligned} &\inf\left\{\int c(x,y)dF(x,y); F \in \mathcal{F}_H\right\} \\ &= \sup\left\{\int f d(H)(x); |f'(x)| \le \max(1, h(|x-a|)), \forall x\right\} \\ &= \sup\left\{\int f'(x)(H)(x)dx; |f'(x)| \le \max(1, h(|x-a|)), \forall x\right\} \\ &= \int \max(1, h(|x-a|))|H|(x)dx. \qquad \square \end{aligned}$$

On the basis of the idea of this proof, we next consider more generally

$$\text{(4.4)} \qquad c(x,y) = |x-y|\zeta(x,y) \quad \left(\text{i.e. } \zeta(x,y) = \frac{c(x,y)}{|x-y|}\right).$$

THEOREM 6 (Generalized Kantorovich–Rubinstein problem). *Assume that for any* $x < t < y, \zeta(t,t) \le \zeta(x,y), \zeta(x,y)$ *symmetric and continuous on the diagonal and also that* $t \to \zeta(t,t)$ *is locally bounded; then*

$$\text{(4.5)} \qquad \inf\left\{\int c(x,y)dF(x,y); F \in \mathcal{F}_H\right\} = \int \zeta(t,t)|H|(t)dt.$$

Proof. Let $\mathcal{F} = \{f : f(x) - f(y) \le c(x,y), \text{ for all } x, y\}$ and let $\mathcal{F}^* = \{f \text{ absolutely continuous and } |f'(t)| \le \zeta(t,t), \text{ for all } t\}$; then $\mathcal{F} \subset \mathcal{F}^*$ as for $f \in \mathcal{F}$, we have $[f(x) - f(y)]/|x-y| \le \zeta(x,y)$ and, therefore, $\overline{\lim}_{y\to x}[f(x) - f(y)]/|x-y| \le \zeta(x,x)$. Also $\underline{\lim}_{y\to x}[f(x) - f(y)]/|x-y| = -\overline{\lim}[f(y) - f(x)]/|x-y| \ge -\overline{\lim}\,\zeta(y,x) = -\zeta(x,x)$. As $\zeta(t,t)$ is locally bounded, f is locally Lipschitz, absolutely continuous, and the inequalities above imply that $|f'(t)| \le \zeta(t,t)$ almost surely. If, conversely, $f \in \mathcal{F}^*$, then $f(x) - f(y) = \int_x^y f'(t)dt$, and therefore, $|f(x) - f(y)| \le \int_x^y |f'(t)|dt \le \int_x^y \zeta(t,t)dt \le |x-y|\zeta(x,y) = c(x,y)$. The dual representation (4.2) again implies (4.3) as in the proof of Proposition 5. □

It is very interesting to observe that restrictions on the difference of the marginals allow this general explicit result without "special" assumptions on c. Note that the solution only depends on the behavior of c at the diagonal, a property that is observed in the minimal network flow problems. Note that from Theorem 6 one obtains the remarkable consequence that

$$\inf\left\{\int |x-y|^p dF(x,y); F \in \mathcal{F}_H\right\} = 0 \tag{4.6}$$

for all $p > 1$, which confirms that cost functions as in Theorem 5 are of the right order.

We next consider an extension to the multivariate case $\mathbb{R}^n$ with the class of cost functions

$$c_p(x,y) = \|x-y\|_p = \left(\sum_{i=1}^n |x_i - y_i|^p\right)^{1/p}, \qquad 1 \le p < \infty.$$

Let F_1, F_2 be n-dimensional distribution functions and let for $H := F_1 - F_2$; $\mathcal{F}_H$ denotes the class of all $2n$-dimensional (joint) distribution functions F with n-dimensional marginals $\tilde{F}_1, \tilde{F}_2$ such that $\tilde{F}_1 - \tilde{F}_2 = H$. Denote

$$A_p(H) := \inf\left\{\int_{\mathbb{R}^{2n}} \|x-y\|_p dF(x,y); F \in \mathcal{F}_H\right\},$$

the value of the optimal multivariate transportation costs. Let $1/q + 1/p = 1$ and assume that F_1, F_2 have densities f_1, f_2 with respect to the Lebesgue measure $h := f_1 - f_2$.

THEOREM 7. (Multivariate transportation problem). (a) *For the value of the optimal transportation costs we have the upper bound*

$$A_p(H) \le B_p(H) := \int_{\mathbb{R}^n} \|y\|_p |J_H(y)|dy, \tag{4.8}$$

where $J_H(y) := \int_0^1 t^{-(n+1)} h(y/t)dt$.

(b) *If there exists a continuous function* $g : \mathbb{R}^n \to \mathbb{R}^1$, *almost everywhere differentiable and satisfying for* $p = 1$

$$\nabla g(y) = (\operatorname{sgn}(y_i J_H(y))) \quad \text{a.e.}, \tag{4.9}$$

respectively for $p > 1$,

$$\nabla g(y) = (\operatorname{sgn}(y_i J_H(y)))\left(\frac{|y_i|}{\|y\|_q}\right)^{q/p}, \tag{4.10}$$

then equality in (4.8) *holds.*

Proof. (a) From the duality theorem in Rachev and Shortt [10]

$$A_p(H) = \sup\left\{ \left| \int_{\mathbb{R}^n} f dH \right| ; |f(x) - f(y)| \le \|x - y\|_p \right\}.$$

From the Radermacher theorem we infer that any Lipschitz function f is almost everywhere differentiable, and as $\sup\{\langle \nabla f(y), a\rangle; \|a\|_p = 1\} = \|\nabla f(y)\|_q$, we obtain from the Lipschitz condition that $\|\nabla f(y)\|_q \le 1$ almost everywhere. Using a Taylor expansion

$$f(y) = f(0) + \int_0^1 \langle \nabla f(ty), y\rangle \, dt,$$

we conclude that

$$\begin{aligned} A_p(H) &\le \sup\left\{ \left| \int_{\mathbb{R}^n} \int_0^1 \langle \nabla f(ty), y\rangle \, dt \, h(y) \, dy \right| ; \|\nabla f(y)\|_q \le 1 \text{ a.e.} \right\} \\ &= \sup\left\{ \left| \int_{\mathbb{R}^n} \int_0^1 \langle \nabla f(y), y\rangle \frac{1}{t^{n+1}} h\left(\frac{y}{t}\right) dt \, dy \right| ; \|\nabla f(y)\|_q \le 1 \text{ a.e.} \right\} \\ &\le \sup\left\{ \int_{\mathbb{R}^n} \|y\|_p \, |J_H(y)| \, \|\nabla f(y)\|_q \, dy \, \|\nabla f(y)\|_q \le 1 \text{ a.e.} \right\} \\ &\le \int_{\mathbb{R}^n} \|y\|_p \, |J_H(y)| \, dy. \end{aligned} \tag{4.11}$$

(b) In the inequalities

$$|\langle x, y\rangle| \le \sum |x_i y_i| \le \|x\|_p \, \|y\|_q, \qquad \|x\|_p = 1,$$

equality is attained for $p > 1$ if and only if

$$x_i = \operatorname{sgn} y_i \frac{|y_i|^{q/p}}{\|y\|_q^{q/p}} = y_i \frac{|y_i|^{q/p-1}}{\|y\|_q^{q/p}},$$

while for $p = 1$ equality holds if and only if $\operatorname{sgn} x_i = \operatorname{sgn} y_i$. This implies part (b) of the Theorem. □

Remark 5. Conditions (4.9), (4.10) are fulfilled in dimension 1 so that the bound (4.8) is sharp and coincides with (4.3). A simple sufficient for $p = 1$ for (4.9) is given by

$$J_H \ge 0 \quad \text{a.e.}, \tag{4.12}$$

which is a stochastic ordering condition. More generally we can allow a "simple" structure of the set $\{J_H \ge 0\}$. We remark that the optimal multivariate transportation problem is a longstanding open problem also in the discrete case.

5. Upper bounds on the total transport mass. Let $\Gamma(x, y)$ be a "distribution function" of a finite measure and define for two fixed distribution functions F_1, F_2 on $\mathbb{R}^1$ the transportation problem:

$$H_\Gamma(x, y) := \sup\{F(x, y); F(x_i, y_i) \le \Gamma(x_i, y_i), i \in I, F \in \mathcal{F}(F_1, F_2)\}, \tag{5.1}$$

where $(x_i, y_i)_{i\in I} \subset \mathbb{R}^2$ may be finite or not. From the Fréchet-bounds in (1.6), we have the following conditions ensuring the nontriviality of the problem:

$$\Gamma(x_i, y_i) \ge (F_1(x_i) + F_2(y_i) - 1)_+, \quad \forall i \in I. \tag{5.2}$$

Problem (5.1) is an extension of a problem treated by Barnes and Hoffman [2] in the finite discrete case and by Olkin and Rachev [7] in the general case. In these papers it was assumed that $F(x,y) \le \Gamma(x,y)$ for all (x,y). Problem (5.2) is motivated by capacitated transportation problems with linearly ordered supply and demand nodes (cf. [2]). Several examples of this problem and extensions to further restrictions on the support of solutions ("staircase supports") are discussed in Hoffman and Veinott [6]. An application to a graph partitioning problem is given in Barnes and Hoffman [2].

THEOREM 8. *Let assumption* (5.2) *be fulfilled and define*

$$F^*(x,y) := \inf_{\substack{x_i \le x \\ y_i \le y}} \{\Gamma(x_i,y_i) + (F_1(x) - F_1(x_i)) + (F_2(y) - F_2(y_i))\} \wedge \min\{F_1(x), F_2(y)\} \tag{5.3}$$

(*with the convention that the infimum is zero, if there do not exist* $x_i \le x, y_i \le y$).

(a) $H_\Gamma(x,y) \le F^*(x,y), \forall\, x,y.$

(b) *If* F^* *is a df, then*

$$H_\Gamma(x,y) = F^*(x,y) \quad \textit{and} \quad F^* \textit{ is a solution of } (5.1). \tag{5.4}$$

(c) (cf. [2], [7]). *If* $\{(x_i,y_i), i \in I\} = \mathbb{R}^2$, *then* F^* *is a df.*

Proof. (a) For $x_i \le x, y_i \le y$, we have for any admissible F using rv's X, Y with $F_{X,Y} = F, F(x,y) = P(X \le x_i, Y \le y) + P(x_i < X \le x, Y \le y) = P(X \le x_i, Y \le y_i) + P(X \le x_i, y_i < Y \le y) + P(x_i < X \le x, Y \le y) \le \Gamma(x_i,y_i) + F_1(x) - F_1(x_i) + F_2(y) - F_2(y_i)$. Furthermore, by the Fréchet bounds (1.6), $F(x,y) \le \min\{F_1(x), F_2(y)\}$. Therefore, $F(x,y) \le F^*(x,y)$.

(b) If F^* is a *df*, then $F^* \in \mathcal{F}(F_1,F_2)$. For the proof observe that for $(x_i,y_i) \le (x,y)$ by (5.2), $\Gamma(x_i,y_i) + F_1(x) - F_1(x_i) + F_2(y) - F_2(y_i) \ge (F_1(x) + F_2(y) - 1)_+$ and so by definition of F^*, $(F_1(x) + F_2(y) - 1)_+ \le F^*(x,y) \le \min\{F_1(x), F_2(y)\}$, which implies by (1.6) that $F^* \in \mathcal{F}(F_1,F_2)$. Since $F^*(x_i,y_i) \le \Gamma(x_i,y_i)$, F^* is an admissible *df*, and, therefore, by (a) a solution of (5.1).

(c) For the proof of (c) we refer to [7]. □

Remark 6. (a) Parts (a) and (b) of Theorem 7 remain valid for any function $\Gamma(x,y) \ge 0$. The difficult part to verify is that F^* is a *df*. But it seems to be clear from the proof that, even in case when F^* is not a *df*, part (a) gives a good upper bound. An indication for this conclusion is part (c) of the theorem.

(b) From (5.4) one obtains for Monge functions c with the regularity condition

$$\int c(x,y_0)F_1(dx) + \int c(x_0,y)F_2(dy) < \infty$$

for some $x_0, y_0 \in \mathbb{R}$ that

$$\inf\left\{\int c(x,y)dF(x,y); F(x,y) \le \Gamma(x,y), \forall\, x,y, F \in \mathcal{F}(F_1,F_2)\right\} = \int c(x,y)dF^*(x,y). \tag{5.5}$$

(c) In the discrete case the solution F^* of (5.5) can be determined by the Barnes-Hoffman greedy algorithm (see [2], [6], [7]). In fact, if $a_i = F_1(x_i) - F_1(x_i-), i \in M =$

$\{1,\dots,m\}, j \in N = \{1,\dots,n\}, b_i = F_2(y_j) - F_2(y_j-), j \in N = \{1,\dots,n\}, \sum_{i\in M} a_i = \sum_{j\in N} b_j = 1, \sigma_{ij} = \Gamma(x_i, y_j)$, then

$$F^*(x_i, y_j) = \sum_{r=1}^{i}\sum_{s=1}^{j} p_{rs},$$

where p_{rs} are recursively defined:

$$p_{11} = \min(a_1, b_1, \sigma_{11});$$

$$p_{ij} = \min\left\{a_i - \sum_{s=1}^{j-1} p_{is}, b_j - \sum_{r=1}^{i-1} p_{rj}, \sigma_{ij} - \sum_{\substack{r\le i \\ (r,s)\neq(i,j)}}\sum_{s\le j} p_{rs}\right\}$$

if p_{rs} is determined for $r \le i < m$ and $s \le j < n$; and

$$p_{ij} = \min\left\{a_i - \sum_{s=1}^{j-1} p_{is}, b_j - \sum_{r=1}^{i-1} p_{rj}\right\}$$

if $i = m$ or $j = n$.

(d) $F(x, y)$ can be viewed as the analogue of the upper Fréchet bound in the set $\mathcal{F}(F_1, F_2)$ under the side constraint $F^*(x, y) \le \Gamma(x, y)$. To obtain a similar analogue for the lower Fréchet bound, consider

$$\max\{G(x, y) : G(x_i, y_i) \le \Delta(x_i, y_i), i \in I, G \in \mathcal{G}(F_1, F_2)\},$$

where $\mathcal{G}(F_1, F_2)$ is the set of all probabilities

$$G(x, y) = G_\mu(x, y) = \mu((-\infty, x] \times [y, \infty)),$$

$x, y \in \mathbb{R}$ of probability measures μ having marginal df's F_1 and F_2, and where Δ determines a positive measure δ by $G_\delta = \Delta$. Then the above maximum is attained at

$$\begin{aligned}\tilde{G}(x, y) = \inf_{\substack{x_i \le x \\ y_i \ge y}} &\{\Delta(x_i, y_i) + F_1(x) - F_1(x_i) + F_2(y_i-) \\ &- F_2(y-)\} \wedge F_1(x) \wedge (F_2(y_i-) - F_2(y-))\end{aligned} \tag{5.6}$$

if and only if $\Delta(x_i, y_i) \ge \max(0, F_1(x_i) - F_2(y_i-)), i \in I$ and $\tilde{G}$ generates a measure. If $\{(x_i, y_i), i \in I\} = \mathbb{R}^2$, then $\tilde{G}$ defines an optimal measure $\tilde{\mu}$ by $G_{\tilde{\mu}} = \tilde{G}$. Moreover under the same regularity conditions as in (b)

$$\begin{aligned}&\sup\left\{\int c(x, y)\mu(dx, dy); G_\mu \in \mathcal{G}, G_\mu(x, y) \le \Delta(x, y), x, y \in \mathbb{R}\right\} \\ &= \int c(x, y)\tilde{\mu}(dx, dy),\end{aligned}$$

(cf. [7] and Theorem 8).

(e) Consider the discrete version of the extremal problem in (d): Find

$$\max\left\{\sum_{i\in M}\cdot\sum_{j\in N} c_{ij}p_{ij}, \text{ subject to } \sum_{j\in N} p_{ij} = a_i \sum_{i\in M} p_{ij} = b_j \right.$$

$$\left. \text{and } \sum_{r\le i}\sum_{s\ge j} p_{rs} \le \Delta(x_i, y_j), i \in M, j \in N\right\},$$

where

$$\sum_{j\in N} b_j = \sum_{i\in M} a_i = 1.$$

Then the solution is determined by

$$\tilde{G}(x_i, y_j) = \sum_{r=1}^{i}\sum_{s=j}^{n} p_{rs}$$

$$= \min_{\substack{1\le r\le i\\ j\le s\le n}} \{\Delta(x_i, y_j) + (a_{r+1} + \cdots + a_i) + (b_j + \cdots + b_{s-1})\} \wedge \sum_{r=1}^{i} a_r \wedge \sum_{s=j}^{n} b_s,$$

or in other words by the following greedy algorithm:

$$p_{1n} = \min\{a_i, b_n, \Delta(x_1, y_n)\},$$

$$p_{ij} = \min\left\{a_i - \sum_{s=j+1}^{n} p_{is}, b_j - \sum_{r=1}^{i-1} p_{rj}, \Delta(x_i, y_j) - \sum_{\substack{r\le i\\ (r,s)\neq(i,j)}}\sum_{s\ge j} p_{rs}\right\},$$

if p_{rs} is determined for $r \le i \le m-1$ and $s \ge j > 1$; and

$$p_{ij} = \min\left\{a_i - \sum_{s=j+1}^{n} p_{is}, b_j - \sum_{r=1}^{i-1} p_{rj}\right\}$$

if $i = m$ or $j = 1$ (cf. [7]). □

Consider more generally a finite measure μ on $(\mathbb{R}^2, \mathcal{B}^2)$ and define for two probability measures P_1, P_2 on $(\mathbb{R}^1, \mathcal{B}^1)$ and $A_i \times B_i \in \mathcal{B}^1 \otimes \mathcal{B}^1, i \in I$,

$$M^{\mu}(P_1, P_2) = \{P \in M^1(P_1, P_2); P(A_i \times B_i) \le \mu(A_i \times B_i), i \in I\}, \tag{5.7}$$

where $M^1(P_1, P_2)$ denotes the set of all probability measures P on $\mathbb{R}^2$ with marginals P_1, P_2. As in (5.2), we assume

$$\mu(A_i \times B_i) \ge (P_1(A_i) + P_2(B_i) - 1)_+. \tag{5.8}$$

THEOREM 9. *Under assumption* (5.8) *define*

$$\begin{aligned} P^*(A \times B) = \inf_{\substack{A_i\subset A\\ B_i\subset B}} &\{\mu(A_i \times B_i) + (P_1(A) - P_1(A_i)) \\ &+ (P_2(B) - P_2(B_i))\} \wedge \min(P_1(A), P_2(B)), A, B \in \mathcal{B}^1. \end{aligned} \tag{5.9}$$

Then

$$h_{\mu}(A \times B) := \sup\{P(A \times B); P \in M^{\mu}(P_1, P_2)\} \le P^*(A \times B). \tag{5.10}$$

If P^* defines a measure, then

$$h_{\mu}(A \times B) = P^*(A \times B) \quad \textit{and } P^* \textit{ is a solution of } (5.9). \tag{5.11}$$

The proof of Theorem 9 is similar to that of Theorem 8. In contrast to Theorem 8 it allows us to consider "local" bounds in the transportation problem. Observe that in the finite discrete case bounds of the type

$$x_{ij} \le \mu_{ij} \quad \text{for some } (i, j) \tag{5.12}$$

are of this "local" type. So far in the literature there are no results concerning the solution of problem (5.6) respectively (5.12) with local bounds.

6. Local bounds for the transportation plans. While in the preceding sections the additional constraints were formulated mainly in terms of the *df*'s we now consider local constraints formulated for the densities. These restrictions of the local type of course are in some respect much stronger than those in §2 and generally they are much more difficult to handle.

Our first result deals with a transportation problem with the cost function

$$c(x,y) = I(x \neq y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y; \end{cases} \tag{6.1}$$

i.e., the cost of transportation is one for any unit that has to be moved and zero otherwise. c does not satisfy a Monge-type condition. We formulate this problem in a general measure space $(S,\mathcal{U})$ only assuming that

$$\{(x,y) : x \neq y\} \in \mathcal{U} \otimes \mathcal{U}. \tag{6.2}$$

Let $M_f(S), M_f(S \times S)$ denote the set of all finite measures on $(S,\mathcal{U})$ respectively. $(S \times S, \mathcal{U} \otimes \mathcal{U})$ and let for $\mu \in M_f(S \times S), \pi_i\mu, i = 1,2$, denote the marginals of μ. This transportation problem leads to an extension of Dobrushins result on optimal couplings.

THEOREM 10. (Optimal couplings with local restrictions). *Assume that* (6.2) *holds and let* $\mu_1, \mu_2 \in M_f(S)$ *with* $\mu_1(S) \leq \mu_2(S)$. *Then*

(a)

$$\begin{aligned} &\inf\{\mu((x,y); x \neq y); \mu \in M_f(S \times S), \pi_1\mu \geq \mu_1, \pi_2\mu \leq \mu_2\} \\ &= \lambda^-(S) := \sup_{C \in \mathcal{U}} (\mu_1(C) - \mu_2(C)). \end{aligned} \tag{6.3}$$

(b) *The infimum in* (6.3) *is attained for*

$$\mu^*(A \times B) = \gamma(A \cap B) + \frac{\lambda^-(A)\lambda^+(B)}{\lambda^+(S)}, \tag{6.4}$$

where $\lambda^+(A) = \sup_{C \subset A}(\mu_2 - \mu_1)(C), \lambda^-(A) = \sup_{C \subset A}(\mu_1 - \mu_2)(C)$ *and* $\gamma(A) = \mu_2(A) - \lambda^+(A) = \mu_1(A) - \lambda^-(A)$.

Proof. For any $\mu \in M_f(S \times S), \mu(x \neq y) \geq \sup_C \mu(C \times (S \backslash C)) = \sup_C \{\mu(C \times S) - \mu(C \times C)\} \geq \sup_C \{\mu(C \times S) - \mu(S \times C)\} \geq \sup_C \{\mu_1(C) - \mu_2(C)\} = \sup_C \{\lambda^-(C) - \lambda^+(C)\} = \lambda^-(\operatorname{supp} \lambda^-) = \lambda^-(S)$. On the other hand, $\mu^*(A \times S) = \gamma(A) + \lambda^-(A)\lambda^+(S)/\lambda^+(S) = \mu_1(A)$ and $\mu^*(S \times B) = \gamma(B) + \lambda^-(S)\lambda^+(B)/\lambda^+(S) \leq \gamma(B) + \lambda^+(B) = \mu_2(B)$ and $\mu^*(x \neq y) = \int I(x \neq y)(\gamma(dx,dy) + \lambda^-(dx)\lambda^+(dy)/\lambda^+(S)) = \int I(x \neq y) \lambda^-(dx)\lambda^+(dy)/\lambda^+(S) = \lambda^-(S)\lambda^+(S)/\lambda^+(S) = \lambda^-(S)$. □

Consider next finite measures μ_1, μ_2 on $\mathbb{R}$ with densities h_1, h_2 with respect to a dominating measure μ on $\mathbb{R}^1$. Define

$$\mathcal{P}_{\mu_1}^{\mu_2} := \{P \in M^1(\mathbb{R}^2, \mathcal{B}^2); \pi_1 P \geq \mu_1, \pi_2 P \leq \mu_2\}. \tag{6.5}$$

Any $P \in \mathcal{P}_{\mu_1}^{\mu_2}$ has marginals $P_1 = \pi_1 P, P_2 = \pi_2 P$ with densities $f_1 \geq h_1$ and $f_2 \leq h_2$ with respect to μ. We assume first that $1 = \mu_1(\mathbb{R}^1) \leq \mu_2(\mathbb{R}^1)$, i.e., μ_1 is a probability measure and so $f_1 = h_1$.

PROPOSITION 11. *Define* $z_0 = \inf\{y : \int_{(y,\infty)} h_2 d\mu \leq 1\}$,

$$f_2^*(y) = \begin{cases} h_2(y) & \text{if } y > z_0 \\ \dfrac{1 - \displaystyle\int_{(z_0,\infty)} h_2(u)du}{\mu(z_0)} & \text{if } y = z_0 \text{ and } \mu\{z_0\} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.6}$$

and P^* the corresponding probability measure with μ-density f_2^*.

(a) $\sup\{\bar{F}_P(x,y); P \in \mathcal{P}_{\mu_1}^{\mu_2}\} = 1 - \max(F_{\mu_1}(x), F_{P^*}(y))$, *for all* x, y, *where* $\bar{F}_P(x,y) = P([x,\infty) \times [y,\infty))$ *is the survival function.*

(b) *The* sup *in* (a) *is attained for the distribution* $F^* = F_{X^*,Y^*}$, *where* $X^* = F_{\mu_1}^{-1}(U)$, $Y^* = F_{P^*}^{-1}(U)$.

(c) *If* c *is a cost function, which is componentwise antitone and satisfies the Monge condition, then*

$$\inf\left\{\int c(x,y)dF_P(x,y); P \in \mathcal{P}_{\mu_1}^{\mu_2}\right\} = \int c(x,y)dF^*(x,y). \tag{6.7}$$

Proof. (a), (b) For $P \in \mathcal{P}_{\mu_1}^{\mu_2}$ with marginals F_{μ_1}, G_2 we know that $\bar{F}_P(x,y) \le P(F_{\mu_1}^{-1}(U) \ge x, G_2^{-1}(U) \ge y) = P(U \ge \max(F_{\mu_1}(x), G_2(y))) = 1 - \max(F_{\mu_1}(x), G_2(y))$. By our construction of P^* we see that $F_{P^*}(y) \le G_2(y)$, for all y, and therefore, $\bar{F}_P(x,y) \le 1 - \max(F_{\mu_1}(x), F_{P^*}(y))$.

(c) The conditions on the cost function c were considered in [11]. In that terminology $-c$ is a Δ-monotone function. Therefore, (c) follows from (a), (b), and [11]. □

The "antitone" assumption in (c) of Proposition 11 does not have a good interpretation in terms of costs. Under some additional assumptions on the bounding measures we can construct solutions for more natural cost functions. Let again μ_1 have densities h_i with respect to μ, $1 = \mu_1(\mathbb{R}^1) \le \mu_2(\mathbb{R}^1)$.

THEOREM 12. *Assume that for some* $y_0 \in \mathbb{R}_1$ *we have*

$$h_1(u) \le h_2(u) \quad \text{for } u < y_0 \text{ and } h_1(u) \ge h_2(u) \text{ for } u \ge y_0. \tag{6.8}$$

Define $x_0 = \inf\{y : \int_{(y,\infty)} h_1(u)d\mu(u) \ge \int_{(y,\infty)} h_2(u)d\mu(u)\}$ *and define*

$$f_2(u) := \begin{cases} h_2(u) & \text{if } u > x_0 \\ \dfrac{\int_{[x,\infty)} h_1(u)d\mu(u) - \int_{(x_0,\infty)} h_2(u)d\mu(u)}{\mu(x_0)} & \text{if } u = x_0 \text{ and } \mu\{x_0\} > 0, \\ h_1(u) & \text{if } u < x_0. \end{cases} \tag{6.9}$$

Then for any cost function c *satisfying the Monge condition* (1.3) *and the unimodality condition* (2.16) *holds*:

$$\inf\left\{\int c(x,y)dF_P(x,y); P \in \mathcal{P}_{\mu_1}^{\mu_2}\right\} = \int_0^1 c(F_{\mu_1}^{-1}(u), F_2^{-1}(u))du, \tag{6.10}$$

where F_2 *is the df of the measure with density* f_2 *with respect to* μ. *The optimal distribution is induced by the rv's* $X^* = F_{\mu_1}^{-1}(U), Y^* = F_2^{-1}(U)$.

Proof. For any $P \in \mathcal{P}_{\mu_1}^{\mu_2}$ with marginals F_{μ_1}, G_2, we have by the Monge condition: $\int c(x,y)dF_P(x,y) \ge \int_0^1 c(F_{\mu_1}^{-1}(u), G_2^{-1}(u))du$. By our construction of F_2 we find that

$$\begin{aligned} &G_2(y) \ge F_2(y) \ge F_{\mu_1}(y) \quad \text{for all } y \ge x_0 \quad \text{and} \\ &F_2(y) = F_{\mu_1}(y) \quad \text{for all } y \le x_0; \end{aligned} \tag{6.11}$$

(6.11) implies that $F_{\mu_1}^{-1}(u) \ge F_2^{-1}(u) \ge G_2^{-1}(u)$ for $u > F_2(x_0)$ and $F_2^{-1}(u) = F_{\mu_1}^{-1}(u)$ for $u \le F_2(x_0)$. Our assumptions on c imply that $c(F_{\mu_1}^{-1}(u), G_2^{-1}(u)) \ge c(F_{\mu_1}^{-1}(u), F_2^{-1}(u))$ for all u. □

Remark 7. It is not difficult to extend the solution of Theorem 12 to the case $\mu_1(\mathbb{R}^1) < 1$ and the conditions $f_1 \geq h_1, f_2 \leq h_2$, for the densities of an admissible plan P, if we still have assumption (6.8). Again choose x_0 as in (6.9) and define

$$f_2(x) = \begin{cases} h_2(x), x > z_0, \\ \dfrac{1 - \int_{(z_0,\infty)} h_2(x)d\mu(x)}{\mu(z_0)} & \text{if } x = z_0 \text{ and } \mu(z_0) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{6.12}$$

where $z_0 = \inf\{y : \int_{(y,\infty)} h_2(x)d\mu(x) \leq 1\}$. Define $y_0 = \inf\{y : \int_{(y,\infty)} h_2(x)d\mu(x) \leq \int_{(y,\infty)} h_1(x)d\mu(x)\}$,

$$f_1(x) = \begin{cases} h_1(x) & \text{if } x > y_0, \\ f_2(x) & \text{if } x < y_0, \\ \dfrac{\int_{[y_0,\infty)} (h_2(x) - h_1(x))d\mu(x)}{\mu(y_0)} & \text{if } \mu(y_0) > 0. \end{cases} \tag{6.13}$$

Then we have for c as in Theorem 12

$$\inf\left\{\int c(x,y)dF_P(x,y); \pi_1 P \geq \mu_1, \pi_2 P \leq \mu_2\right\} = \int_0^1 c(F_1^{-1}(u), F_2^{-1}(u))du, \tag{6.14}$$

where F_i have densities f_i with respect to $\mu, i = 1, 2$. □

Acknowledgments. We are indebted to the referees for several valuable suggestions, among them the relevance of the transportation problem with fixed sum of marginals as considered in §3 for the "fractional b-matching problem" in the context of undirected graphs (cf. Neuhauser, G. L. and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization*, Wiley). The authors are grateful to Michel Balinski for many helpful discussions.

REFERENCES

[1] E. J. Anderson and P. Nash, *Linear Programming in Infinite Dimensional Spaces. Theory and Applications*, Wiley, New York, 1987.
[2] E. R. Barnes and A. J. Hoffman, *On transportation problems with upper bounds on leading rectangles*, SIAM J. Discr. Meth., 6 (1985), pp. 487–496.
[3] M. S. Bazaraa and J. J. Javis, *Linear Programming and Network Flows*, Wiley, New York, 1977.
[4] S. Cambanis, G. Simons, and W. Stout, *Inequalities for $Ek(X, Y)$ when the marginals are fixed*, Z. Wahrsch. verw. Gebiete 36 (1976), pp. 285–294.
[5] G. Dall'Aglio, *Fréchet classes and compatibility of distribution functions*. Sympos. Math., 9 (1972), pp. 131–150.
[6] A. Hoffman and A. Veinott, Jr., *Staircase transportation problems with superadditive rewards and cumulative capacities*, Math. Programming, 1992, to appear.
[7] I. Olkin and S. T. Rachev, *Marginal problems with additional constraints*, Tech. report 270, Dept. of Statistics, Stanford Univ., Stanford, CA, 1990.
[8] S. T. Rachev, *The Monge-Kantorovich mass transference problem and its stochastic applications*, Theory Probab. Appl., 29 (1984), pp. 647–676.
[9] ———, *Extreme functionals in the space of probability measures*, LNM 1155, 320–348, Springer-Verlag, 1985.
[10] S. T. Rachev and R. M. Shortt, *Duality theorems for Kantorovich-Rubinstein and Wasserstein functionals*, Dissertationes Math. CCXCIX (1990).
[11] L. Rüschendorf, *Inequalities for the expectation of Δ-monotone functions*, Z. Wahrsch. verw. Gebiete, 54 (1980), pp. 341–349.
[12] A. H. Tchen, *Inequalities for distributions with given marginals*, Ann. Probab., 8 (1980), pp. 814–827.
[13] W. Whitt, *Bivariate distributions with given marginals*, Ann. Statist., 4 (1976), pp. 1280–1289.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 690–727, May 1994

006

THE FREE BOUNDARY OF THE MONOTONE FOLLOWER*

MARIA B. CHIAROLLA† AND ULRICH G. HAUSSMANN†

Abstract. This paper identifies the free boundary arising in the two-dimensional *monotone follower, cheap control problem.* It proves that if a region of inaction $\mathcal{A}$ is of locally finite perimeter (LFP), then $\mathcal{A}$ can be replaced by a new region of inaction $\tilde{\mathcal{A}}$ whose boundary is locally C^1 (up to sets of lower dimension). It then gives conditions under which the hypothesis (LFP) holds. Furthermore, under these conditions even higher regularity of the free boundary is obtained, namely $C^{2,\alpha}$, except perhaps at a single corner point.

Key words. monotone stochastic control, variational inequality, free boundary, locally finite perimeter, measure theoretic boundary

AMS subject classifications. 93E20, 49L99

1. Introduction.

The *monotone follower problem* is a stochastic control problem in which the state, a diffusion process, is controlled by a monotone, nondecreasing process. The one-dimensional case has been studied by several authors, e.g., [9] where the diffusion reduces to Brownian motion and [5] where the diffusion has affine drift and diffusion coefficient independent of the state of the motion. It has been shown that the optimal control is singular with respect to Lebesgue measure as a function of time and is characterized by a region of inaction $\mathcal{A}$ and its complement; in this case the "free" boundary $\partial\mathcal{A}$, i.e., the boundary of $\mathcal{A}$, reduces to a point. When the state is outside $\mathcal{A}$, an optimal control makes it jump to $\partial\mathcal{A}$; when the state is in $\mathcal{A}$, the control is inactive. At $\partial\mathcal{A}$ an optimal control acts like the "local time" of the state process at $\partial\mathcal{A}$, as it forces the process to stay inside $\mathcal{A}$ with an instantaneous action at the boundary.

In more than one dimension several problems arise; one of these is the question of smoothness of the free boundary. For example, in [13] the regularity of the boundary is crucial for the construction of the optimal control process that is obtained as the solution of a Skorokhod problem (see [10]). In [15] the regularity of $\partial\mathcal{A}$ is studied in the n-dimensional case, but strong hypotheses on the data are needed in order to differentiate the Bellman equation to obtain n obstacle problems, the union of whose coincidence sets is the region of action $\mathcal{A}^c$. Then C^1-regularity in a neighborhood of any point where the coincidence set has positive Lebesgue density follows from a result of Caffarelli [2].

In this paper we identify under mild conditions the free boundary $\partial\mathcal{A}$ of the two-dimensional *monotone follower problem* in the *cheap control* case (i.e., when the cost functional does not depend explicitly on the control). Then we study its regularity. In particular, we recover the result of [15] in the two-dimensional, cheap case. Our results are restricted to two dimensions for technical reasons, but work on the n-dimensional, non-cheap case is in progress. In a companion paper [4] we construct an optimal control that acts in the manner described above for the one-dimensional case.

In §2 we state the control problem and briefly recall some of the properties of the value function that will be used in the remaining sections. In §3 we identify a region of inaction $\mathcal{A}$ and the three regions into which the region of action $(\mathcal{A})^c$ splits. We also characterize $\partial\mathcal{A}$ as the union of the "graphs" of two functions. Then, in §4, we assume that $\mathcal{A}$ is of locally finite perimeter (LFP) and we show that $\mathcal{A}$ can be replaced by a new region of inaction $\tilde{\mathcal{A}}$ with boundary $\partial\tilde{\mathcal{A}}$ which is countably 1-rectifiable. In §5 we give conditions under which the local finiteness of the perimeter (i.e., (LFP)) of $\mathcal{A}$ holds. In §6 we upgrade the regularity of

* Received by the editors June 15, 1992; accepted for publication (in revised form) December 9, 1992.
† Mathematics Department, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z2.

the boundary as well as that of the value function under the conditions of §5. Finally in §7 we discuss the monotonicity of the boundary of $\mathcal{A}$. An Appendix concludes the paper.

It should be pointed out that when this work was in progress, only an earlier version of [15] was available to the authors; it did not include some of the results of the new version. The motivation for our §6 was to complete the results of that earlier version of [15].

These results form a portion of the Ph.D. dissertation of the first author.

2. Statement of the problem. Let the state X_t of a two-dimensional system be governed by the Itô equation

$$\begin{cases} dX_t = g\,dt + \sigma \cdot dW_t + dk_t, \\ X_0 = x + k_0, \end{cases} \tag{2.1}$$

where $x, g \in \mathbb{R}^2, \sigma$ is a constant 2×2 matrix, W_t is a standard two-dimensional Brownian motion on an underlying probability space $\{\Omega, \mathcal{F}, P\}$ endowed with a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ satisfying the usual conditions, and the control $\{k_t\}_{t\geq 0}$ is a progressively measurable process with cadlag nondecreasing components k_t^1, k_t^2, such that $k_0^1 \geq 0, k_0^2 \geq 0$. The set of such processes is denoted by V_+. We write X_t^k for the state to denote the dependence on the control. The cost associated with each initial position $x \in \mathbb{R}^2$ and each control $k \in V_+$ is given by

$$J_x(k) = E\left\{\int_0^\infty f(X_t^k)e^{-\rho t}\,dt\right\}. \tag{2.2}$$

The value function is

$$\hat{u}(x) = \inf\{J_x(k) : k \in V_+\}. \tag{2.3}$$

Here $\rho > 0$ is a discount factor, and the cost rate f is strictly convex, non-negative and satisfies the following conditions:

$$f(x) \to +\infty \quad \text{as } |x| \to \infty; \tag{2.4}$$

there exist $p > 1$ and constants $0 < r \leq C_0, C_1, C_2$ such that for any $\lambda \in (0,1)$, any $x \in \mathbb{R}^2$ and any x' such that $|x'| \leq 1$:

$$r|x^+|^p - C_0 \leq f(x) \leq C_0(1 + |x|)^p; \tag{2.5}$$

$$|f(x) - f(x + x')| \leq C_1(1 + f(x) + f(x + x'))^{1-1/p}|x'|; \tag{2.6}$$

$$0 < f(x + \lambda x') + f(x - \lambda x') - 2f(x) \leq C_2\lambda^2(1 + f(x))^q; \tag{2.7}$$

where $q = (1 - 2/p)^+$ and $x^+ := (x_1^+, x_2^+)$ if $x = (x_1, x_2) \in \mathbb{R}^2$, with $x_i^+ := \max\{0, x_i\}$. An example is $f(x) = |x|^p$ with $p \in \mathbb{N}, p$ even. We set

- $B(x, r)$ is the open ball with center x and radius r.
- $\Lambda^* := \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0, x_2 \geq 0\}$.
- $C^{m,\alpha}(\Omega)$ is the set of m times continuously differentiable functions : $\Omega \mapsto \mathbb{R}$ whose mth order partial derivatives are locally Hölder continuous of order α, where Ω is open, m is a non-negative integer, and $\alpha \in [0, 1]$.
- G_p is the set of locally Lipschitz continuous functions v on $\mathbb{R}^2$ such that $|v(x)| \leq C(1 + |x|)^p$, and $|\nabla v(x)| \leq C(1 + |x|)^{p-1}$ for almost every x, for some constant C, with $p > 1$.

- $Au(x) := -\frac{1}{2}\,\mathrm{trace}[\sigma\sigma^* D^2u(x)] - g\cdot\nabla u(x) + \rho u(x)$ with D^2u the Hessian matrix of u.

We recall the following main properties of the value function $\hat{u}$ (cf. [3, Thms. 2.1 and 2.11, Prop. 2.2]).

THEOREM 2.1. *There exist constants $\hat{r}, \hat{C}_0, \hat{C}_1, \hat{C}_2$ such that for each $\lambda \in (0,1)$ and each x' with $|x'| \le 1$, the function $\hat{u}(x)$ satisfies* (2.8)–(2.10) *below*

$$\hat{r}|x^+|^p - \hat{C}_0 \le \hat{u}(x) \le \hat{C}_0(1+|x|)^p, \tag{2.8}$$

$$|\hat{u}(x) - \hat{u}(x+x')| \le \hat{C}_1(1+|x|+|x+x'|)^{p-1}|x'|, \tag{2.9}$$

$$0 \le \hat{u}(x+\lambda x') + \hat{u}(x-\lambda x') - 2\hat{u}(x) \le \hat{C}_2\lambda^2(1+|x|)^{(p-2)^+}. \tag{2.10}$$

Hence the optimal cost $\hat{u}$ is in $W^{2,\infty}_{\mathrm{loc}}(\mathbb{R}^2)$. In particular, there exists a version of $\hat{u}$ in $C^{1,1}(\mathbb{R}^2)$.

THEOREM 2.2. *The optimal cost $\hat{u}$ is the maximal solution of*

$$\begin{cases} Au \le f \quad \text{a.e. in } \mathbb{R}^2; \\ \dfrac{\partial u}{\partial x_1} \ge 0, \qquad \dfrac{\partial u}{\partial x_2} \ge 0 \quad \text{a.e. in } \mathbb{R}^2; \\ (Au - f)\dfrac{\partial u}{\partial x_1}\dfrac{\partial u}{\partial x_2} = 0 \quad \text{a.e. in } \mathbb{R}^2; \\ u \in G_p, \qquad \| D^2u \| \in L^\infty_{\mathrm{loc}}. \end{cases}$$

3. Identification of the free boundary. Motivated by the above theorem we say that an open set $\mathcal{A}$ is *a set of inaction* for $\hat{u}$ if

$$\begin{aligned} A\hat{u}(x_1,x_2) &= f(x_1,x_2) \quad \text{a.e. } (x_1,x_2) \in \mathcal{A} \\ \hat{u}_{x_1}(x_1,x_2)\hat{u}_{x_2}(x_1,x_2) &= 0 \quad \text{a.e. } (x_1,x_2) \in \mathcal{A}^c. \end{aligned}$$

The obvious set of inaction is

$$\mathcal{A} = \{(x_1,x_2) : \hat{u}_{x_1}(x_1,x_2) > 0, \hat{u}_{x_2}(x_1,x_2) > 0\}.$$

In this section we study $\mathcal{A}$, its complement and its boundary.

The following theorem is proved in the Appendix. (Notice that this result was already implicit in [11, Thm. 4.1], although there the proof is quite confusing and seems to contain several gaps.)

THEOREM 3.1. *There exist two functions $\psi_2(x_1)$ and $\psi_1(x_2)$ such that*

$$\forall\, x_1 \in \mathbb{R}: \quad \begin{cases} \hat{u}_{x_2}(x_1,x_2) = 0 & \forall\, x_2 \le \psi_2(x_1), \\ \hat{u}_{x_2}(x_1,x_2) > 0 & \forall\, x_2 > \psi_2(x_1); \end{cases} \tag{3.1}$$

$$\forall\, x_2 \in \mathbb{R}: \quad \begin{cases} \hat{u}_{x_1}(x_1,x_2) = 0 & \forall\, x_1 \le \psi_1(x_2), \\ \hat{u}_{x_1}(x_1,x_2) > 0 & \forall\, x_1 > \psi_1(x_2); \end{cases} \tag{3.2}$$

i.e., the functions $\psi_2(x_1)$ and $\psi_1(x_2)$ are defined by

$$\begin{cases} \psi_2(x_1) = \inf\{x_2 : \hat{u}_{x_2}(x_1,x_2) > 0\}, \\ \psi_1(x_2) = \inf\{x_1 : \hat{u}_{x_1}(x_1,x_2) > 0\}. \end{cases} \tag{3.3}$$

We remark that Menaldi and Robin (cf. [11, Thm. 4.1]) claimed that the functions ψ_i are nonincreasing; however, their proof is incorrect. We study the monotonicity of ψ_i in §7 and give a counter-example to the Menaldi–Robin claim.

We now define

$$\begin{cases} R_0 := \{(x_1, x_2) : x_1 \le \psi_1(x_2), x_2 \le \psi_2(x_1)\}, \\ R_1 := \{(x_1, x_2) : x_1 \le \psi_1(x_2), \psi_2(x_1) < x_2\}, \\ R_2 := \{(x_1, x_2) : \psi_1(x_2) < x_1, x_2 \le \psi_2(x_1)\}; \end{cases} \tag{3.4}$$

$$\partial := \partial(R_0 \cup R_1 \cup R_2); \tag{3.5}$$

$$\begin{cases} \partial_0 := \partial R_0 \cap \partial, \\ \partial_1 := \partial R_1 \cap \partial, \\ \partial_2 := \partial R_2 \cap \partial; \end{cases} \tag{3.6}$$

$$\begin{aligned} \mathcal{A} := (R_0 \cup R_1 \cup R_2)^c &= \{(x_1, x_2) : \psi_1(x_2) < x_1, \psi_2(x_1) < x_2\} \\ &= \{(x_1, x_2) : \hat{u}_{x_1}(x_1, x_2) > 0, \hat{u}_{x_2}(x_1, x_2) > 0\}. \end{aligned} \tag{3.7}$$

Clearly $\mathcal{A}$ is open since $\hat{u}_{x_1}$ and $\hat{u}_{x_2}$ are continuous; $(R_0 \cup R_1 \cup R_2)^c$ is the region of inaction (i.e., the region where $A\hat{u} = f$ holds almost everywhere), and $\partial_0 = \partial R_0 \cap \partial\mathcal{A}$. Note that $\mathcal{A} \neq \emptyset$ as this follows from Theorem 3.1 and (2.8). Since $f, \hat{u}, \hat{u}_{x_i}$ are continuous and the "dynamic programming equation" $A\hat{u} = f$ holds almost everywhere in $\mathcal{A}$, this last equality can be interpreted to hold everywhere in $\mathcal{A}$ and tr $[\sigma\sigma^* D^2 u]$ can be taken to be defined everywhere in $\mathcal{A}$ by continuity.

Note that from the definition of ψ_i, the convexity of $\hat{u}$ and the fact that $\hat{u}_{x_i} \ge 0$ follows

$$\begin{aligned} &\forall\, (\bar{x}_1, \bar{x}_2) \in R_1 : (-\infty, \bar{x}_1] \times \{\bar{x}_2\} \subset R_1, \\ &\forall\, (\bar{x}_1, \bar{x}_2) \in R_2 : \{\bar{x}_1\} \times (-\infty, \bar{x}_2] \subset R_2. \end{aligned}$$

LEMMA 3.2. *Let* $\alpha_0 = \inf\{\hat{u}(x_1, x_2) : (x_1, x_2) \in \mathbb{R}^2\}$. *Then*

(i) $R_0 = \{(x_1, x_2) : \nabla\hat{u}(x_1, x_2) = 0\} = \{(x_1, x_2) : \hat{u}(x_1, x_2) = \alpha_0\}$;

(ii) *for every* $P \in R_0, (P - \Lambda^*) \subset R_0$;

(iii) *in* R_1 *one has* $\hat{u}_{x_2} =$ const *along horizontal line segments*;

(iv) *in* R_2 *one has* $\hat{u}_{x_1} =$ const *along vertical line segments*;

(v) *in* int(R_1) *one has* $\hat{u}_{x_2x_2} =$ const *along almost all horizontal line segments*;

(vi) *in* int(R_2) *one has* $\hat{u}_{x_1x_1} =$ const *along almost all vertical line segments*.

Proof. (i) From the definition of ψ_i follows $\hat{u}_{x_i} = 0$ in int(R_0). Now (i) follows by the continuity of $\hat{u}_{x_i}$ and the convexity of $\hat{u}$.

(ii) Let $P = (x_1, x_2) \in R_0$. Then $\hat{u}_{x_i} \ge 0\,(i = 1, 2)$ implies

$$\alpha_0 = \hat{u}(P) \ge \hat{u}(Q) \ge \alpha_0, \quad \forall\, Q \in P - \Lambda^*.$$

Hence $Q \in R_0$ by (i).

(iii) Let $\tilde{P} = (\tilde{x}_1, \tilde{x}_2) \in R_1$. Then $\hat{u} =$ const on $(-\infty, \tilde{x}_1] \times \{\tilde{x}_2\}$ (by the definition of ψ_1), so for every fixed $\delta > 0$, one has

$$\begin{cases} \hat{u}(\cdot\,, \tilde{x}_2 + \delta) - \hat{u}(\cdot\,, \tilde{x}_2) \uparrow & \text{in } (-\infty, \tilde{x}_1) \\ \hat{u}(\cdot\,, \tilde{x}_2) - \hat{u}(\cdot\,, \tilde{x}_2 - \delta) \downarrow & \text{in } (-\infty, \tilde{x}_1), \end{cases}$$

since $\hat{u}_{x_1} \ge 0$. It follows that $\hat{u}_{x_2}(\cdot\,, \tilde{x}_2)$ is constant on $(-\infty, \tilde{x}_1]$. Similarly (iv) follows.

The same arguments prove (v) and (vi) wherever $\hat{u}_{x_ix_i}$ exists. □

LEMMA 3.3. *The function ψ_i is upper semicontinuous* (u.s.c.), $i = 1, 2$.

Proof. Let us recall that $\psi_1(z) = \inf\{x_1 : \hat{u}_{x_1}(x_1, z) > 0\}$ is defined for every $z \in \mathbb{R}$ and is finite (cf. Theorem 3.1). Let $z \in \mathbb{R}, \varepsilon > 0$, then

$$\psi_1(z) - \psi_1(y) > \bar{x} - \varepsilon - \psi_1(y)$$

for some $\bar{x}$ such that $\hat{u}_{x_1}(\bar{x}, z) > 0$, and this holds for any $y \in \mathbb{R}$. Now from the continuity of $\hat{u}_{x_1}$ follows $\hat{u}_{x_1}(\bar{x}, y) > 0$ if $|y - z| < \delta$, for some $\delta > 0$. Therefore,

$$\psi_1(z) - \psi_1(y) > -\varepsilon \quad \text{if } |y - z| < \delta,$$

i.e., ψ_1 is u.s.c. □

Note that Lemma 3.3 implies that $R_0 \cup R_i$ is closed $(i = 1, 2)$.

For a function $h : \mathbb{R}^2 \to \mathbb{R}$ and a point $P_0 \in \mathbb{R}^2$, we set

$$h(P_0-) := \lim_{\substack{P \to P_0 \\ P \in R_0}} h(P), \qquad h(P_0+) := \lim_{\substack{P \to P_0 \\ P \in \mathcal{A}}} h(P), \tag{3.8}$$

if these limits exist. Then we have the following proposition.

PROPOSITION 3.4. (i) $\partial_0 \subset \{(x_1, x_2) : f(x_1, x_2) = \rho\alpha_0\} \cap \{(x_1, x_2) : \hat{u}(x_1, x_2) = \alpha_0\}$;

(ii) $\rho\alpha_0 = A\hat{u}(P_0-) = A\hat{u}(P_0+) = f(P_0)$ *for every* $P_0 \in \partial_0$;

(iii) $\emptyset \neq \partial(R_0 \cap R_1) \cap \partial(R_0 \cup R_2) \subset \partial_0$;

(iv) ∂_0 *is a singleton.*

Proof. Let $P_0 \in \partial_0$, then $P_0 \in R_0$ and so $\hat{u}(P_0) = \alpha_0$, according to Lemma 3.2. Moreover, since $\hat{u}$ is constant in R_0 and $\hat{u}_{x_1}, \hat{u}_{x_2}$ are continuous, we have

$$\rho\alpha_0 = \rho\hat{u}(P_0) = A\hat{u}(P_0-) \leq f(P_0).$$

On the other hand, since $\hat{u}$ is convex, $\operatorname{tr}[\sigma\sigma^* D^2\hat{u}(P_0+)] \geq 0$; therefore,

$$f(P_0) = A\hat{u}(P_0+) = -\frac{1}{2}\operatorname{tr}\left[\sigma\sigma^* D^2\hat{u}(P_0+)\right] + \rho\hat{u}(P_0) \leq \rho\hat{u}(P_0).$$

Thus,

$$A\hat{u}(P_0+) = f(P_0) = \rho\hat{u}(P_0) = \rho\alpha_0 = A\hat{u}(P_0-).$$

Now we show that $\partial(R_0 \cup R_1)$ and $\partial(R_0 \cup R_2)$ intersect. Suppose not, then (since ψ_i is a function of $x_j, i \neq j$) there are three cases (see Fig. 1):

(a) There exists $P_1 = (x_1, x_2) \in \operatorname{graph}(\psi_1)$ such that
  i) there exists $P_2 = (y_1, y_2) \in \partial(R_0 \cup R_2) \cap ((-\infty, x_1] \times \{x_2\})$,
  ii) there exists $z \leq x_1$ such that $\psi_2(z) < x_2$.

(b) There exists $Q_1 = (x_1, x_2) \in \operatorname{graph}(\psi_2)$ such that
  i) there exists $Q_2 = (y_1, y_2) \in \partial(R_0 \cup R_1) \cap (\{x_1\} \times (-\infty, x_2])$,
  ii) there exists $z \leq x_2$ such that $\psi_1(z) < x_1$.

(c) i) $\partial(R_0 \cup R_2) \cap \operatorname{subgraph}(\psi_1) = \emptyset$,
  ii) $\partial(R_0 \cup R_1) \cap \operatorname{subgraph}(\psi_2) = \emptyset$.

Let us assume that (a) holds. Then, $\nabla\hat{u}(P_2) = 0$ by (3.4) (recall that $R_0 \cup R_i$ are closed); so $P_2 \in R_0$. Also, $\hat{u}(P_2) = \hat{u}(P_1)$ (by the definition of ψ_1) since $y_1 < x_1 \leq \psi_1(x_2)$. Therefore, we have $\hat{u}(P_1) = \alpha_0$ (cf. Lemma 3.2), i.e., $P_1 \in R_0$. Hence $(z, x_2) \in R_0$ since $z \leq x_1$, and thus $\hat{u}(z, x_2) = \alpha_0$. But this is impossible because $\hat{u}$ is no longer constant above the graph of ψ_2, by the definition of ψ_2. So (a) cannot occur.

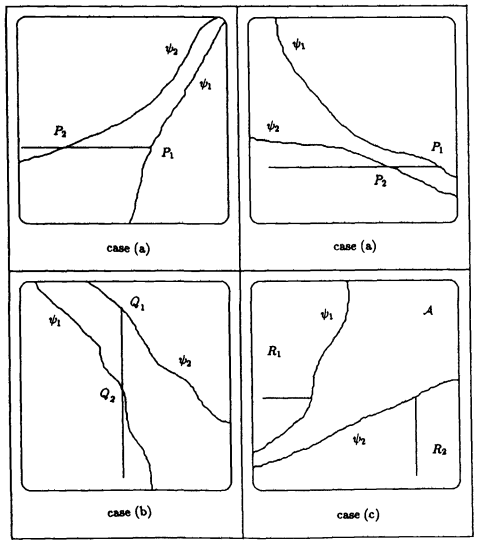


FIG. 1. *Possible cases* (a), (b), *and* (c) *occurring in the proof of Proposition* 3.4, *taking* ψ_1, ψ_2 *continuous.*

Similarly we can show that (b) cannot occur. Therefore $\partial(R_0 \cup R_1)$ must be to the left and above $\partial(R_0 \cup R_2)$, i.e., (c) must hold. So, since we are assuming $\partial(R_0 \cup R_1) \cap \partial(R_0 \cup R_2) = \emptyset$, there must exist a sequence $\{(x_n^1, x_n^2) : n \in \mathbb{N}\}$ of points of graph(ψ_1) such that $x_n^2 \to -\infty, x_n^1 \to -\infty$, and $x_n^2 > \psi_2(x_n^1)$; i.e., for each $n \in \mathbb{N}$,

$$\begin{cases} \hat{u}_{x_1}(x_n^1, x_n^2) & = 0, \\ \hat{u}_{x_2}(x_n^1, x_n^2) & > 0, \\ |(x_n^1, x_n^2)| \to +\infty & \text{as } n \to \infty. \end{cases}$$

From $A\hat{u} = f$ almost everywhere in $\mathcal{A}$ and tr $[\sigma\sigma^* D^2\hat{u}] \geq 0$ follows

$$-g \cdot \nabla\hat{u} + \rho\hat{u} \geq f \quad \text{in } \mathcal{A}$$

(recall that $\hat{u}, f, \hat{u}_{x_1}, \hat{u}_{x_2}$ are continuous). Hence, by continuity, the same is true on graph(ψ_1), and so we have

$$g_2^+ \hat{u}_{x_2}(x_n^1, x_n^2) + \rho\hat{u}(x_n^1, x_n^2) \geq f(x_n^1, x_n^2), \quad \text{for } n \in \mathbb{N}.$$

Now we recall that $f(x^1, x^2) \to +\infty$ as $|(x^1, x^2)| \to \infty$ and so we get

$$g_2^+ \hat{u}_{x_2}(x_n^1, x_n^2) + \rho\hat{u}(x_n^1, x_n^2) \to +\infty \quad \text{as } n \to +\infty. \tag{3.9}$$

However, $\hat{u}(x_n^1, x_n^2) \to +\infty$ as $n \to \infty$ would imply $\hat{u} \equiv +\infty$ since $\hat{u}$ is convex (and $\hat{u}_{x_1} \geq 0, \hat{u}_{x_2} \geq 0$) and since $x_n^2 \to -\infty$. Then $\hat{u}(x_n^1, x_n^2) \to +\infty$ is impossible. Similarly, $\hat{u}_{x_2}(x_n^1, x_n^2) \to +\infty$ as $n \to \infty$ cannot hold for that would imply $\hat{u}_{x_2} \equiv +\infty$ on graph(ψ_1); in fact, $\hat{u}_{x_2}(x_n^1, \cdot) \uparrow$ (by convexity) and $\hat{u}_{x_2} =$ const along horizontal line segments to the left of graph(ψ_1) (by Lemma 3.2(iii)) imply $\hat{u}_{x_2}(\psi_1(\cdot), \cdot) \uparrow$. Clearly $\hat{u}_{x_2} \equiv +\infty$ on graph(ψ_1) contradicts the polynomial growth of $\hat{u}_{x_2}$. Hence (3.9) is false and we must conclude that $\partial(R_0 \cup R_1) \cap \partial(R_0 \cup R_2) \neq \emptyset$.

Now let $P = (x_1, x_2) \in \partial(R_0 \cup R_1) \cap \partial(R_0 \cup R_2)$; thus $P \in \partial(R_0 \cup R_1 \cup R_2)$ and $P \in \partial(R_0)$. Hence $P \in \partial_0$ (cf. (3.6)).

Finally let $B = \{P : f(P) < \rho\alpha_0\}$. If $B = \emptyset$, then $\partial_0 = \operatorname{argmin} f$ and hence is a singleton. Otherwise B is a strictly convex set such that $\partial B = \{P : f(P) = \rho\alpha_0\}$; $R_0 = \{P : \hat{u}(P) = \alpha_0\}$ (cf. Lemma 3.2(i)) is a convex set and so is $\operatorname{int}(R_0)$. Also, in $\operatorname{int}(R_0)$ one has $\rho\alpha_0 = \rho\hat{u} = A\hat{u} \le f$, which implies $B \cap \operatorname{int}(R_0) = \emptyset$. Then, there exists a hyperplane separating B and $\operatorname{int}(R_0)$; hence $\partial B \cap R_0$ is at most a singleton (since B is strictly convex). Now (iv) follows from (i) and (iii). □

Remark 3.5. We only use the strict convexity of f to conclude that ∂_0 is a singleton. If we only assume f to be convex, then the above argument shows that ∂_0 is a line segment if $B \neq \emptyset$. If $B = \emptyset$, then ∂_0 is part of the boundary of the convex set $\operatorname{argmin} f \cap R_0$ and can be replaced by the line segment joining its end points. This gives a new R_0 and $\mathcal{A}$ but changes little else.

PROPOSITION 3.6. *If* $x_1^0 := \lim_{x_2 \to -\infty} \psi_1(x_2)$ *and* $x_2^0 := \lim_{x_1 \to -\infty} \psi_2(x_1)$, *then*

(i) x_1^0 *and* x_2^0 *exist and are finite*;

(ii) $\partial_0 = \{P_0\} := \{(x_1^0, x_2^0)\}$ *and* $R_0 = \partial_0 - \Lambda^*$;

(iii) ψ_i *is constant on* $(-\infty, x_j^0]$, $i \neq j, i, j = 1, 2$.

Proof. Let $\bar{x} \in \partial_1 \backslash \partial_0$; then $\bar{x}_1 = \psi_1(\bar{x}_2)$, $\bar{x}_2 > \psi_2(\bar{x}_1)$. Consider the line segment $S = \{(x_1, x_2) : x_1 \le \bar{x}_1, x_2 = \bar{x}_2\}$. Then $\hat{u}$ is constant on S (since $\hat{u}_{x_1} = 0$ on S by the definition of ψ_1). If $S \cap R_0 \neq \emptyset$, then $\hat{u} = \alpha_0$ on S (cf. Lemma 3.2) and hence $\hat{u} = \alpha_0$ on $(\bar{x}_1, \bar{x}_2) - \Lambda^*$ (since $\hat{u}_{x_i} \ge 0$). Hence $\hat{u} = \alpha_0$ on an open set in R_1, which is impossible by the definition of ψ_2. Therefore, $S \cap R_0 = \emptyset$, i.e., R_0 has to lie below any horizontal line through $\partial_1 \backslash \partial_0$. Similarly it must lie to the left of any vertical line through $\partial_2 \backslash \partial_0$. Since ∂_1 and ∂_2 meet at ∂_0, we have $R_0 \subset \partial_0 - \Lambda^*$. Now the assertions (i) and (ii) follow from the definition of R_0, Lemma 3.2, and Proposition 3.4 (iv).

To prove (iii), it suffices to observe that $\psi_i(x_j) = \psi_i(x_j^0)$ on $(-\infty, x_j^0]$, $i \neq j$, since $R_0 = \partial_0 - \Lambda^*$ by (ii). □

LEMMA 3.7. *The function* ψ_i *is locally bounded* $(i = 1, 2)$.

Proof. Fix $i = 1$. Recall that ψ_1 is u.s.c. (cf. Lemma 3.3), so ψ_1 is bounded above on compacta. Hence it suffices to show that ψ_1 is bounded below on any compact set $K \subset [x_2^0, +\infty)$ (cf. Proposition 3.6). Assume not, then there exists $z_1 \in K$ such that $\underline{\lim}_{\substack{y \to z_1 \\ y \in K}} \psi_1(y) = -\infty$; so there exists a sequence $y_n \to z_1$ such that $y_n \in K$ and $\psi_1(y_n) \to -\infty$ as $n \to \infty$. Also, since $K \subset [x_2^0, +\infty)$ we have

$$\{(\psi_1(y_n), y_n) : n \in \mathbb{N}\} \subset \partial_1.$$

By Theorem 2.2, $A\hat{u} = f$ almost everywhere in $\mathcal{A}$, but $\operatorname{tr}[\sigma\sigma^* D^2\hat{u}] \ge 0$ almost everywhere (since $\hat{u}$ is convex); hence

$$\rho\hat{u} \ge f + g \cdot \nabla\hat{u} \quad \text{a.e. in } \mathcal{A},$$

and by continuity this inequality can be interpreted to hold everywhere in $\operatorname{cl}(\mathcal{A})$. In particular, one has

$$\rho\hat{u} - g_2\hat{u}_{x_2} \ge f \quad \text{on } \partial_1.$$

Therefore,

$$f(\psi_1(y_n), y_n) \le \rho\hat{u}(\psi_1(y_n), y_n) - g_2\hat{u}_{x_2}(\psi_1(y_n), y_n),$$

but then, in the limit as $n \to \infty$, the left-hand side (LHS) will diverge to $+\infty$ since $f(x_1, x_2) \to +\infty$ as $|(x_1, x_2)| \to +\infty$, while the right-hand side (RHS) will remain bounded since

$$\begin{cases} \hat{u}(\psi_1(y_n), y_n) \le \hat{u}(\psi_1(z_1), y_n) \to \hat{u}(\psi_1(z_1), z_1), & \\ \hat{u}_{x_2}(\psi_1(y_n), y_n) \le \hat{u}_{x_2}(\psi_1(y_n), z_1) = \hat{u}_{x_2}(\psi_1(z_1), z_1) & \text{if } y_n \uparrow z_1, \\ \hat{u}_{x_2}(\psi_1(y_n), y_n) \text{ is nonincreasing with } n & \text{if } y_n \downarrow z_1, \end{cases}$$

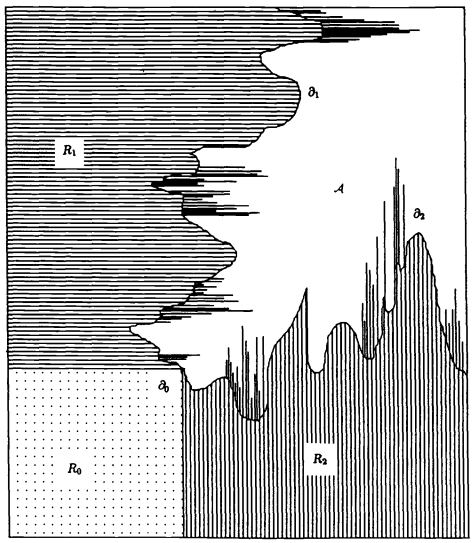


FIG. 2. *Sketch of the region of inaction $\mathcal{A}$ and its complement $R_0 \cup R_1 \cup R_2$.*

as follows from the fact that in R_1 both $\hat{u}$ and $\hat{u}_{x_2}$ are constant along horizontal line segments (cf. Lemma 3.2), and $\hat{u}_{x_1}, \hat{u}_{x_2} \geq 0$, a contradiction. (For ψ_2 the proof is the same.) □

LEMMA 3.8. *The function ψ_i is continuous at x_j^0, $i \neq j$ $(i, j = 1, 2)$. Hence, in particular,*

$$\emptyset \neq \text{graph}(\psi_1) \cap \text{graph}(\psi_2) = \partial_0 = \{P_0\}.$$

Proof. Let us fix $i = 1, j = 2$ for simplicity. By Lemma 3.3, ψ_1 is u.s.c.; therefore, if $\psi_1(x_2^0+) := \underline{\lim}_{z \downarrow x_2^0} \psi_1(z)$, we have $\psi_1(x_2^0+) \leq \psi_1(x_2^0)$. On the other hand, since $x_2^0 = \psi_2(x_1^0) = \psi_2(x_1)$ for any $x_1 \in (-\infty, x_1^0)$ (cf. Proposition 3.6), $\psi_1(x_2^0+) < \psi_1(x_2^0)$ would imply $[\psi_1(x_2^0+), \psi_1(x_2^0)] \times \{x_2^0\} \subset \partial_0$ (cf. (3.6) for the definition of ∂_0), but this is impossible since ∂_0 is a singleton (cf. Proposition 3.4). So $\psi_1(x_2^0+) = \psi_1(x_2^0)$; also, $\psi_1(x_2^0-) = \psi_1(x_2^0)$ by Proposition 3.6(iii). The last assertion of the lemma follows from Propositions 3.4 and 3.6. □

Hence we have established that the functions ψ_i cause the plane to split into the four regions R_0, R_1, R_2, and $\mathcal{A}$, as shown in Fig. 2.

Remark 3.9. We can show that $\nabla f(P_0)$ points into $-\Lambda^*$. This is the two-dimensional counterpart of the fact that in the one-dimensional case $f_x(x_0) \leq 0$ if $\{x_0\} = \partial\mathcal{A}$. To prove this remark observe that Theorem 2.2 and Lemma 3.2 imply $f(P) \geq A\hat{u}(P) = \rho\alpha_0$ for $P \in \text{int}(R_0)$. By continuity it follows that $f(P) \geq \rho\alpha_0$ for $P \in R_0$; but $f(P_0) = \rho\alpha_0$ by Proposition 3.4; hence P_0 minimizes f over R_0. It follows that $-\nabla f(P_0)$ (if it is not zero) is an outward normal to R_0 at P_0 and the result follows by Proposition 3.6(ii).

THEOREM 3.10. *For the free boundary $\partial\mathcal{A}$ one has*

(i) $\partial_0 = \{(x_1^0, x_2^0)\}; \nabla\hat{u}(x_1^0, x_2^0) = (0, 0)$.

(ii) $\partial_1 \cap \partial_2 = \partial_0$.

(iii) $\partial\mathcal{A} = \partial_1 \cup \partial_2$.

(iv) $\nabla\hat{u} = \begin{cases} (0, \hat{u}_{x_2}) & \text{on } \partial_1 \backslash \partial_0; \\ (\hat{u}_{x_1}, 0) & \text{on } \partial_2 \backslash \partial_0; \end{cases}$ *therefore* $\nabla\hat{u} \neq 0$ *on* $\partial\mathcal{A} \backslash \partial_0$.

Proof. Statements (i), (iii), and (iv) follow from the previous results.

(ii) Clearly $P \in \partial_1 \cap \partial_2$ implies $P \in R_0$, and so $P \in \partial_0$. On the other hand, it is obvious that $\partial_0 \subset \partial_1 \cap \partial_2$. □

Remark 3.11. Clearly we also have

$$\left.\begin{aligned} \hat{u}_{x_2x_1} &= 0 = \hat{u}_{x_1x_2} \\ \hat{u}_{x_1x_1} &= 0 \end{aligned}\right\} \quad \text{in int}(R_1),$$

and

$$\left.\begin{aligned} \hat{u}_{x_2x_1} &= 0 = \hat{u}_{x_1x_2} \\ \hat{u}_{x_2x_2} &= 0 \end{aligned}\right\} \quad \text{in int}(R_2).$$

Therefore for every x_i there exist functions $c_i(x_i), c_{ii}(x_i), i = 1, 2$, such that

$$A\hat{u}(x_1, x_2) = \begin{cases} -\frac{1}{2}a_{22}c_{22}(x_2) - g_2c_2(x_2) + \rho\hat{u}(x_1, x_2) & \text{a.e. in int}(R_1), \\ -\frac{1}{2}a_{11}c_{11}(x_1) - g_1c_1(x_1) + \rho\hat{u}(x_1, x_2) & \text{a.e. in int}(R_2), \end{cases}$$

and $\hat{u}(x_1, x_2)$ is constant in the variable x_1 in int(R_1) and in the variable x_2 in int(R_2).

4. Regularity of the free boundary. For the construction and approximation of the optimal control it is important to have some regularity of $\partial\mathcal{A}$. From the theory of sets of locally finite perimeter, it follows that the "measure theoretic boundary" of $\mathcal{A}$, a subset of $\partial\mathcal{A}$, is regular, cf. Theorem 4.8. Although we cannot determine what regularity, if any, the remaining part of $\partial\mathcal{A}$ possesses, we can modify $\mathcal{A}$ to a new set of inaction, $\tilde{\mathcal{A}}$, so that $\partial\tilde{\mathcal{A}}$ will have substantial regularity, cf. Theorem 4.32. We continue this discussion in Remark 4.9. We start with a few definitions (cf. [16, §5]).

DEFINITION 4.1. *For $\mathcal{O} \subset \mathbb{R}^2$ open, a function $u \in L^1(\mathcal{O})$ whose partial derivatives in the sense of distributions are measures with finite total variation in $\mathcal{O}$ is called a function of bounded variation, i.e., $u \in BV(\mathcal{O})$. If $u \in BV(\Omega)$ for every bounded open set Ω such that $cl(\Omega) \subset \mathcal{O}$, then we say that $u \in BV_{\text{loc}}(\mathcal{O})$.*

Thus, $u \in BV(\mathcal{O})$ if there exists a constant $C > 0$ such that

$$\left|\int_{\mathcal{O}} u(x)\frac{\partial\xi(x)}{\partial x_i}\,dx\right| \le C\,\|\xi\|_{L^\infty(\mathcal{O})} \quad \text{for } i = 1, 2 \text{ and all } \xi \in C_0^\infty(\mathcal{O}). \tag{4.1}$$

If $u \in BV(\mathcal{O})$, then its generalized gradient Du is a vector valued measure whose total variation is finite and given by

$$\|Du\|(\mathcal{O}) = \sup\left\{\sum_{i=1}^{2}\int_{\mathcal{O}} u(x)\frac{\partial\xi_i(x)}{\partial x_i}\,dx : \xi_i \in C_0^\infty(\mathcal{O}), \sum_{i=1}^{2}|\xi_i(x)|^2 \le 1, \text{ for } x \in \mathcal{O}\right\}. \tag{4.2}$$

Remark 4.2. Clearly, if $u \in C^1(\mathcal{O})$, then $\|Du\|(\mathcal{O}) = \int_{\mathcal{O}} |\nabla u|\,dx$ (this follows from (4.2) after an integration by parts). On the other hand, if $u \in W^{1,1}(\mathcal{O})$, then $\|Du\|(\mathcal{O}) = \int_{\mathcal{O}} |(u_{x_1}, u_{x_2})|\,dx$ where u_{x_i} is the weak derivative of u with respect to x_i. Therefore, $W^{1,1}(\mathcal{O}) \subset BV(\mathcal{O})$, but the two spaces are not equal (an example of a function in $BV(\mathcal{O})\setminus W^{1,1}(\mathcal{O})$ is given in Remark 4.5).

Finally, we point out that if $\mathcal{O} \subset \mathbb{R}$, then $u \in BV(\mathcal{O})$ (which is defined in a manner analogous to the two-dimensional case) has a more appealing characterization, namely (c) of the following remark.

Remark 4.3. If $\psi : \mathbb{R} \to \mathbb{R}$ is in $BV(\mathbb{R})$, there are other definitions equivalent to (the one-dimensional analogue of) Definition 4.1 (cf. [8, p. 26]), i.e., the following statements are equivalent:

(a) the derivative of ψ (in the distribution sense) is a finite measure;
(b) ψ can be approximated in L^1 by C^∞ functions with uniformly bounded variation;
(c) there exists a function $\tilde{\psi}$ such that $\tilde{\psi} = \psi$ almost everywhere and $\tilde{\psi}$ has bounded variation (in the usual sense) i.e., $V(\tilde{\psi}) = \sup\{V_a^b(\tilde{\psi}) : a < b, a, b \in \mathbb{R}\} < \infty$ where the variation of $\tilde{\psi}$ on $[a, b]$, $V_a^b(\tilde{\psi})$, is defined as

$$V_a^b(\tilde{\psi}) = \sup\left\{\sum_{i=1}^{m} |\tilde{\psi}(t_i) - \tilde{\psi}(t_{i-1})| : m \in \mathbb{N} \text{ and } a = t_0 < t_1 < \cdots < t_m = b\right\}. \tag{4.3}$$

Moreover, if ψ satisfies any of the conditions (a)–(c) above, then the total variation $\|D\psi\|(\mathbb{R})$ in the sense of (the one-dimensional analogue of) (4.2) is also given by

$$\|D\psi\|(\mathbb{R}) = \inf\{V(\tilde{\psi}) : \tilde{\psi} = \psi \text{ a.e.}\}. \tag{4.4}$$

DEFINITION 4.4. *A Borel set $E \subset \mathbb{R}^2$ is said to have* locally finite perimeter *if for every bounded open set $\Omega \subset \mathbb{R}^2$, the characteristic function of E, 1_E, is a function of bounded variation in Ω. Then the perimeter of E in Ω is defined as*

$$P(E, \Omega) = \|D1_E\|(\Omega) < \infty, \tag{4.5}$$

i.e.,

$$P(E, \Omega) := \sup\left\{\sum_{i=1}^{2} \int_E \frac{\partial \xi_i(x)}{\partial x_i}\, dx : \xi_i \in C_0^\infty(\Omega), \sum_{i=1}^{2} |\xi_i(x)|^2 \le 1, \text{ for } x \in \Omega\right\}. \tag{4.6}$$

Remark 4.5. If E is a bounded open set with C^2 boundary, then E is of finite perimeter and $P(E, \Omega)$ is the arc length of $\Omega \cap \partial E$ in the classical sense (cf. [16, Remark 5.4.2]) by an application of the Gauss–Green theorem. Hence $1_E \in BV(\Omega)$ but $1_E \notin W^{1,1}(\Omega)$!

We recall (cf. (3.4) and (3.7)) that

$$\mathcal{A} = \{\hat{u}_{x_1} > 0, \hat{u}_{x_2} > 0\}; \tag{4.7}$$

$$R_0 = \{\hat{u}_{x_1} = 0 = \hat{u}_{x_2}\}; \tag{4.8}$$

$$R_1 = \{\hat{u}_{x_1} = 0, \hat{u}_{x_2} > 0\}; \tag{4.9}$$

$$R_2 = \{\hat{u}_{x_1} > 0, \hat{u}_{x_2} = 0\}. \tag{4.10}$$

Then we make the following assumption.

(LFP). The sets $\{\hat{u}_{x_i} = 0\}, i = 1, 2$, are of locally finite perimeter.

Note that $\hat{u}_{x_i} \in BV_{\text{loc}}(\mathbb{R}^2)$ since

$$\hat{u}_{x_i} \in W^{1,1}_{\text{loc}}(\mathbb{R}^2); \tag{4.11}$$

hence almost all level sets of $\hat{u}_{x_i}$ are sets of locally finite perimeter (cf. [16, Thm. 5.4.4]).

DEFINITION 4.6. *If $E \subset \mathbb{R}^2$ is a Lebesgue measurable set, the* measure-theoretic boundary *of E is defined by*

$$\partial_M E = \{x : \bar{D}(E, x) > 0\} \cap \{x : \bar{D}(E^c, x) > 0\}, \tag{4.12}$$

where

$$\bar{D}(E, x) = \overline{\lim_{r \to 0}} \frac{|E \cap B(x, r)|}{|B(x, r)|}, \tag{4.13}$$

$|\cdot|$ being the Lebesgue measure in $\mathbb{R}^2$, and $B(x,r)$ being the open ball with center x and radius r. (If $\overline{\lim} = \underline{\lim}$ in (4.13), we denote their common value by $D(E,x)$.)

The $\partial_M E$ is a subset of the topological boundary ∂E and the points of ∂E where a tangent exists are in $\partial_M E$. If $E \subset \mathbb{R}^2$ is a set of locally finite perimeter, then its measure-theoretic boundary $\partial_M E$ is equal (up to a set of H^1 measure zero) to the reduced boundary of E, which is essentially the set of points of ∂E at which a measure-theoretic tangent exists. In this case a fundamental result of De Giorgi shows that the reduced boundary of E is a countably 1-rectifiable set (see the definition below), hence $\partial_M E$ is itself countably 1-rectifiable (cf. [16, Cor. 5.6.8, Lem. 5.9.5, and Thm. 5.7.3]).

DEFINITION 4.7. *A subset A of $\mathbb{R}^2$ is* countably 1-rectifiable *if*

$$A \subset \left(\bigcup_{i=1}^{\infty} f_j(A_j) \right) \cup A_0, \tag{4.14}$$

where $H^1(A_0) = 0$ and $f_j : A_j \to \mathbb{R}^2$, $A_j \subset \mathbb{R}$, is a countable collection of Lipschitz maps (H^1 being the one-dimensional Hausdorff measure on $\mathbb{R}^2$).

The following is an equivalent formulation of Definition 4.7 (cf. [16, Lem. 3.7.2]):

$$\begin{gathered} A \subset \mathbb{R}^2 \text{ is countably 1-rectifiable if} \\ A \subset \bigcup_{i=1}^{\infty} M_i \cup N, \end{gathered} \tag{4.15}$$

where $H^1(N) = 0$ and each M_i is a one-dimensional embedded C^1 submanifold of $\mathbb{R}^2$.

Since the union of two sets of locally finite perimeter is a set of locally finite perimeter ([8, Remark 1.7]) we obtain the following theorem.

THEOREM 4.8. *Assume* (LFP). *Then*

(i) *the region of inaction $\mathcal{A}$ is of locally finite perimeter;*

(ii) *$\partial_M \mathcal{A}$ is countably 1-rectifiable, i.e., $\partial_M \mathcal{A} \subset \bigcup_{i=1}^{\infty} M_i \cup N$ where*

- *$H^1(N) = 0$ (H^1 being the one-dimensional Hausdorff measure on $\mathbb{R}^2$),*
- *each M_i is a one-dimensional embedded C^1 submanifold of $\mathbb{R}^2$.*

Remark 4.9. We now know that $\partial_M \mathcal{A}$ is regular, so we turn to $\partial\mathcal{A} \backslash \partial_M \mathcal{A}$. As it is difficult to say much about this set, we show that it is possible to redefine $\mathcal{A}$ in order to obtain a new region of inaction $\tilde{\mathcal{A}}$ such that $\partial\tilde{\mathcal{A}}$ equals $\partial_M \tilde{\mathcal{A}}$ at least up to sets of lower dimension. More precisely, the difficulty lies in $\partial\mathcal{A}$ having possibly uncountably many "hairs" extending into $\mathcal{A}$ (there cannot be any extending into $\mathcal{A}^c$ because the ψ_i are u.s.c.), but we show that up to negligible sets the ψ_i have no discontinuities. We then "shave" these hairs off to obtain $\tilde{\mathcal{A}}$. In so doing we will not have lost any of $\partial_M \mathcal{A}$ and in fact $\partial\tilde{\mathcal{A}} \backslash \partial_M \mathcal{A}$ is negligible. Since all points of $\partial\mathcal{A}$ where a tangent vector exists are also points of the measure-theoretic boundary $\partial_M \mathcal{A}$, it is natural to start the study of $\partial\mathcal{A} \backslash \partial_M \mathcal{A}$ by examining those points of the boundary where the tangent fails to exist. Fortunately we have already obtained a parametric representation of $\mathcal{A}$ (cf. (4.7)) in terms of the functions ψ_1, ψ_2 so that the problem is now reduced to the study of the differentiability of ψ_1 and ψ_2. (LFP) and the local boundedness of ψ_i (cf. Lemma 3.7) will lead us to the existence of two functions, $\tilde{\psi}_1$ and $\tilde{\psi}_2$, differentiable almost everywhere and such that $\psi_i = \tilde{\psi}_i$ almost everywhere, $i = 1, 2$.

Throughout the remainder of this section we assume that (LFP) holds. Then, the set $R_0 \cup R_i$ $(i = 1, 2)$ is a set of locally finite perimeter (cf. (4.8)–(4.10)); hence the gradient of $1_{R_0 \cup R_i}$ is a vector valued measure whose total variation over any bounded open set $\mathcal{O} \subset \mathbb{R}^2$, $\|D1_{R_0 \cup R_i}\|(\mathcal{O})$, is finite (cf. (4.2) and (4.6)).

THEOREM 4.10. *Assume* (LFP), *then $\psi_i \in BV_{\text{loc}}(\mathbb{R})$ for $i = 1, 2$.*

Proof. Fix $i = 1$. Let Ω be a bounded open set in $\mathbb{R}$; then it follows from Lemma 3.7 that there exists $T > 0$ such that

$$(-\infty, -T) \times \Omega \subset ((R_0 \cup R_1) \cap (\mathbb{R} \times \Omega)) \subset (-\infty, T) \times \Omega;$$

therefore,

$$\|D1_{R_0 \cup R_1}\|(\mathbb{R} \times \Omega) = \|D1_{R_0 \cup R_1}\|((-T, T) \times \Omega),$$

and the RHS is finite since $R_0 \cup R_1$ is a set of locally finite perimeter. (Note that, in general, if E is a set of locally finite perimeter in $\mathbb{R}^2$, the total variation $\|D1_E\|(\mathbb{R} \times \Omega)$ need not be finite as $\mathbb{R} \times \Omega$ is not bounded.) Hence (LFP) together with $\|D1_{R_0 \cup R_1}\|(\mathbb{R} \times \Omega) < \infty$ allows us to apply Teorema 1.10 in [12, p. 525], and conclude $\psi_1 \in BV_{\text{loc}}(\mathbb{R})$. The proof for $i = 2$ is the same. □

The following result is crucial. Let Δ_ψ be the set of points of discontinuity of $\psi : \mathbb{R} \mapsto \mathbb{R}$, and denote by λ the Lebesgue measure on $\mathbb{R}$.

THEOREM 4.11. *If $\psi \in BV_{\text{loc}}(\mathbb{R})$ is upper semicontinuous, then $\lambda(\Delta_\psi) = 0$.*

Proof. It suffices to show that the set of points where $\psi|_{[a,b]}$ is discontinuous has measure zero for all finite intervals $[a, b]$. So let us fix $a < b$ and let $\psi \in BV([a, b])$; we again set

$$\Delta_\psi = \{x \in [a, b] : x \text{ is a point of discontinuity of } \psi\}. \tag{4.16}$$

Assume $\lambda(\Delta_\psi) > 0$. Also, $\psi \in BV([a, b])$ and Remark 4.3 (c) imply the existence of a function $\tilde{\psi}$ of bounded variation (i.e., $V_a^b(\tilde{\psi}) < \infty$) such that $\psi = \tilde{\psi}$ almost everywhere in $[a, b]$. Therefore, since $\tilde{\psi}$ may have at most countably many discontinuities,

$$\exists N \subset [a, b] \quad \text{such that} \begin{cases} \psi = \tilde{\psi} & \text{on } [a, b] \backslash N, \\ \tilde{\psi} \text{ is continuous} & \text{on } [a, b] \backslash N, \\ \lambda(N) = 0. \end{cases} \tag{4.17}$$

Thus, $\lambda(\Delta_\psi \backslash N) > 0$ and we set

$$\Delta'_\psi = \Delta_\psi \backslash N; \tag{4.18}$$

moreover, we may assume the elements of Δ'_ψ to be points of discontinuity of the second kind, since those of the first kind are at most countable (this is true for every real function) and hence may be assumed to be elements of N.

Let $y \in \Delta'_\psi$. Let $\varepsilon > 0$ be fixed and such that $\varepsilon < \psi(y) - \underline{\lim}_{z \to y} \psi(z)$. This is possible since ψ is u.s.c., and we are assuming that y is a point of discontinuity of the second kind, so $\psi(y) \geq \overline{\lim}_{z \to y} \psi(z) > \underline{\lim}_{z \to y} \psi(z)$. By continuity, since $y \in [a, b] \backslash N$, there exists $n_0 \in \mathbb{N}$ such that

$$|z - y| < 1/n_0, \qquad z \in [a, b] \backslash N \Rightarrow |\psi(z) - \psi(y)| < \varepsilon. \tag{4.19}$$

On the other hand, since $\psi(y) - \varepsilon > \underline{\lim}_{z \to y} \psi(z)$, it follows from (4.19) that

$$\exists x_0 \in N \cap B(y, 1/n_0) \quad \text{such that } \psi(x_0) < \psi(y) - \varepsilon. \tag{4.20}$$

Let $m_0 \in \mathbb{N}$ be such that $B(x_0, 1/m_0) \subset B(y, 1/n_0)$, then from $\lambda(N) = 0$ follows

$$\forall\, m \geq m_0 \quad \exists q_m \in ([a, b] \backslash N) \cap B(x_0, 1/m); \tag{4.21}$$

thus $\{q_m\} \subset [a, b] \backslash N$, $q_m \to x_0$ as $m \to \infty$, and we have

$$m \geq m_0 \Rightarrow q_m \in B(y, 1/n_0). \tag{4.22}$$

So (4.19) and (4.22) imply

$$m \geq m_0 \Rightarrow |\psi(q_m) - \psi(y)| < \varepsilon; \tag{4.23}$$

therefore,

$$\psi(y) - \varepsilon \leq \varliminf_{m\to\infty} \psi(q_m), \tag{4.24}$$

but $q_m \to x_0$ as $m \to \infty$ and ψ is u.s.c., so we must also have

$$\varliminf_{m\to\infty} \psi(q_m) \leq \overline{\lim_{z\to x_0}} \psi(z) \leq \psi(x_0). \tag{4.25}$$

Then (4.24) an (4.25) contradict (4.20) and the theorem is proved. □

Remark 4.12. The hypothesis that ψ is u.s.c. is absolutely crucial in Theorem 4.11, as shown by the simple example $\psi = 1_{\mathbb{Q}}$, where $\mathbb{Q}$ is the set of all rational numbers. In fact, $1_{\mathbb{Q}} \in BV(\mathbb{R})$ since $1_{\mathbb{Q}}$ is almost everywhere equal to the function of constant value zero, but $\Delta_{1_{\mathbb{Q}}} = \mathbb{R}$!

COROLLARY 4.13. *Assume* (LFP). *Then* $\lambda(\Delta_{\psi_i}) = 0, i = 1, 2$.

Proof. This follows trivially from Theorem 4.11, Theorem 4.10, and Lemma 3.3. □

COROLLARY 4.14. *Assume* (LFP). *Then the free boundary* $\partial\mathcal{A}$ *has two-dimensional Lebesgue measure zero, i.e.,*

$$|\partial\mathcal{A}| = |\partial_1 \cup \partial_2| = 0. \tag{4.26}$$

Proof. We have

$$\partial_1 \subset \{(x_1, x_2) \in \mathbb{R} \times [x_2^0, +\infty) : x_2 \in \Delta_{\psi_1}, x_1 \in [\varliminf_{z\to x_2} \psi_1(z), \psi_1(x_2)]\}$$
$$\cup \{(x_1, x_2) \in \mathbb{R} \times [x_2^0, +\infty) : x_2 \in (\Delta_{\psi_1})^c, x_1 = \psi_1(x_2)\};$$

therefore, Fubini's theorem implies $|\partial_1| = 0$ since $\lambda(\Delta_{\psi_1}) = 0$ by Corollary 4.13 and

$$\forall\, x_2 \in (\Delta_{\psi_1})^c \cap [x_2^0, +\infty) : \lambda(\{\psi_1(x_2)\}) = 0.$$

Similarly, $|\partial_2| = 0$ and (4.26) follows. □

Now we set

$$\tilde{\mathcal{A}} = \text{int}(\text{cl}(\mathcal{A})), \tag{4.27}$$

$$\tilde{\partial} = \partial\tilde{\mathcal{A}}, \tag{4.28}$$

$$\begin{cases} \tilde{\partial}_1 = \partial R_1 \cap \tilde{\partial}, \\ \tilde{\partial}_2 = \partial R_2 \cap \tilde{\partial}. \end{cases} \tag{4.29}$$

We need the following lemma whose proof is given in the Appendix.

LEMMA 4.15.

$$\begin{cases} \partial_i \backslash \tilde{\partial}_i = R_i \cap \tilde{\mathcal{A}}, & i = 1, 2, \\ \partial_0 \subset \tilde{\partial}. \end{cases}$$

COROLLARY 4.16. *Assume* (LFP). *Then the set* $\partial\mathcal{A}\backslash\partial\tilde{\mathcal{A}}$ *has empty intersection with* $\partial_M\mathcal{A}$, *the measure theoretic boundary of* $\mathcal{A}$, *i.e.,*

$$\partial_M\mathcal{A} \subset \partial\tilde{\mathcal{A}}. \tag{4.30}$$

Proof. If $P \in \partial_i \backslash \tilde{\partial}_i$, then $P \in \tilde{\mathcal{A}}$, an open set, so $B(P, r) \subset \tilde{\mathcal{A}}$ for r sufficiently small. Now the result follows immediately from (cf. Lemma 4.15)

$$|R_i \cap \tilde{\mathcal{A}}| = |\partial_i \backslash \tilde{\partial}_i| \leq |\partial_i| = 0$$

and the fact that all the elements of $\partial_M \mathcal{A}$ are points of positive Lebesgue density for both $\mathcal{A}$ and its complement (cf. Definition 4.6). □

We have shown that $\partial \backslash \tilde{\partial}$ is a subset of int(cl($\mathcal{A}$)) whose two-dimensional measure is zero. Moreover, it follows from Lemma 4.15 that

$$\tilde{\mathcal{A}} = \mathcal{A} \cup (\partial \backslash \tilde{\partial}); \tag{4.31}$$

hence $\tilde{\mathcal{A}}$ is also a region of inaction for $\hat{u}$.

Now we want to obtain a representation of $\tilde{\mathcal{A}}$ similar to the one provided by (3.7) for $\mathcal{A}$, i.e., we will show

$$\tilde{\mathcal{A}} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > \tilde{\psi}_1(x_2), x_2 > \tilde{\psi}_2(x_1)\}, \tag{4.32}$$

where $\tilde{\psi}_1, \tilde{\psi}_2$ can be selected to be u.s.c. (just as ψ_1, ψ_2 were u.s.c.) and of locally bounded variation in the usual sense (instead, ψ_1, ψ_2 were only elements of $BV_{\text{loc}}(\mathbb{R})$).

We start by showing that $\tilde{\psi}_1, \tilde{\psi}_2$ can be uniquely chosen to be u.s.c. among all the functions provided by Remark 4.3, (c).

We set

$$\text{bv}(\psi_i) = \{\phi : \mathbb{R} \to \mathbb{R} \mid \phi = \psi_i \text{ a.e.}, V_a^b(\phi) < \infty \, \forall \, a, b \in \mathbb{R}, a < b\} \tag{4.33}$$

where the total variation $V_a^b(\phi)$ is defined by (4.3), and $i = 1, 2$. Let

$$\{\text{u.s.c.}\} = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is upper semicontinuous}\}. \tag{4.34}$$

For any $\phi \in \text{bv}(\psi_i)$ we define $\bar{\phi} : \mathbb{R} \to \mathbb{R}$ by

$$\bar{\phi}(x) = \overline{\lim_{z \to x}} \, \phi(z), \qquad x \in \mathbb{R}. \tag{4.35}$$

LEMMA 4.17. *Assume* (LFP). *If* $\phi \in \text{bv}(\psi_i)$, *then* $\bar{\phi} \in \text{bv}(\psi_i) \cap \{\text{u.s.c.}\}$. *Moreover,*

$$\bar{\phi}(x) = \overline{\lim_{z \to x}} \, \bar{\phi}(z) \qquad x \in \mathbb{R}. \tag{4.36}$$

Proof. Surely $\text{bv}(\psi_i) \neq \emptyset$ as $\psi_i \in BV_{\text{loc}}(\mathbb{R})$ (cf. Remark 4.3). Let $\phi \in \text{bv}(\psi_i)$, then ϕ is continuous almost everywhere and so

$$\bar{\phi}(x) = \overline{\lim_{z \to x}} \, \phi(z) = \lim_{z \to x} \phi(z) = \phi(x) \quad \text{for a.e. } x \in \mathbb{R},$$

i.e.,

$$\bar{\phi}(x) = \phi(x) = \psi_i(x) \quad \text{for a.e. } x \in \mathbb{R}.$$

From (4.35) follows

$$\forall \varepsilon > 0 \quad \exists \delta_0 > 0 \quad \text{such that } 0 < |z - x| < \delta_0 \Rightarrow \phi(z) \leq \bar{\phi}(x) + \varepsilon;$$

then, if $z \in B(x, \delta_0)$ there exists δ_z such that $B(z, \delta_z) \subset B(x, \delta_0)$, and one has

$$0 < |y - z| < \delta_z \Rightarrow \phi(y) \leq \bar{\phi}(x) + \varepsilon;$$

i.e.,

$$\sup_{0<|y-z|<\delta_z} \phi(y) \le \bar{\phi}(x) + \varepsilon;$$

hence

$$\bar{\phi}(z) = \overline{\lim_{y\to z}}\, \phi(y) \le \bar{\phi}(x) + \varepsilon.$$

So we have

$$\forall\, \varepsilon > 0 \quad \exists \delta_0 > 0 \quad \text{such that } 0 < |z - x| < \delta_0 \Rightarrow \bar{\phi}(z) \le \bar{\phi}(x) + \varepsilon;$$

therefore

$$\overline{\lim_{z\to x}}\, \bar{\phi}(z) \le \bar{\phi}(x), \tag{4.37}$$

i.e., $\bar{\phi} \in \{\text{u.s.c.}\}$.

We now show that $\bar{\phi}$ is of locally bounded variation; in fact, if $a, b \in \mathbb{R}$ are points of continuity of ϕ such that $a < b$, then

$$V_a^b(\bar{\phi}) \le V_a^b(\phi). \tag{4.38}$$

Let $a, b \in \mathbb{R}$ be as above; let $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ such that $x_i - x_{i-1} = (b-a)/n$, $i = 1, 2, \ldots, n$. Also, let $\varepsilon > 0$, then from

$$\bar{\phi}(x_i) = \overline{\lim_{y\to x_i}}\, \phi(y)$$

follows

(i) there exists $\delta_i > 0$ such that $0 < |y - x_i| < \delta_i \Rightarrow \phi(y) \le \bar{\phi}(x_i) + \varepsilon/(2n)$, and with $\delta < \min\{\delta_1, \delta_2, \ldots, \delta_{n-1}, (b-a)/2n\}$,

(ii) there exist $z_1, z_2, \ldots, z_{n-1}$ such that $0 < |z_i - x_i| < \delta$ and $\phi(z_i) \ge \bar{\phi}(x_i) - \varepsilon/(2n)$.

So if we set $z_0 = a$, $z_n = b$ we obtain a new partition $a = z_0 < z_1 < \cdots < z_{n-1} < z_n = b$ of $[a, b]$ such that

$$|\bar{\phi}(x_i) - \phi(z_i)| \le \frac{\varepsilon}{2n} \quad \text{if } i = 1, 2, \ldots, n-1,$$

and

$$|\bar{\phi}(x_j) - \phi(x_j)| = |\phi(x_j) - \phi(x_j)| = 0 \quad \text{if } j = 0, n$$

since a, b are points of continuity of ϕ; hence $\bar{\phi}$ equals ϕ there. Then,

$$\begin{aligned}
&\sum_{i=1}^{n} |\bar{\phi}(x_i) - \bar{\phi}(x_{i-i})| \\
&\qquad \le \sum_{i=1}^{n} \{|\bar{\phi}(x_i) - \phi(z_i)| + |\phi(z_i) - \phi(z_{i-i})| + |\bar{\phi}(x_{i-1}) - \phi(z_{i-1})|\} \\
&\qquad \le \sum_{i=1}^{n} \left\{\frac{\varepsilon}{2n} + |\phi(z_i) - \phi(z_{i-1})| + \frac{\varepsilon}{2n}\right\} \\
&\qquad \le \varepsilon + \sum_{i=1}^{n} |\phi(z_i) - \phi(z_{i-1})| \\
&\qquad \le \varepsilon + V_a^b(\phi);
\end{aligned}$$

therefore, $V_a^b(\bar{\phi}) \le \varepsilon + V_a^b(\phi)$, and since $\varepsilon > 0$ is arbitrary (4.38) follows. Thus,

$$V_a^b(\bar{\phi}) < +\infty \quad \text{for every } a, b \in \mathbb{R}, a < b$$

(if a, b are points of discontinuity of ϕ, then we can always find $a' < a$ and $b' > b$ such that a', b' are points of continuity of ϕ, and we have $V_a^b(\bar{\phi}) \le V_{a'}^{b'}(\bar{\phi}) \le V_{a'}^{b'}(\phi) < \infty$).

It remains to show that the reverse inequality of (4.37) holds too so that (4.36) is verified. We observe that

$$\overline{\lim_{z\to y}}\, \phi(z) = \Big(\lim_{z\to y^+} \phi(z)\Big) \vee \Big(\lim_{z\to y^-} \phi(z)\Big)$$

as well as

$$\overline{\lim_{y\to x}}\, \bar{\phi}(z) = \Big(\lim_{y\to x^+} \bar{\phi}(y)\Big) \vee \Big(\lim_{y\to x^-} \bar{\phi}(y)\Big),$$

since ϕ and $\bar{\phi}$ are of locally bounded variation; hence they admit one-sided limits. Therefore,

$$\begin{aligned}
\overline{\lim_{y\to x}}\, \bar{\phi}(y) &= \Big(\lim_{y\to x^+} \bar{\phi}(y)\Big) \vee \Big(\lim_{y\to x^-} \bar{\phi}(y)\Big) \\
&\ge \Big(\lim_{y\to x^+} \Big(\lim_{z\to y^+} \phi(z)\Big)\Big) \vee \Big(\lim_{y\to x^-} \Big(\lim_{z\to y^-} \phi(z)\Big)\Big) \\
&= \Big(\lim_{z\to x^+} \phi(z)\Big) \vee \Big(\lim_{z\to x^-} \phi(z)\Big);
\end{aligned}$$

i.e.,

$$\overline{\lim_{y\to x}}\, \bar{\phi}(y) \ge \overline{\lim_{z\to x}}\, \phi(z) = \bar{\phi}(x). \tag{4.39}$$

Hence (4.37) and (4.39) imply (4.36), and the lemma is completely proved. □

PROPOSITION 4.18. *Assume* (LFP). *Then, the set*

$$\Big\{\phi \in \operatorname{bv}(\psi_i) \cap \{\text{u.s.c.}\} : \phi(x) = \overline{\lim_{z\to x}}\, \phi(z), x \in \mathbb{R}\Big\}$$

is a singleton $(i = 1, 2)$.

Proof. Assume not and let $\phi_1, \phi_2 \in \operatorname{bv}(\psi_i) \cap \{\text{u.s.c.}\}$ such that $\phi_j(x) = \overline{\lim}_{z\to x} \phi_j(z)$, $j = 1, 2$, and suppose there exists $y_0 \in \mathbb{R}$ such that $\phi_1(y_0) < \phi_2(y_0)$. Since $\phi_j \in \operatorname{bv}(\psi_i)$, ϕ_j admits one-sided limits and so one has

$$\phi_j(x) = \overline{\lim_{z\to x}}\, \phi_j(z) = \Big(\lim_{z\to x^+} \phi_j(z)\Big) \vee \Big(\lim_{z\to x^-} \phi_j(z)\Big). \tag{4.40}$$

Then, there are "essentially" two cases.

Case 1. $\phi_1(y_0) = \lim_{z\to y_0^+} \phi_1(z), \phi_2(y_0) = \lim_{z\to y_0^+} \phi_2(z)$.

Case 2. $\phi_1(y_0) = \lim_{z\to y_0^+} \phi_1(z), \phi_2(y_0) = \lim_{z\to y_0^-} \phi_2(z)$.

In Case 1, let $\varepsilon > 0$ be such that $\phi_1(y_0) + \varepsilon < \phi_2(y_0) - \varepsilon$; then

$$\exists \delta_1, \delta_2 > 0 \quad \text{such that } 0 < z - y_0 < \delta_1 \wedge \delta_2 \Rightarrow \begin{cases} |\phi_1(z) - \phi_1(y_0)| < \varepsilon, \\ |\phi_2(z) - \phi_2(y_0)| < \varepsilon; \end{cases}$$

i.e.,

$$0 < z - y_0 < \delta_1 \wedge \delta_2 \Rightarrow \phi_1(z) < \phi_1(y_0) + \varepsilon < \phi_2(y_0) - \varepsilon < \phi_2(z),$$

but this is impossible since $\phi_1 = \phi_2 = \psi_i$ almost everywhere.

In Case 2 one has

$$\lim_{z \to y_0^-} \phi_1(z) = \underline{\lim}_{z \to y_0} \phi_1(z) \leq \lim_{z \to y_0^+} \phi_1(z) = \phi_1(y_0) < \phi_2(y_0) = \lim_{z \to y_0^-} \phi_2(z);$$

i.e.,

$$\lim_{z \to y_0^-} \phi_1(z) < \lim_{z \to y_0^-} \phi_2(z).$$

Then, the same arguments as for Case 1 show that it is possible to find $\varepsilon > 0$ and $\delta_1, \delta_2 > 0$ such that

$$z < y_0, z \in B(y_0, \delta_1 \wedge \delta_2) \Rightarrow \phi_1(z) < \phi_1(y_0) + \varepsilon < \phi_2(y_0) - \varepsilon < \phi_2(z),$$

and again we end up contradicting $\phi_1 = \phi_2 = \psi_i$ almost everywhere. □

Proposition 4.18 justifies the following definition.

DEFINITION 4.19. *For $i = 1, 2$ we define $\tilde{\psi}_i$ as the unique element of the set*

$$\left\{\phi \in \text{bv}(\psi_i) \cap \{\text{u.s.c.}\} \mid \phi(x) = \overline{\lim_{z \to x}}\, \phi(z), x \in \mathbb{R}\right\}.$$

LEMMA 4.20. *Assume* (LFP). *For every $x \in \mathbb{R}$ one has $\tilde{\psi}_i(x) \leq \psi_i(x), i = 1, 2$.*

Proof. Clearly $\tilde{\psi}_i = \psi_i$ almost everywhere, so let $y_0 \in \mathbb{R}$ be a point where $\tilde{\psi}_i \neq \psi_i$ and let us assume that

$$\tilde{\psi}_i(y_0) > \psi_i(y_0). \tag{4.41}$$

Also, we may assume (for example) that

$$\tilde{\psi}_i(y_0) = \lim_{z \to y_0^-} \tilde{\psi}_i(z), \tag{4.42}$$

since $\tilde{\psi}_i$ satisfies (4.40). Now let $\gamma > 0$ be such that

$$\lim_{z \to y_0^-} \tilde{\psi}_i(z) > \gamma > \psi_i(y_0);$$

then

$$\exists \delta > 0 \quad \text{such that } z < y_0, |z - y_0| < \delta \Rightarrow \tilde{\psi}_i(z) > \gamma;$$

hence

$$\inf_{0 < y_0 - z < \delta} \tilde{\psi}_i(z) \geq \gamma > \psi_i(y_0). \tag{4.43}$$

On the other hand, from the upper semicontinuity of ψ_i follows

$$\gamma > \overline{\lim_{z \to y_0}}\, \psi_i(z),$$

and so

$$\exists \delta_0 > 0 \quad \text{such that} \quad \sup_{0 < |z - y_0| < \delta_0} \psi_i(z) < \gamma. \tag{4.44}$$

Thus, (4.43) and (4.44) imply

$$\sup_{0<y_0-z<\delta\wedge\delta_0} \psi_i(z) < \inf_{0<y_0-z<\delta\wedge\delta_0} \tilde{\psi}_i(z),$$

i.e.,

$$0 < y_0 - z < \delta \wedge \delta_0 \Rightarrow \psi_i(z) < \tilde{\psi}_i(z)$$

and this contradicts $\psi_i = \tilde{\psi}_i$ almost everywhere. Therefore, (4.41) must be false and the lemma is proved. □

LEMMA 4.21. *Assume* (LFP). *For every* $x \in \mathbb{R}$ *one has* $\underline{\lim}_{z\to x} \psi_i(z) \leq \tilde{\psi}_i(x)$, $i = 1, 2$.

Proof. Assume not; then there exists $\bar{x} \in \mathbb{R}$ such that

$$\underline{\lim}_{x\to\bar{x}} \psi_i(z) > \tilde{\psi}_i(\bar{x}); \tag{4.45}$$

but $\tilde{\psi}_i$ satisfies (4.40), so there is no loss of generality if we assume (for example)

$$\tilde{\psi}_i(\bar{x}) = \lim_{z\to\bar{x}^+} \tilde{\psi}_i(z). \tag{4.46}$$

Now let $\gamma > 0$ be such that $\underline{\lim}_{z\to\bar{x}} \psi_i(z) > \gamma > \tilde{\psi}_i(\bar{x})$; then from (4.46) follows

$$\exists \delta_1 > 0 \quad \text{such that } 0 < z - \bar{x} < \delta_1 \Rightarrow \tilde{\psi}_i(z) < \gamma;$$

hence

$$\sup_{0<z-\bar{x}<\delta_1} \tilde{\psi}_i(z) \leq \gamma. \tag{4.47}$$

On the other hand, since $\underline{\lim}_{z\to\bar{x}} \psi_i(z) > \gamma$, we have

$$\exists \delta_2 > 0 \quad \text{such that } \inf_{0<|z-\bar{x}|<\delta_2} \psi_i(z) > \gamma. \tag{4.48}$$

Therefore, if $\delta = \delta_1 \wedge \delta_2$, (4.47) and (4.48) imply

$$\sup_{0<z-\bar{x}<\delta} \tilde{\psi}_i(z) < \inf_{0<z-\bar{x}<\delta} \psi_i(z),$$

i.e.,

$$0 < z - \bar{x} < \delta \Rightarrow \tilde{\psi}_i(z) < \psi_i(z),$$

but this is impossible since $\tilde{\psi}_i = \psi_i$ almost everywhere. □

Remark 4.22. From Lemma 4.20 and Lemma 4.21 one has

$$\underline{\lim}_{z\to x} \psi_i(z) \leq \tilde{\psi}_i(x) = \overline{\lim}_{z\to x} \tilde{\psi}_i(z) \leq \overline{\lim}_{z\to x} \psi_i(z) \leq \psi_i(x), \tag{4.49}$$

for every $x \in \mathbb{R}, i = 1, 2$. In particular, Lemma 4.20 implies that

$$\text{graph}(\tilde{\psi}_1) \cap \text{graph}(\tilde{\psi}_2) = \{P_0\}, \tag{4.50}$$

since $\tilde{\psi}_i \equiv \psi_i$ on $(-\infty, x_j^0]$ (by Proposition 3.6, (iii), Lemma 3.8, and the definition of $\tilde{\psi}_i$).

PROPOSITION 4.23. *Assume* (LFP). *For the non-dense part of R_i in $\tilde{\mathcal{A}}$, $\partial_i\backslash\tilde{\partial}_i$, one has*

$$\text{(4.51)}\qquad \begin{cases} \partial_1\backslash\tilde{\partial}_1 \subset \left\{(x_1,x_2) : x_1 \in \left(\underline{\lim}_{z\to x_2} \psi_1(z), \psi_1(x_2)\right], x_2 > \psi_2(x_1)\right\}, \\ \partial_2\backslash\tilde{\partial}_2 \subset \left\{(x_1,x_2) : x_2 \in \left(\underline{\lim}_{z\to x_1} \psi_2(z), \psi_2(x_1)\right], x_1 > \psi_1(x_2)\right\}. \end{cases}$$

Proof. Let us recall that $\partial_1\backslash\tilde{\partial}_1 = R_1 \cap \tilde{\mathcal{A}}$ (cf. Lemma 4.15). Let $P = (x_1, x_2) \in \partial_1\backslash\tilde{\partial}_1$; then $P \in R_1$ and hence (cf. (4.9))

$$\text{(4.52)}\qquad x_1 \le \psi_1(x_2), \qquad x_2 > \psi_2(x_1).$$

Also, $P \in \tilde{\mathcal{A}}$ and $\tilde{\mathcal{A}}$ is open, so there is an open ball $B(P,r) \subset \tilde{\mathcal{A}}$, but then (cf. Corollary 4.14)

$$\text{(4.53)}\qquad |B(P,r) \cap R_1| \le |\tilde{\mathcal{A}} \cap R_1| = |\partial_1\backslash\tilde{\partial}_1| \le |\partial_1| = 0.$$

Claim. $x_1 > \underline{\lim}_{z\to x_2} \psi_1(z)$.
In fact, if not, then $x_1 \le \underline{\lim}_{z\to x_2} \psi_1(z)$. Thus,

$$\text{(4.54)}\qquad \forall \varepsilon > 0 \quad \exists \delta > 0 \quad \text{such that} \inf_{0<|z-x_2|<\delta} \psi_1(z) > x_1 - \varepsilon.$$

On the other hand, since ψ_2 is u.s.c., from $x_2 > \psi_2(x_1)$ follows

$$\text{(4.55)}\qquad \exists \varepsilon_0 < r/2 \quad \text{such that} \sup_{0<|t-x_1|<2\varepsilon_0} \psi_2(t) < x_2,$$

and from this we have

$$\text{(4.56)}\qquad \psi_2(x_1 - \varepsilon_0) < x_2 \quad \text{and} \quad (x_1 - \varepsilon_0, x_2) \in B(P,r).$$

Now (4.54) with $\varepsilon = \varepsilon_0$ implies

$$\text{(4.57)}\qquad \exists \delta_0 > 0 \quad \text{such that } 0 < |z - x_2| < \delta_0 \Rightarrow \psi_1(z) > x_1 - \varepsilon_0;$$

hence (4.56) and (4.57) imply

$$0 < z - x_2 < \delta_0 \Rightarrow \begin{cases} \psi_1(z) > x_1 - \varepsilon_0, \\ \psi_2(x_1 - \varepsilon_0) < z; \end{cases}$$

therefore (cf. (4.9))

$$0 < z - x_2 < \delta_0 \Rightarrow (x_1 - \varepsilon_0, z) \in R_1.$$

Also, since $(x_1 - \varepsilon_0, x_2) \in B(P,r)$,

$$\exists \delta_1 > 0 \quad \text{such that } 0 < z - x_2 < \delta_1 \Rightarrow (x_1 - \varepsilon_0, z) \in B(P,r);$$

so for $\delta = \delta_1 \wedge \delta_0$ we have

$$\text{(4.58)}\qquad 0 < z - x_2 < \delta \Rightarrow (x_1 - \varepsilon_0, z) \in R_1 \cap B(P,r),$$

but then, also,

$$\left\{(t,z) \in B(P,r) : x_1 - r < t \le x_1 - \varepsilon_0, x_2 < z < x_2 + \delta\right\} \subset R_1;$$

hence

$$|R_1 \cap B(P,r)| > \frac{(r-\varepsilon_0)\delta}{2} \neq 0, \tag{4.59}$$

and this is impossible because of (4.53). So the claim follows. The claim and (4.52) prove $(4.51)_1$. (The proof of $(4.51)_2$ is the same.) □

PROPOSITION 4.24. *Assume* (LFP). *If* $\bar{x}_j \in \pi_j(\partial_i \backslash \tilde{\partial}_i)$ *where* π_j *is the orthogonal projection on the* x_j*-axis, then*

$$(\partial_i \backslash \tilde{\partial}_i) \cap \left(\left(\varliminf_{z \to \bar{x}_j} \psi_i(z), \tilde{\psi}_i(\bar{x}_j) \right] \times \{\bar{x}_j\} \right) = \emptyset \quad \text{for } i \neq j, \quad i,j = 1,2.$$

Proof. Assume not and take $i = 1, j = 2$ for simplicity. Let $P = (\bar{x}_1, \bar{x}_2) \in \partial_1 \backslash \tilde{\partial}_1$ with

$$\bar{x}_1 \in \left(\varliminf_{z \to \bar{x}_2} \psi_1(z), \tilde{\psi}_1(\bar{x}_2) \right].$$

As in the proof of Proposition 4.23 (cf. (4.53)), from $P \in \partial_1 \backslash \tilde{\partial}_1$ and $\partial_1 \backslash \tilde{\partial}_1 = R_1 \cap \tilde{\mathcal{A}}$ follows that for some $r > 0$, $B(P,r) \subset \tilde{\mathcal{A}}$ and

$$|B(P,r) \cap R_1| = 0. \tag{4.60}$$

Thus, we can select $\gamma \in (\bar{x}_1 - r, \bar{x}_1 + r)$ such that

$$1_{R_1}(\gamma, \cdot) = 0 \qquad \text{a.e. in } B_\gamma(P,r). \tag{4.61}$$

where

$$B_\gamma(P,r) = \{z : (\gamma, z) \in B(P,r)\}.$$

In particular, we may fix

$$\gamma \in (\bar{x}_1 - r, \bar{x}_1 + r) \cap \left(\varliminf_{z \to \bar{x}_2} \psi_1(z), \tilde{\psi}_1(\bar{x}_2) \right),$$

and we may assume (for example)

$$\tilde{\psi}_1(\bar{x}_2) = \varliminf_{z \to x_{2-}} \tilde{\psi}_1(z), \tag{4.62}$$

since $\tilde{\psi}_1$ satisfies (4.40) and Proposition 4.18. Then, (4.62) and $\gamma < \tilde{\psi}_1(\bar{x}_2)$ imply

$$\exists \delta > 0 \quad \text{such that } 0 < \bar{x}_2 - z < \delta \Rightarrow \gamma < \tilde{\psi}_1(z). \tag{4.63}$$

On the other hand, (4.61) implies

$$(\gamma, z) \notin R_1 \quad \text{for a.e. } z \in B_\gamma(P,r); \tag{4.64}$$

therefore it must be

$$(\gamma, z) \in \mathcal{A} \quad \text{for a.e. } z \in B_\gamma(P,r), \tag{4.65}$$

since locally R_1 is the complement of $\mathcal{A}$. (In fact, since $R_1 \cap (R_2 \cup R_0) = \emptyset$, we can always assume $B(P,r) \cap (R_2 \cup R_0) = \emptyset$, with r smaller if necessary.) So from (4.65) we deduce

$$\gamma > \psi_1(z), \quad z > \psi_2(\gamma) \quad \text{for a.e. } z \in B_\gamma(P,r). \tag{4.66}$$

But $\tilde{\psi} \leq \psi_1$ by Lemma 4.20; hence (4.63) and (4.66) imply

$$\gamma > \psi_1(z) \geq \tilde{\psi}_1(z) > \gamma \quad \text{for a.e. } z \in (\bar{x}_2 - \eta, \bar{x}_2) \tag{4.67}$$

for $\eta = \min\{r, \delta\}$, and we have a contradiction. □

COROLLARY 4.25. *Assume* (LFP). *For the non-dense part of R_i in $\tilde{\mathcal{A}}, \partial_i \backslash \tilde{\partial}_i$, one has*

$$\begin{cases} \partial_1 \backslash \tilde{\partial}_1 \subset \{(x_1, x_2) : x_1 \in (\tilde{\psi}_1(x_2), \psi_1(x_2)], x_2 > \psi_2(x_1)\}, \\ \partial_2 \backslash \tilde{\partial}_2 \subset \{(x_1, x_2) : x_2 \in (\tilde{\psi}_2(x_1), \psi_2(x_1)], x_1 > \psi_1(x_2)\}. \end{cases} \tag{4.68}$$

Proof. This follows from Proposition 4.23 and Proposition 4.24. □

We can improve Corollary 4.25. In fact, we now show that the inclusions in (4.68) are equalities.

PROPOSITION 4.26. *Assume* (LFP). *The non-dense part of R_i in $\tilde{\mathcal{A}}, \partial_i \backslash \tilde{\partial}_i$, may be characterized as*

$$\begin{cases} \partial_1 \backslash \tilde{\partial}_1 = \{(x_1, x_2) : x_1 \in (\tilde{\psi}_1(x_2), \psi_1(x_2)], x_2 > \psi_2(x_1)\}, \\ \partial_2 \backslash \tilde{\partial}_2 = \{(x_1, x_2) : x_2 \in (\tilde{\psi}_2(x_1), \psi_2(x_1)], x_1 > \psi_1(x_2)\}. \end{cases} \tag{4.69}$$

Proof. Let $P = (\bar{x}_1, \bar{x}_2) \in \{(x_1, x_2) : x_1 \in (\tilde{\psi}_1(x_2), \psi_1(x_2)], x_2 > \psi_2(x_1)\}$; then $\bar{x}_1 > \tilde{\psi}_1(\bar{x}_2)$, and let us assume that $\psi_1(\bar{x}_2) = \lim_{z \to \bar{x}_2^+} \tilde{\psi}_1(z)$ (this is possible because of (4.40)). So we have

$$\bar{x}_1 > \tilde{\psi}_1(\bar{x}_2) = \lim_{z \to \bar{x}_2^+} \tilde{\psi}_1(z) = \overline{\lim_{z \to \bar{x}_2}} \tilde{\psi}_1(z) \geq \underline{\lim_{z \to \bar{x}_2}} \tilde{\psi}_1(z) = \lim_{z \to \bar{x}_2^-} \tilde{\psi}_1(z);$$

therefore, if $\gamma > 0$ is such that $\bar{x}_1 > \gamma > \tilde{\psi}_1(\bar{x}_2)$, then

$$\exists \delta > 0 \quad \text{such that } |z - \bar{x}_2| < \delta \Rightarrow \tilde{\psi}_1(z) < \gamma < \bar{x}_1.$$

So it is possible to find a ball $B(P, r)$ with $r < \bar{x}_1 - \gamma, r < \delta$, such that

(i) $\{(\tilde{\psi}_1(z), z) : z \in (\bar{x}_2 - \delta, \bar{x}_2 + \delta)\} \cap B(P, r) = \emptyset$,

(ii) $B(P, r) \cap (R_2 \cup \text{int}(R_0)) = \emptyset$

(note that (ii) follows from the fact that $P \in R_1$ and $R_1 \cap (R_2 \cup R_0) = \emptyset$ as in the proof of (4.65)). Now (i) and $\psi_1 > \tilde{\psi}_1$ only on a null set imply

$$\lambda(\{z \in (\bar{x}_2 - \delta, \bar{x}_2 + \delta) : \psi_1(z) > \gamma\}) = 0;$$

hence from Fubini's Theorem, (ii), and the definition of R_1 follows

$$|B(P, r) \cap R_1| = 0,$$

so that P is in the non-dense part of R_1 in $\tilde{\mathcal{A}}, \partial_1 \backslash \tilde{\partial}_1$. Hence $(4.69)_1$ follows from Corollary 4.25. (The proof of $(4.69)_2$ is the same.) □

Remark 4.27. We point out that there may be other points $P \in \partial_i$ such that $D(R_i, P) = 0$; these are points where ψ_i has a cusp pointing into $\mathcal{A}$, and they too have zero Lebesgue density with respect to R_i. The difference between such points and those in $\partial_i \backslash \tilde{\partial}_i$ is that the latter ones verify a condition even stronger than $D(R_i, P) = 0$, namely (4.53), i.e.,

$$\exists r > 0 \quad \text{s.t. } |R_i \cap B(P, r)| = 0. \tag{4.70}$$

We are now ready to characterize the boundary of the new region of inaction $\tilde{\mathcal{A}}$.

THEOREM 4.28. *Assume* (LFP). *The new region of inaction* $\tilde{\mathcal{A}} = \operatorname{int}(\operatorname{cl}(\mathcal{A}))$ *is given by*

$$\tilde{\mathcal{A}} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 > \tilde{\psi}_1(x_2), x_2 > \tilde{\psi}_2(x_1)\} \tag{4.71}$$

with $\tilde{\psi}_1$ *and* $\tilde{\psi}_2$ *as in Definition* 4.19.

Proof. It suffices to recall (cf. (4.31)) that

$$\tilde{\mathcal{A}} = \mathcal{A} \cup (\partial\mathcal{A} \backslash \partial\tilde{\mathcal{A}}) = \mathcal{A} \cup \bigcup_{i=1}^{2} \partial_i \backslash \tilde{\partial}_i,$$

so (4.71) follows from (4.7), $\tilde{\psi}_i \leq \psi_i, \tilde{\psi}_i = \psi_i$ almost everywhere, Proposition 4.26, and the fact that

$$\{(x_1, x_2) : \psi_1(x_2) \geq x_1 > \tilde{\psi}_1(x_2), \psi_2(x_1) \geq x_2 > \tilde{\psi}_2(x_1)\} = \emptyset \tag{4.72}$$

(in fact, $x_1 \leq \psi_1(x_2)$ and $x_2 \leq \psi_2(x_1)$ imply $(x_1, x_2) \in R_0 = \{(x_1^0, x_2^0)\} - \Lambda^*$, i.e., $x_i \in (-\infty, x_i^0], i = 1, 2$, but there $\psi_i = \tilde{\psi}_i$ (cf. Remark 4.22)). □

Hence $\partial\tilde{\mathcal{A}}$ is essentially obtained by adding all the finite line segments corresponding to the jumps of $\tilde{\psi}_i$ to the graph of $\tilde{\psi}_i|_{[x_j^0, \infty)}, i = 1, 2, j = 1, 2, j \neq i$.

DEFINITION 4.29. *Let* $\{\zeta_j\}_{j=1}^{\infty}$ *and* $\{\xi_j\}_{j=1}^{\infty}$ *be the points of discontinuity of* $\tilde{\psi}_1$ *and* $\tilde{\psi}_2$, *respectively. Then we set*

$$\begin{cases} \tilde{\psi}_1[\underline{j}] = \underline{\lim}_{z \to \zeta_j} \tilde{\psi}_1(z), \\ \tilde{\psi}_1[j] = \overline{\lim}_{z \to \zeta_j} \tilde{\psi}_1(z) = \tilde{\psi}_1(\zeta_j), \end{cases} \tag{4.73}$$

and similarly

$$\begin{cases} \tilde{\psi}_2[\underline{j}] = \underline{\lim}_{z \to \xi_j} \tilde{\psi}_2(z), \\ \tilde{\psi}_2[j] = \overline{\lim}_{z \to \xi_j} \tilde{\psi}_2(z) = \tilde{\psi}_2(\xi_j), \end{cases} \tag{4.74}$$

for every $j \in \mathbb{N}$.

PROPOSITION 4.30. *Assume* (LFP). *Then the new free boundary* $\partial\tilde{\mathcal{A}}$ *is given by*

$$\begin{aligned} \partial\tilde{\mathcal{A}} = \tilde{\partial}_1 \cup \tilde{\partial}_2 \; = & \left(\operatorname{graph}(\tilde{\psi}_1|_{[x_2^0, +\infty)}) \cup \bigcup_{j=1}^{\infty} [\tilde{\psi}_1[\underline{j}], \tilde{\psi}_1[j]) \times \{\zeta_j\} \right) \\ & \cup \left(\operatorname{graph}(\tilde{\psi}_2|_{[x_1^0, +\infty)}) \cup \bigcup_{j=1}^{\infty} \{\xi_j\} \times [\tilde{\psi}_2[\underline{j}], \tilde{\psi}_2[j]) \right). \end{aligned} \tag{4.75}$$

Proof. This is obvious from Theorem 4.28. □

As we observed at the beginning of this section, all points on the boundary $\partial\tilde{\mathcal{A}}$ where a tangent vector exists belong to the measure-theoretic boundary $\partial_M\tilde{\mathcal{A}}$. Here we show that this is in fact the case for almost every point of $\partial\tilde{\mathcal{A}}$. This result is an obvious consequence of the rectifiability of the boundaries $\tilde{\partial}_1$ and $\tilde{\partial}_2$. (Recall that $\tilde{\partial}_i$ is rectifiable since $\tilde{\psi}_i$ is a function of bounded variation.)

PROPOSITION 4.31. *Assume* (LFP). *Then the topological boundary $\partial\tilde{\mathcal{A}}$ and the measure-theoretic boundary $\partial_M\tilde{\mathcal{A}}$ are the same except for a set of one-dimensional Hausdorff measure zero, i.e.,*

$$H^1(\partial\tilde{\mathcal{A}}\backslash\partial_M\tilde{\mathcal{A}}) = 0. \tag{4.76}$$

Proof. It suffices to show that there exists a definite tangent to $\partial\tilde{\mathcal{A}}$ almost everywhere with respect to the one-dimensional Hausdorff measure H^1 in $\mathbb{R}^2$. But $\tilde{\psi}_1$ and $\tilde{\psi}_2$ are functions of locally bounded variation (in the usual sense); hence $\tilde{\partial}_1$ and $\tilde{\partial}_2$ are locally rectifiable curves, and the measure H^1 coincides with the arc-length s. Also, a result due to Tonelli (cf. [14]) guarantees that the classical formula

$$(s'(t))^2 = (x'(t))^2 + (y'(t))^2$$

is valid almost everywhere with respect to the parameter t for every rectifiable curve (if locally, $x = x(t)$, $y = y(t)$ is a parametric representation of the curve). In particular, if we choose the arc-length s as parameter, we obtain

$$(x'(s))^2 + (y'(s))^2 = 1 \qquad \text{a.e.}$$

and hence $x'(s), y'(s)$ exist almost everywhere (and are not both zero almost everywhere) assuring the existence of a definite tangent almost everywhere with respect to s. □

Finally, from Proposition 4.31 and Theorem 4.8 we obtain the regularity of the entire boundary of the new region of inaction.

THEOREM 4.32. *Assume* (LFP). *Then the new region of inaction $\tilde{\mathcal{A}}$ is of locally finite perimeter and its boundary $\partial\tilde{\mathcal{A}}$ is countably 1-rectifiable, i.e.,*

$$\partial\tilde{\mathcal{A}} \subset \bigcup_{i=1}^{\infty} M_i \cup N, \tag{4.77}$$

where $H^1(N) = 0$ and each M_i is a one-dimensional embedded C^1 submanifold of $\mathbb{R}^2$. □

5. Verification of (LFP). In the previous section we obtained the regularity of the free boundary arising in the control problem defined by (2.1) under the assumption (LFP); that is, we assumed the region of inaction $\mathcal{A}$ to be of locally finite perimeter. We aim to show that such an assumption is, after all, reasonable and verifiable. We restrict ourselves to the case where the diffusion matrix σ is nondegenerate. Such a condition naturally implies the coercivity (see below) of the bilinear form $a(u, v)$ associated with the operator Au, and this will allow us to show (LFP) by means of a localization of a result obtained by Brezis and Kinderlehrer [1] in the framework of variational inequalities with obstacles relative to locally coercive vector fields.

In addition to the assumptions stated in §2, we now assume the following:

$$\sigma\sigma^* \text{ is positive definite;} \tag{5.1}$$

$$f \in C^2(\mathbb{R}^2); \tag{5.2}$$

$$f_{x_i} \text{ and } \nabla(f_{x_i}) \text{ never vanish simultaneously } (i = 1, 2). \tag{5.3}$$

(It should be noticed that we already had $f \in C^{1,1}(\mathbb{R}^2)$ as this follows from the growth conditions (2.5)–(2.7) by using the same arguments as in Theorem 2.1.) Let $W_0^{1,2}(\Omega)$ be the closure of $C_0^\infty(\Omega)$ in $W^{1,2}(\Omega)$, for any open set $\Omega \subset \mathbb{R}^2$.

DEFINITION 5.1 ([6, p. 15]). *A bilinear form $a(u,v)$ is said to be coercive on $W_0^{1,2}(\Omega)$ if*

$$\exists \nu > 0 \quad \text{such that } a(u,u) \geq \nu \|u\|_{1,2}^2 \quad \text{for every } u \in W_0^{1,2}(\Omega), \tag{5.4}$$

where $\|\cdot\|_{1,2}$ is the norm in $W^{1,2}(\Omega)$.

In particular, we consider the bilinear form $a(u,v)$ associated with the operator Au, i.e.,

$$a(u,v) := \int_\Omega \left\{ \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} u_{x_i} v_{x_j} - \sum_{i=1}^{2} g_i u_{x_i} v + \rho uv \right\} dx \tag{5.5}$$

for $u, v \in W^{1,2}(\Omega)$, with Ω open in $\mathbb{R}^2$ (to be chosen later). Let us recall a few known results.

LEMMA 5.2 ([15, Lemma 4.3]). *Assume* (5.1). *Let $\Omega \subset \mathbb{R}^2$ be an open ball and let $a(u,v)$ be defined by* (5.5). *Then, $a(u,v)$ is coercive on $W_0^{1,2}(\Omega)$.*

Proof. This follows from $\rho > 0$, $\sigma\sigma^*$ positive definite and $\int_\Omega u u_{x_i} dx = 0$ for $u \in W_0^{1,2}(\Omega), i = 1,2$. □

DEFINITION 5.3. *Let a be as in* (5.5) *and let*

$$\mathbb{K}(\Omega) := \{v \in W^{1,2}(\Omega) : v \geq 0 \text{ a.e. in } \Omega\}. \tag{5.6}$$

We say that w is a local solution of the variational inequality

$$a(w, v-w) \geq (f_{x_i}, v-w) \qquad \forall v \in \mathbb{K}(\Omega), \tag{5.7}$$

if $w \in \mathbb{K}(\Omega)$ and we have

$$a(w, \eta(v-w)) \geq \int_\Omega f_{x_i} \eta (v-w)\, dx \quad \forall v \in \mathbb{K}(\Omega), \eta \in C_0^\infty(\Omega), \eta \geq 0. \tag{5.8}$$

THEOREM 5.4 ([15, Theorem 4.5]). *Assume* (5.1) *and* (5.2). *Let Ω be an open ball such that* $\mathrm{cl}(\Omega) \subset S_i$ *where*

$$S_i := \{x \in \mathbb{R}^2 : \hat{u}_{x_j}(x) > 0, j \neq i\}, \tag{5.9}$$

then $\hat{u}_{x_i}$ is a local solution of (5.7), $i = 1,2$.

THEOREM 5.5 ([6, problem 5, p. 30 and problem 1, p. 44]; [15, Thm. 4.6]). *Assume* (5.1) *and* (5.2). *Let Ω be an open ball such that* $\mathrm{cl}(\Omega) \subset S_i$, *then*

(i) $\hat{u}_{x_i} \in W^{2,\infty}(\Omega)$;

(ii) $A\hat{u}_{x_i} \geq f_{x_i}, \hat{u}_{x_i} \geq 0, (A\hat{u}_{x_i} - f_{x_i})\hat{u}_{x_i} = 0$ *almost everywhere in* Ω.

Clearly,

$$S_i = R_i \cup \mathcal{A}, \qquad i = 1,2;$$

hence Theorem 5.5(i) implies (for $i = 1,2$)

$$\hat{u}_{x_i} \in C^{1,1}(\Omega) \tag{5.10}$$

if Ω is an open ball such that $\mathrm{cl}(\Omega) \subset R_i \cup \mathcal{A}$.

The following lemma is proved in the Appendix.

LEMMA 5.6 ([6, problem 5, p. 30]). *Assume* (5.1) *and* (5.2). *If w is a local solution of* (5.7), *then*

$$
\begin{aligned}
a(\gamma w, v-\gamma w) \geq & \int_\Omega f_{x_i}\gamma(v-\gamma w)\,dx \\
& - \int_\Omega \left\{ \frac{1}{2}\sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} w\gamma_{x_i x_j} + \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} w_{x_i}\gamma_{x_j} \right\} (v-\gamma w)\,dx \\
& - \int_\Omega \sum_{i=1}^{2} g_i \gamma_{x_i} w(v-\gamma w)\,dx
\end{aligned} \tag{5.11}
$$

for every $v \in \mathbb{K}_0(\Omega)$ and $\gamma \in C_0^\infty(\Omega), \gamma = 1$ on $\Omega', 0 \leq \gamma \leq 1$ in Ω ($\mathrm{cl}(\Omega') \subset \Omega$), where

$$
\mathbb{K}_0(\Omega) := \{v \in W_0^{1,2}(\Omega) : v \geq 0 \text{ a.e.}\}. \tag{5.12}
$$

Remark 5.7. We point out that, if w is a local solution of (5.7), then γw is the unique solution of (5.11) since $a(u,v)$ is coercive (cf. [6, Thm. 2.7, p. 15]).

Let us define the bilinear form

$$
\tilde{a}(u,v) := \int_\Omega \frac{1}{2}\sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} u_{x_i} v_{x_j}\,dx \tag{5.13}
$$

for $u, v \in W^{1,2}(\Omega)$, Ω open in $\mathbb{R}^2$; also, we set

$$
\begin{aligned}
\tilde{F}_r := & f_{x_r}\gamma - \frac{1}{2}\sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij}\hat{u}_{x_r}\gamma_{x_i x_j} \\
& - \frac{1}{2}\sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij}(\hat{u}_{x_r})_{x_i}\gamma_{x_j} + \sum_{i=1}^{2} g_i (\hat{u}_{x_r})_{x_i}\gamma \\
& - \rho\hat{u}_{x_r}\gamma, \qquad r = 1,2.
\end{aligned} \tag{5.14}
$$

Then, if Ω is an open ball such that $\mathrm{cl}(\Omega) \subset R_r \cup \mathcal{A}\,(r = 1,2)$ and $\gamma \in C_0^\infty(\Omega), \gamma = 1$ on $\Omega', 0 \leq \gamma \leq 1$ in Ω, $\mathrm{cl}(\Omega') \subset \Omega$, from Theorem 5.4 it follows that $w = \hat{u}_{x_r}$ satisfies (5.11), which can be restated as

$$
\tilde{a}(\gamma\hat{u}_{x_r}, v-\gamma\hat{u}_{x_r}) \geq \int_\Omega \tilde{F}_r(v-\gamma\hat{u}_{x_r})\,dx, \qquad \forall\, v \in \mathbb{K}_0(\Omega), \tag{5.15}
$$

where $\hat{u}_{x_r} \in W^{2,\infty}(\Omega)$ (by Theorem 5.5(i)), and hence $\tilde{F}_r \in W^{1,\infty}(\Omega)$ (by (5.14)). Clearly, the bilinear form $\tilde{a}(u,v)$ is coercive and $\gamma\hat{u}_{x_r} \in \mathbb{K}_0(\Omega)$; hence $\gamma\hat{u}_{x_r}$ is the unique solution of (5.15) (cf. Remark 5.7). Also, if we denote by

$$
\tilde{A}w = -\frac{1}{2}\sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} w_{x_i x_j} \tag{5.16}
$$

whenever the RHS makes sense (in the sense of distributions), then (5.15) may be formulated as

$$
(\tilde{A}(\gamma\hat{u}_{x_r}), v-\gamma\hat{u}_{x_r}) \geq \int_\Omega \tilde{F}_r(v-\gamma\hat{u}_{x_r})\,dx, \qquad \forall\, v \in \mathbb{K}_0(\Omega); \tag{5.17}
$$

now we observe that $\gamma\hat{u}_{x_r}$ is a Lipschitz function (in fact, $\gamma\hat{u}_{x_r} \in C^{1,1}(\Omega)$ by Sobolev embedding theorem); hence we can restrict (5.17) to $\mathcal{L}_0(\Omega)$, the convex set of Lipschitz functions v satisfying $v|_{\partial\Omega} = 0$ and $v \geq 0$ almost everywhere in Ω; i.e.,

$$(\tilde{A}(\gamma\hat{u}_{x_r}), v - \gamma\hat{u}_{x_r}) \geq \int_\Omega \tilde{F}_r(v - \gamma\hat{u}_{x_r})\,dx, \qquad \forall\, v \in \mathcal{L}_0(\Omega). \tag{5.18}$$

The variational inequality (5.18) is now in the setting of the problem studied by Brezis and Kinderlehrer [1], except for the fact that our $\tilde{F}_r \in W^{1,\infty}(\Omega)$, while theirs is in $C^1(\mathrm{cl}(\Omega))$. However, it is easy to see that all their estimates still hold in our case since they depend only on $\|\tilde{F}_r\|_{1,\infty}$ (where $\|\cdot\|_{1,\infty}$ is the norm in $W^{1,\infty}(\Omega)$); hence Brezis and Kinderlehrer's Theorem 4 applies and provides us with

$$\tilde{A}(\gamma\hat{u}_{x_r}) \in BV_{\mathrm{loc}}(\Omega), \qquad r = 1, 2. \tag{5.19}$$

Since $\tilde{A}(\gamma\hat{u}_{x_r}) = A(\gamma\hat{u}_{x_r}) + \sum_{i=1}^2 g_i(\gamma\hat{u}_{x_r})_{x_i} - \rho\gamma\hat{u}_{x_r}$ and $\gamma\hat{u}_{x_r} \in C^{1,1}(\Omega)$ (as we observed above), from (5.19) follows

$$A(\gamma\hat{u}_{x_r}) \in BV_{\mathrm{loc}}(\Omega), \qquad r = 1, 2; \tag{5.20}$$

but $\gamma \equiv 1$ on Ω', $\mathrm{cl}(\Omega') \subset \Omega$; hence we have

$$A\hat{u}_{x_r} \in BV_{\mathrm{loc}}(\Omega), \qquad r = 1, 2. \tag{5.21}$$

Now (5.21) and $f_{x_r} \in C^1(\mathbb{R}^2)$ (by 5.2)) imply $(f_{x_r} - A\hat{u}_{x_r}) \in BV_{\mathrm{loc}}(\Omega)$. Therefore, from Theorem 5.5(ii) will follow

$$1_{R_r \cap \Omega} = \frac{f_{x_r} - A\hat{u}_{x_r}}{f_{x_r}} \qquad \text{a.e. in } \Omega \tag{5.22}$$

if we establish the following two facts: $f_{x_r} \neq 0$ in Ω and $A\hat{u}_{x_r} = 0$ almost everywhere in $R_r \cap \Omega$. Then, (5.22) will imply $1_{R_r \cap \Omega} \in BV_{\mathrm{loc}}(\Omega)$, i.e. we will obtain the local finiteness of the perimeter of $R_r, r = 1, 2$, and hence that of $\mathcal{A}$. So we need to show that $f_{x_r} \neq 0$ in a neighborhood of the free boundary ∂_r. This follows from a generalization of Lemma 7.3, p. 195 of [6], which makes essential use of the hypotheses (5.1)–(5.3). With our notation such a result is stated as Theorem 5.8.

THEOREM 5.8 ([15, Thm. 4.8 and Cor. 4.9]). *Assume* (5.1), (5.2), *and* (5.3). *Then*

$$f_{x_r} < 0 \quad \textit{on the free boundary determined by Theorem } 5.5\text{(ii)}, r = 1, 2; \tag{5.23}$$

hence

$$f_{x_r} < 0 \quad \textit{on } \partial_r \backslash \partial_0, \qquad r = 1, 2. \tag{5.24}$$

Remark 5.9. Clearly $\partial_r \backslash \partial_0$ and the free boundary determined by (ii) of Theorem 5.5 coincide; in fact they both are characterized by the function ψ_r and the set S_r (cf. (5.9)).

We point out in [4, Remark 6.1] that the strict convexity of f makes (5.3) unnecessary in the above theorem.

LEMMA 5.10. *Assume* (5.1) *and* (5.2). *Then*

$$\hat{u}_{x_r} = 0 \quad \textit{and} \quad \nabla\hat{u}_{x_r} = 0 \quad \textit{on } R_r, \quad r = 1, 2; \tag{5.25}$$

$$A\hat{u}_{x_r} = 0 \quad \text{a.e. } \textit{in } R_r, \quad r = 1, 2. \tag{5.26}$$

Proof. Since $\hat{u}_{x_r} \in C^{1,1}(S_r)$ (cf. (5.10)) and $\hat{u}_{x_r}$ attains its minimum value in R_r, (5.25) follows. Now (5.26) is immediate if the boundary ∂_r has zero two-dimensional Lebesgue measure; otherwise from the fact that the $\hat{u}_{x_r x_i x_j}$'s exist almost everywhere and from the properties of R_r (for $j \neq r$, the x_j-sections of R_r are half-lines) it follows $\hat{u}_{x_r x_r x_r} = \hat{u}_{x_r x_r x_j} = \hat{u}_{x_r x_j x_r} = 0$ almost everywhere in $\partial_r \backslash \partial_0$; then $\hat{u}_{x_r x_j x_j}(P) = 0$ for almost every P in $\partial_r \backslash \partial_0$ follows by taking the limit of the Newton quotient along a sequence in $\{x_r = x_r(P)\} \cap (\partial_r \backslash \partial_0)$ and by recalling that there $\hat{u}_{x_r x_j} \equiv 0$ (such a sequence exists for almost every P if $|\partial_r| > 0$). □

Remark 5.11. As a matter of fact, it is possible to show that $|\partial_r| = 0$ under the conditions (5.1) and (5.2) (cf. [6, Thm. 3.4 and Thm. 3.5, p. 155]).

We are now ready to prove (LFP).

THEOREM 5.12. *Assume* (5.1) *and* (5.2). *Then the region of inaction $\mathcal{A}$ is of locally finite perimeter, i.e.,* (LFP) *is verified.*

Proof. Since $f_{x_r} \in C(\mathbb{R}^2)$ and $f_{x_r}|_{\partial_r \backslash \partial_0} < 0$, we can cover ∂_r with open balls Ω so that (5.22) holds there, i.e.,

$$1_{R_r \cap \Omega} = \frac{f_{x_r} - A\hat{u}_{x_r}}{f_{x_r}} \qquad \text{a.e. in } \Omega.$$

Then, $1_{R_r} \in BV_{\text{loc}}(\Omega), r = 1, 2$ (by Theorem 5.8, (5.2), and (5.21)); also $\partial\mathcal{A} = \partial_1 \cup \partial_2$; hence the assertion of the theorem follows. □

Clearly all the results of §4 hold in the present setting (since (LFP) holds for $\mathcal{A}$). However, under assumptions (5.1) and (5.2), Proposition 4.30 can be considerably improved thanks to the greater regularity of $\hat{u}_{x_r}$, as we show in the next section.

6. Higher regularity of the boundary. In this section we assume (5.1) and (5.2) and we show (with the notation of the previous sections) that the function ψ_i which defines the boundary ∂_i is Lipschitz continuous, and therefore $\psi_i = \tilde{\psi}_i$. The proof is a generalization of a result concerning the regularity of the free boundary of a filtration problem (cf. [6, Thm. 6.1, p. 177]). The arguments of the proof are based on PDE methods (and for these we need to assume (5.1) and (5.2)), and also on the geometry of the problem (that is, on the results of §3).

The results of this section parallel results in [15], but these results were not included in the earlier version of [15] available to the authors at the time of writing.

We need the following lemma that provides us with some basic properties of $\hat{u}_{x_r}$. Let us recall that $\hat{u}_{x_r} \in C^{1,1}(S_r)$ (cf. (5.10)).

LEMMA 6.1 ([6, Lem. 3.2 and Cor. 3.3, p. 155]). *Assume* (5.1) *and* (5.2). *Let Ω be an open ball in S_r that intersects $\partial_r \backslash \partial_0$, let $P \in \mathcal{A} \cap \Omega$ be such that* $\operatorname{dist}(P, \partial_r \backslash \partial_0) < \delta$, $\operatorname{dist}(P, \partial\Omega) \geq \varepsilon_0 > 0$, *and let $M > 0$ be such that*

$$|D_{ij}(\hat{u}_{x_r})| \leq M \quad \text{in } \Omega,$$

for all i, j. Then

$$\hat{u}_{x_r}(P) \leq M\delta^2 \leq C\delta^2, \tag{6.1}$$

$$|\nabla \hat{u}_{x_r}(P)| \leq C\delta, \tag{6.2}$$

where $C = C(\varepsilon_0, \|\hat{u}_{x_r}\|_{C^{1,1}(\Omega)}), r = 1, 2$.

THEOREM 6.2. *Assume* (5.1) *and $f \in C^3(\mathbb{R}^2)$. Then the functions ψ_1 and ψ_2 are continuous everywhere and locally Lipschitz away from the corner point ∂_0. In particular, the free*

boundary $\partial\mathcal{A}$ is given by

$$\partial\mathcal{A} = \partial_1 \cup \partial_2 = (\text{graph}(\psi_1|_{[x_2^0,+\infty)})) \cup (\text{graph}(\psi_2|_{[x_1^0,+\infty)})). \tag{6.3}$$

Proof (see also [6, Thm. 6.1, p. 177]). Fix $i = 1$ for simplicity and recall that

$$\hat{u}_{x_1} \in C^{1,1}(\Omega), \qquad i = 1, 2,$$

for every open ball $\Omega \subset \mathcal{A}$ (cf. (5.10)); hence

$$\hat{u} \in C^{2,1}(\Omega), \tag{6.4}$$

and so also

$$\hat{u} \in C^2(\mathcal{A}). \tag{6.5}$$

Then, from (6.4) and $f \in C^{2,\alpha}(\Omega)$ (by 5.1) follows

$$\hat{u} \in C^{4,\alpha}(\Omega) \quad \text{for every open ball } \Omega \subset \mathcal{A}, \tag{6.6}$$

by elliptic regularity (cf. [7, Prop. 6.17, p. 109]). This enables us to differentiate the Bellman equation once more; i.e., we get

$$A\hat{u}_{x_1x_1} = f_{x_1x_1} \quad \text{in } \mathcal{A},$$

and $f_{x_1x_1} \geq 0$ (by convexity) implies

$$A\hat{u}_{x_1x_1} \geq 0 \quad \text{in } \mathcal{A}. \tag{6.7}$$

Also, $\hat{u}_{x_1x_1} \geq 0$ (again by convexity) and $\hat{u}_{x_1x_1} = 0$ on $\partial_1\backslash\partial_0$ (by (5.25)). Now let Ω be an open ball in $S_1 = R_1 \cup \mathcal{A}$ such that $\Omega \cap \partial_1 \neq \emptyset$; then we apply the strong maximum principle (cf. [7, Thm. 3.5, p. 35]) to

$$\begin{cases} A\hat{u}_{x_1x_1} \geq 0 & \text{in } \Omega \cap \mathcal{A}, \\ \hat{u}_{x_1x_1} = 0 & \text{on } \Omega \cap \partial_1, \\ \hat{u}_{x_1x_1} \geq 0 & \text{in } \Omega, \\ \hat{u}_{x_1x_1} \in C^{2,\alpha}(\Omega \cap \mathcal{A}) \cap C^{0,1}(\Omega), \end{cases}$$

and we conclude ($\rho > 0$ is used here)

$$\hat{u}_{x_1x_1} > 0 \quad \text{in } \Omega \cap \mathcal{A}, \tag{6.8}$$

since if not, then the minimum (which is zero) would be achieved inside Ω and this contradicts the maximum principle unless $\hat{u}_{x_1x_1} = 0$ on $\Omega \cap \mathcal{A}$. In this case $\hat{u}_{x_1} = 0$ on horizontal line segments in $\Omega \cap \mathcal{A}$ contradicting the definition of $\mathcal{A}$.

We recall that $f_{x_1} < 0$ on $\partial_1\backslash\partial_0$ (by Theorem 5.8); hence, by continuity, $f_{x_1} < 0$ in a neighborhood of $\partial_1\backslash\partial_0$, and we assume

$$f_{x_1} < 0 \quad \text{in } \Omega. \tag{6.9}$$

Let $Q = (x_1^Q, x_2^Q) \in \partial_1 \cap \Omega$ and let $R > 0$ be such that there exists $b > 0$ for which

$$D_R := \{(x_1, x_2) \in \mathcal{A} : |x_2 - x_2^Q| < R,\ x_1 < b\} \subset \Omega,$$
$$\text{cl}(D_{2R}) := \text{cl}\{(x_1, x_2) \in \mathcal{A} : |x_2 - x_2^Q| < 2R,\ x_1 < b\} \subset \Omega.$$

Set

$$w := K\hat{u}_{x_1x_1} + H\hat{u}_{x_1x_2} - F\hat{u}_{x_1}$$

for $K > 0, F > 0$, and $|H| \leq 1$ (with the constants K and F to be chosen later). Then we have

$$Aw = Kf_{x_1x_1} + Hf_{x_1x_2} - Ff_{x_1},$$

and from this follows, for F sufficiently large (independently of K and H),

$$Aw > \varepsilon > 0 \quad \text{in } D_{2R}, \tag{6.10}$$

since $f_{x_1x_1}, f_{x_1x_2}$ are bounded in Ω and (6.9) holds. Also

$$w = 0 \quad \text{on } \partial D_R \cap \partial_1, \tag{6.11}$$

and (6.8) implies

$$w > 0 \quad \text{on } \partial D_{R'} \cap \{\text{dist}(\cdot, \partial_1) > \delta\}, \quad \text{for all } R \leq R' \leq 2R, \tag{6.12}$$

if $K \geq K(\delta)$ with $0 < \delta < 1$, a small number to be determined later. If we show

$$w > 0 \quad \text{on } \partial D_R \cap \{\text{dist}(\cdot, \partial_1) \leq \delta\} \cap \mathcal{A}, \tag{6.13}$$

then by applying the maximum principle to (6.10)–(6.13) we will conclude

$$w > 0 \quad \text{in } D_R. \tag{6.14}$$

We show (6.13) by contradiction. Assume (6.13) is false and let $\bar{P} \in \partial D_R \cap \mathcal{A}$ be such that

$$w(\bar{P}) \leq 0 \quad \text{and} \quad \text{dist}(\bar{P}, \partial_1) \leq \delta. \tag{6.15}$$

Now set

$$\tilde{w}(P) = w(P) + \nu|P - \bar{P}|^2, \qquad P \in \text{cl}(D_{2R}),$$

with $\nu > 0$ small enough to guarantee

$$A(\nu|P - \bar{P}|^2) > -\varepsilon$$

for ε as in (6.10) (this can be done because of (5.1), i.e., $\sum(\sigma\sigma^*)_{ij}\xi_i\xi_j \geq \alpha|\xi|^2$, for every $\xi \in \mathbb{R}^2$, with $\alpha > 0$). Thus,

$$A\tilde{w} > 0 \quad \text{in } D_{2R}, \tag{6.16}$$

$$\tilde{w} \geq 0 \quad \text{on } \partial D_{2R} \cap \partial_1, \tag{6.17}$$

and

$$\tilde{w} > 0 \quad \text{on } \partial D_{2R} \cap \{\text{dist}(\cdot, \partial_1) > \delta\} \tag{6.18}$$

as this follows from (6.12). Therefore, if we show

$$\tilde{w} > 0 \quad \text{on } \partial D_{2R} \cap \{\text{dist}(\cdot, \partial_1) \leq \delta\} \cap \mathcal{A}, \tag{6.19}$$

then by applying the strong maximum principle to (6.16)–(6.19) we will get

$$\tilde{w} > 0 \quad \text{in } D_{2R},$$

which contradicts

$$\tilde{w}(\bar{P}) = w(\bar{P}) \leq 0,$$

and hence we will have proved (6.13). So let us show that (6.19) holds. Let $P \in \partial D_{2R} \cap \{\text{dist}(\cdot, \partial_1) \leq \delta\} \cap \mathcal{A}$, and recall that

$$\hat{u}_{x_1} \in C^{1,1}(\Omega), \qquad \hat{u}_{x_1} = 0 \text{ on } \partial_1, \qquad \nabla \hat{u}_{x_1} = 0 \text{ on } \partial_1,$$

by (5.25) and (5.10); then, by Lemma 6.1 follows

$$\hat{u}_{x_1}(P) \leq C\delta^2 \quad \text{and} \quad |\nabla \hat{u}_{x_1}(P)| \leq C\delta.$$

Hence

$$\begin{aligned} \tilde{w}(P) &= K\hat{u}_{x_1x_1}(P) + H\hat{u}_{x_1x_2}(P) - F\hat{u}_{x_1}(P) + \nu|P - \bar{P}|^2 \\ &\geq K\hat{u}_{x_1x_1}(P) - HC\delta - FC\delta^2 + \nu|P - \bar{P}|^2 \\ &\geq K\hat{u}_{x_1x_1}(P) - C\delta(1 + F) + \nu R^2 \end{aligned}$$

since $\delta < 1$ and $|P - \bar{P}| \geq R$. Then, certainly $\tilde{w}(P) > 0$ if we choose $\delta < \nu R^2/[C(1 + F)]$; so (6.19) holds, and hence so does (6.13). Thus (6.14) holds true too, i.e.,

$$w > 0 \quad \text{in } D_R,$$

i.e.,

$$K\hat{u}_{x_1x_1} + H\hat{u}_{x_1x_2} - F\hat{u}_{x_1} > 0 \quad \text{in } D_R;$$

but $\hat{u}_{x_1} > 0$ in $\mathcal{A}$; hence we have

$$K\hat{u}_{x_1x_1} + H\hat{u}_{x_1x_2} > 0 \quad \text{in } D_R, \tag{6.20}$$

which means that $\hat{u}_{x_1}$ increases along lines of slope H/K in D_R, for any H such that $|H| \leq 1$. However, $\hat{u}_{x_1} = 0$ on ∂_1, therefore there exists a cone $\gamma^+(Q)$ with vertex Q, and angle $2\beta = 2\arctan(1/K)$, and with axis parallel to the x_1-axis such that

$$\gamma^+(Q) \cap D_R \subset \mathcal{A}.$$

The same holds for any $P \in \partial_1 \cap V_Q$, V_Q being a small neighborhood of Q, i.e.,

$$\gamma^+(P) \cap V_Q \subset \mathcal{A}, \qquad \forall\, P \in V_Q \cap \partial_1. \tag{6.21}$$

But we do know that R_1 is the subgraph of $\psi_1(x_2)$; hence it must also be

$$\gamma^-(P) \cap V_Q \subset R_1, \qquad \forall\, P \in V_Q \cap \partial_1, \tag{6.22}$$

if

$$\gamma^-(P) := \{(x_1, x_2) : (-x_1, x_2) \in \gamma^+(P)\}.$$

Thus ψ_1 is Lipschitz continuous in $\pi_2(V_Q)$ (with π_2 being the projection onto the x_2-axis); in fact,

$$\frac{|\psi_1(x_2)-\psi_1(y_2)|}{|x_2-y_2|} \le \max\left\{\tan\beta, \tan\left(\frac{\pi}{2}-\beta\right)\right\} = \max\{1/K, K\} \tag{6.23}$$

for every $x_2, y_2 \in \pi_2(V_Q)$. □

Finally, from the Lipschitz continuity follows even greater regularity by the application of classical regularity results (cf. [6, Thm. 6.2, p. 179]).

THEOREM 6.3. *Assume* (5.1) *and* $f \in C^3(\mathbb{R}^2)$. *Then, for* $i = 1,2, j \neq i$, *and with* $\{(x_1^0, x_2^0)\} = \partial_0$,

(i) $\psi_i \in C^1((x_j^0, +\infty))$ *and* $\hat{u}_{x_i} \in C^2(\mathcal{A} \cup (\partial_i \backslash \partial_0))$;

(ii) $f_{x_i} \in C^{m,\alpha}(\{f_{x_i} < 0\})\, (m \in \mathbb{N}, 0 < \alpha < 1) \Rightarrow \psi_i \in C^{m+1,\alpha}((x_j^0, +\infty))$;

(iii) f_{x_i} *is analytic in* $\{f_{x_i} < 0\} \Rightarrow \psi_i$ *is analytic in* $(x_j^0, +\infty)$.

Note that (i) is a consequence of a fundamental result of Caffarelli, which guarantees C^1-regularity of the boundary in a neighborhood of any point of positive Lebesgue density for the coincidence set (cf. [6, Thm. 3.10, p. 162]).

COROLLARY 6.4. *Assume* (5.1) *and* $f \in C^3(\mathbb{R}^2)$. *Then, for* $i = 1,2, j \neq i$, *and with* $\{(x_1^0, x_2^0)\} = \partial_0$,

(i) $\psi_i \in C^0((-\infty, +\infty)) \cap C^{2,\alpha}((x_j^0, +\infty))$;

(ii) $\hat{u} \in C^3(\mathcal{A} \cup (\partial\mathcal{A} \backslash \partial_0))$.

Proof. This follows from $f_{x_i} \in C^2$ and (i) and (ii) of Theorem 6.3, together with the established continuity of ψ_i at the corner point ∂_0 (cf. Lemma 3.8). □

The local finiteness of the perimeter of the region of inaction $\mathcal{A}$ provides us with a lot of information about its boundary. This is a property of geometric character and has never been used before to study the region of inaction, $\mathcal{A}$, of a singular control problem in the setting of the *monotone follower problem*. As §4 shows, this approach turns out to be extremely powerful in the two-dimensional case; in fact, the functions ψ_i being functions of a real variable and having obtained $\psi_i \in BV_{\text{loc}}(\mathbb{R})$, the useful characterization (c) of Remark 4.3 is available for ψ_i. This characterization allows the detailed analysis of the boundary that leads to its regularity. In more than two dimensions there are other characterizations of $BV_{\text{loc}}(\mathbb{R}^{n-1})$ which, however, are not as easy to apply and to work with; nevertheless they still are a useful tool to study the regularity of $\partial\tilde{\mathcal{A}}$ as we show in forthcoming work.

7. Monotonicity of ψ_i. We can examine the monotonicity properties of ψ_i under the strengthened conditions in [15], i.e.,

$$\sigma\sigma^* \text{ is positive definite;} \tag{7.1}$$

$$f \in C^3(\mathbb{R}^2). \tag{7.2}$$

Remark 7.1. Note that in Theorem 6.2 we showed that if (7.1) and (7.3) hold, then ψ_i is continuous everywhere and locally Lipschitz away from the corner point ∂_0 (and hence $\psi_i = \tilde{\psi}_i$).

Then we have the following proposition.

PROPOSITION 7.2. *Assume* (7.1) *and* (7.2). *If*

$$f_{x_1x_2} \ge 0 \quad \text{in } \mathbb{R}^2,$$

then ψ_i *is nonincreasing.*

Proof. We denote by v^ε the solution of the penalized problem (3.15) of [15], then $v^\varepsilon \to \hat{u}$, $v^\varepsilon_{x_i} \to \hat{u}_{x_i}$ uniformly on compacta. Moreover if $w = v^\varepsilon_{x_1 x_2}$, then

$$A_0 w = f_{x_1 x_2}, \tag{7.3}$$

where A_0 is a differential operator that differs from A only by first and 0th order terms. As $f_{x_1 x_2} \geq 0$, it follows from a version of the maximum principle for functions of $C^2(\mathbb{R}^2)$ bounded above by a polynomial that $w = v^\varepsilon_{x_1 x_2} \geq 0$.

This implies, for example, that $v^\varepsilon_{x_2}$ is nondecreasing along horizontal lines. Then, by using the fact that $v^\varepsilon_{x_2} \to \hat{u}_{x_2}$ uniformly on compacta as $\varepsilon \to 0$, we conclude that if $P \in \mathcal{A}$, then the horizontal ray to the right of P is also in $\mathcal{A}$. It follows that ψ_2 is nonincreasing. □

The following corollary is obtained by replacing w by $-w$ in the above proof.

COROLLARY 7.3. *Assume* (7.1) and (7.2), *and*

$$f_{x_1 x_2} \leq 0 \quad \text{in } \mathbb{R}^2. \tag{7.4}$$

Then ψ_i is nondecreasing $(i = 1, 2)$.

A particular example of interest is

$$J_x(k) = E\left\{\int_0^\infty f(X_t^k) e^{-t}\, dt\right\},$$

$$f(x_1, x_2) = x_1^2 + \alpha x_1 x_2 + x_2^2,$$

$$X_t = x + \sqrt{2} W_t + k_t$$

with $\alpha \in (-2, 0)$. We show that here ψ_i cannot be nonincreasing. Suppose to the contrary; then by the above corollary, $\psi_i = c_i$, a constant. Symmetry considerations allow us to conclude that $c_1 = c_2 = c$. We now attempt to solve

$$-\Delta u + u = f, \qquad x > c, y > c$$

$$u_x(c, y) = 0 = u_y(x, c), \qquad x > c, y > c.$$

Shifting the origin to (c, c), changing to polar coordinates, and applying Sturm–Liouville theory produces a solution of the form

$$u(x, y) = \sum_{k=0}^{\infty} a_k(r) \cos(4k\theta),$$

where (r, θ) is the polar form of $(x - c, y - c)$ and

$$a_k'' + \frac{a_k'}{r} - \left[\frac{(4k)^2}{r^2} + 1\right] a_k = g_k$$

$$g_0 = -\left(1 + \frac{\alpha}{\pi}\right) r^2 - \frac{4}{\pi}(2 + \alpha) cr - (2 + \alpha) c^2$$

$$g_k = \frac{2\alpha r^2}{\pi[(2k)^2 - 1]} + \frac{8(2 + \alpha) cr}{\pi[(4k)^2 - 1]}, \qquad k = 1, 2, \ldots.$$

Let us write this as $g_k = \alpha_0^k + \alpha_1^k r + \alpha_2^k r^2$. The solution a_k^h of the homogeneous equation is given in terms of modified Bessel functions and is not admissible if we insist that a_k be finite at $r = 0$ and that a_k'' be finite at $r = \infty$ (cf. (2.10)). A series solution of the nonhomogeneous equation turns out to have the form

$$a_k(r) = \sum_{n=2}^{\infty} b_{k,n} r^n$$

where

$$b_{k,2} = \frac{\alpha_0^k}{2^2 - (4k)^2}, \qquad b_{k,3} = \frac{\alpha_1^k}{3^2 - (4k)^2},$$

$$b_{k,4} = \frac{\alpha_2^k + \frac{\alpha_0^k}{2^2-(4k)^2}}{4^2 - (4k)^2},$$

$$b_{k,n} = \frac{b_{k,n-2}}{n^2 - (4k)^2} \qquad n = 5, 6, \ldots.$$

The difficulty now lies in that the denominators may be zero. This never occurs for $k = 0$, but when $k > 0$ and $n = 4k$, then the recurrence relationship cannot be solved unless the corresponding $b_{k,n}$ are zero, i.e., $b_{k,n} = 0$ for n even! But this only occurs if $\alpha_0^k = 0 = \alpha_2^k$, i.e., $\alpha = 0$ no matter what c is. In this case the variational inequality decomposes into two one-dimensional problems with solution

$$u(x, y) = x^2 + y^2 + 4 - 2(e^{1-x} + e^{1-y}), \qquad c = -1.$$

For non-zero α it follows that a solution of the variational inequality fails to exist no matter what c is if we assume that $\mathcal{A}$ is a quadrant. This shows that there are examples where the ψ_i are not nonincreasing.

A. Appendix.

Proof of Theorem 3.1. We recall that $\hat{u} \in C^1(\mathbb{R}^2)$ (cf. Theorem 2.3).

CLAIM 1. *For every x_2, there exists $\bar{x}_1$ such that for every $x_1 \geq \bar{x}_1$, $\hat{u}_{x_1}(x_1, x_2) > 0$.*

With $\bar{\bar{x}}_1 \in \mathbb{R}$ and $c > 0$ (both to be fixed later), we define the following function:

$$w(x_1, x_2) = \begin{cases} 0 & \text{if } x_1 \leq \bar{\bar{x}}_1, \\ c(x_1 - \bar{\bar{x}}_1) & \text{if } x_1 \geq \bar{\bar{x}}_1. \end{cases}$$

Since $Aw(x_1, x_2) = -g_1 c + \rho c(x_1 - \bar{\bar{x}}_1)$ for $x_1 \geq \bar{\bar{x}}_1$, we can choose $c > 0$ and $\bar{\bar{x}}_1 \in \mathbb{R}$ such that

$$Aw(x_1, x_2) \leq f(x_1, x_2) \quad \text{if } x_1 \geq \bar{\bar{x}}_1.$$

(This can be done because the assumed polynomial growth condition $(2.4)_1$ implies $r|(x_1, x_2)^+|^p - C_0 \leq f(x_1, x_2)$ with C_0 and r independent of x_2.) Since $f \geq 0$, then w is a solution of the quasi-variational inequality (2.42), whose maximal solution is $\hat{u}$ (cf. Theorem 2.9); so we conclude that $w \leq \hat{u}$. Therefore, since $\hat{u}_{x_1} \geq 0$, for every x_2 there exists some point $\bar{x}_1 \geq \bar{\bar{x}}_1$ such that $\hat{u}_{x_1}(\bar{x}_1, x_2) > 0$; from this we now deduce Claim 1 (since $\hat{u}$ is convex).

Similarly, we have that for every x_1 there exists $\bar{x}_2$ such that for every $x_2 \geq \bar{x}_2$, $\hat{u}_{x_2}(x_1, x_2) > 0$.

CLAIM 2. *For every $\tilde{x}_1$, $\hat{u}_{x_i}(\tilde{x}_1, x_2)$ remains bounded as $x_2 \to -\infty$, $(i = 1, 2)$.*

In fact, $\hat{u}_{x_2}(\tilde{x}_1, \cdot)$ is nondecreasing (by convexity) and ≥ 0 (by Theorem 2.11). So $\hat{u}_{x_2}(\tilde{x}_1, x_2) \to \infty$ as $x_2 \to -\infty$ would imply $\hat{u}_{x_2}(\tilde{x}_1, \cdot) \equiv +\infty$, which is impossible. On the other hand, since $\hat{u}$ is convex and non-negative, we have

$$\begin{aligned} \hat{u}(\tilde{x}_1 + h, x_2) &\geq \hat{u}(\tilde{x}_1, x_2) + h\hat{u}_{x_1}(\tilde{x}_1, x_2) \\ &\geq h\hat{u}_{x_1}(\tilde{x}_1, x_2) \quad \text{for every } x_2. \end{aligned}$$

Therefore, $\hat{u}_{x_1}(\tilde{x}_1, x_2) \to +\infty$ as $x_2 \to -\infty$ would imply (for $h > 0$) $\hat{u}(\tilde{x}_1 + h, x_2) \to +\infty$ as $x_2 \to -\infty$, so $\hat{u}(\tilde{x}_1 + h, \cdot) \equiv +\infty$ (since $\hat{u}_{x_2} \geq 0$). This is impossible because of the polynomial growth of $\hat{u}$.

CLAIM 3. *For every x_1 there exists $\tilde{x}_2$ such that $\hat{u}_{x_2}(x_1, \tilde{x}_2) = 0$.*

In fact, if not, then we have that there exists $\tilde{x}_1$ such that $\hat{u}_{x_2}(\tilde{x}_1, x_2) > 0$ for all x_2. So one of the following two cases occurs:

$$\text{Case (a):} \qquad \exists x_{2,n} \to -\infty \quad \text{such that } \hat{u}_{x_1}(\tilde{x}_1, x_{2,n}) > 0;$$

or

$$\text{Case (b):} \qquad \exists \tilde{x}_2 \quad \text{such that } \hat{u}_{x_1}(\tilde{x}_1, x_2) = 0 \quad \forall\, x_2 \leq \tilde{x}_2.$$

In Case (a), by continuity of $\hat{u}_{x_1}$ and $\hat{u}_{x_2}$, we have $\hat{u}_{x_1} > 0$ and $\hat{u}_{x_2} > 0$ on some open set U_n containing $(\tilde{x}_1, x_{2,n})$. Then, by Theorem 2.11, $A\hat{u} = f$ almost everywhere in U_n, i.e.,

$$-\frac{1}{2} \operatorname{tr}\left[\sigma\sigma^* D^2\hat{u}\right] - g \cdot \nabla\hat{u} + \rho\hat{u} = f \quad \text{a.e. in } U_n. \tag{A.1}$$

From (A.1) and $\operatorname{tr}\left[\sigma\sigma^* D^2\hat{u}\right] \geq 0$ almost everywhere (since $\hat{u}$ is convex) follows

$$\rho\hat{u} \geq f + g \cdot \nabla\hat{u} \quad \text{in } U_n, \tag{A.2}$$

so

$$\rho\hat{u}(\tilde{x}_1, x_2) \geq f(\tilde{x}_1, x_2) + g_1\hat{u}_{x_1}(\tilde{x}_1, x_2) + g_2\hat{u}_{x_2}(\tilde{x}_1, x_2) \quad \text{with } (\tilde{x}_1, x_2) \in U_n, \tag{A.3}$$

and in particular

$$\rho\hat{u}(\tilde{x}_1, x_{2,n}) \geq f(\tilde{x}_1, x_{2,n}) + g_1\hat{u}_{x_1}(\tilde{x}_1, x_{2,n}) + g_2\hat{u}_{x_2}(\tilde{x}_1, x_{2,n}). \tag{A.4}$$

Because we are assuming $f(x_1, x_2) \to +\infty$ as $|(x_1, x_2)| \to +\infty$, we have $f(\tilde{x}_1, x_{2,n}) \to +\infty$ as $n \to +\infty$, while $\hat{u}_{x_1}(\tilde{x}_1, x_{2,n})$ and $\hat{u}_{x_2}(\tilde{x}_1, x_{2,n})$ stay bounded as $n \to +\infty$ (by Claim 2). Then, from (A.4) we have

$$\rho\hat{u}(\tilde{x}_1, x_{2,n}) \to +\infty \quad \text{as } n \to +\infty,$$

but $\hat{u}(\tilde{x}_1, \cdot)$ is nondecreasing; therefore it must be $\hat{u}(\tilde{x}_1, \cdot) \equiv +\infty$, and this contradicts the polynomial growth of $\hat{u}$. So Case (a) cannot occur.

Let's now assume that Case (b) holds. Then, because of convexity and $\hat{u}_{x_1} \geq 0$, we have

$$\hat{u}_{x_1}(x_1, x_2) = 0 \quad \text{for every } (x_1, x_2) \in (\tilde{x}_1, \tilde{x}_2) - \Lambda^*. \tag{A.5}$$

Now Claim 1 allows us to define, for $x_2 \leq \tilde{x}_2$,

$$\psi_1(x_2) = \max\{x_1 : \hat{u}_{x_1}(x_1, x_2) = 0\} = \inf\{x_1 : \hat{u}_{x_1}(x_1, x_2) > 0\}.$$

Let's now define the region $\mathcal{T} = \operatorname{int}(\{(x_1, x_2) : x_1 \geq \psi_1(x_2), x_2 \leq \tilde{x}_2\})$. From the definition of ψ_1 follows that $\hat{u}_{x_1} \neq 0$ in $\mathcal{T}$. Moreover, by assumption $\hat{u}_{x_2} \neq 0$ on the line $x_1 = \tilde{x}_1$; also,

on the left of ψ_1, $\hat{u}$ is constant along horizontal line segments and therefore $\hat{u}_{x_2}$ is constant along horizontal line segments. In conclusion, $\hat{u}_{x_2}(\psi_1(x_2), x_2) = \hat{u}_{x_2}(\tilde{x}_1, x_2) > 0$ for $x_2 \leq \tilde{x}_2$. So by continuity of $\hat{u}_{x_2}$ we have $\hat{u}_{x_2} \neq 0$ in $U_{\psi_1} \cap \mathcal{T}$, U_{ψ_1} being a neighborhood of $\partial\mathcal{T}$. Then, $\hat{u}_{x_1} \neq 0$ and $\hat{u}_{x_2} \neq 0$ in $U_{\psi_1} \cap \mathcal{T}$. Now Theorem 2.11 implies that the dynamic programming equation holds in $U_{\psi_1} \cap \mathcal{T}$, i.e., $-\frac{1}{2}\text{tr}\,[\sigma\sigma^* D^2\hat{u}] - g \cdot \nabla\hat{u} + \rho\hat{u} = f$ almost everywhere in $U_{\psi_1} \cap \mathcal{T}$. We know tr $[\sigma\sigma^* D^2\hat{u}] \geq 0$ almost everywhere (since $\hat{u}$ is convex), so we have

$$\rho\hat{u}(\tilde{z}, x_2) \geq f(\tilde{z}, x_2) + g_1\hat{u}_{x_1}(\tilde{z}, x_2) + g_2\hat{u}_{x_2}(\tilde{z}, x_2) \quad \text{if } (\tilde{z}, x_2) \in U_{\psi_1} \cap \mathcal{T}.$$

It follows from the continuity of $\hat{u}_{x_1}$ and the definition of ψ_1 that $\overline{\lim}_{t \to x_2} \psi_1(t) \leq \psi_1(x_2)$; hence $(\psi_1(x_2), x_2) \in \partial\mathcal{T}$ for any $x_2 < \tilde{x}_2$. Therefore, we can take $\lim_{\substack{\tilde{z} \to \psi_1(x_2) \\ (\tilde{z}, x_2) \in \mathcal{T} \cap U_{\psi_1}}}$ and obtain (by continuity)

$$\rho\hat{u}(\psi_1(x_2), x_2) \geq f(\psi_1(x_2), x_2) + g_2\hat{u}_{x_2}(\psi_1(x_2), x_2), \quad \text{with } x_2 < \tilde{x}_2. \tag{A.6}$$

Also, $\hat{u}(\tilde{x}_1, x_2) = \hat{u}(\psi_1(x_2), x_2)$ (by the definition of ψ_1, since (A.5) implies $\tilde{x}_1 \leq \psi_1(x_2)$ for every $x_2 \leq \tilde{x}_2$); $\hat{u}_{x_2}(\tilde{x}_1, \tilde{x}_2) \geq \hat{u}_{x_2}(\tilde{x}_1, x_2)$ (by convexity and $x_2 \leq \tilde{x}_2$); finally, $\hat{u}_{x_2}(\psi_1(x_2), x_2) = \hat{u}_{x_2}(\hat{x}_1, x_2)$ (as we observed above). Then, from (A.6) follows

$$\begin{aligned} \rho\hat{u}(\tilde{x}_1, x_2) &\geq f(\psi_1(x_2), x_2) + g_2\hat{u}_{x_2}(\tilde{x}_1, x_2) \\ &\geq f(\psi_1(x_2), x_2) - |g_2|\hat{u}_{x_2}(\tilde{x}_1, \tilde{x}_2). \end{aligned}$$

Using the fact that

$$f(x_1, x_2) \to +\infty \quad \text{as } |(x_1, x_2)| \to +\infty,$$

we obtain

$$\hat{u}(\tilde{x}_1, x_2) \to +\infty \quad \text{as } x_2 \to -\infty,$$

which is impossible since $\hat{u}(\tilde{x}_1, x_2)$ decreases as $x_2 \to -\infty$ (as $\hat{u}_{x_2} > 0$). Therefore, Case (b) cannot occur and Claim 3 is finally proved.

Similarly we have that for every x_2 there exists $\tilde{x}_1$ such that $\hat{u}_{x_1}(\tilde{x}_1, x_2) = 0$. Then, the functions ψ_i in (3.3) are well defined. □

Proof of Lemma 4.15. Let us fix $i = 1$ for simplicity. Then,

$$\begin{aligned} \partial_1 \backslash \tilde{\partial}_1 &= (\partial R_1 \cap \partial\mathcal{A}) \backslash (\partial R_1 \cap \partial\tilde{\mathcal{A}}) \\ &= \partial R_1 \cap (\partial\mathcal{A} \backslash \partial\tilde{\mathcal{A}}) \\ &= \partial R_1 \cap ([\partial\mathcal{A} \cap (\text{cl}(\tilde{\mathcal{A}}))^c] \cup [\partial\mathcal{A} \cap (\text{cl}((\tilde{\mathcal{A}})^c))^c]), \end{aligned}$$

but $\tilde{\mathcal{A}}$ is open, so $(\text{cl}((\tilde{\mathcal{A}})^c))^c = ((\tilde{\mathcal{A}})^c)^c = \tilde{\mathcal{A}}$ and we have

$$\partial_1 \backslash \tilde{\partial}_1 = [\partial R_1 \cap \partial\mathcal{A} \cap (\text{cl}(\tilde{\mathcal{A}}))^c] \cup [\partial R_1 \cap \partial\mathcal{A} \cap \tilde{\mathcal{A}}].$$

Clearly, $\text{cl}(\tilde{\mathcal{A}}) \subset \text{cl}(\mathcal{A})$; on the other hand, since $\mathcal{A}$ is open, we have $\mathcal{A} = \text{int}(\mathcal{A}) \subset \text{int}(\text{cl}(\mathcal{A})) = \tilde{\mathcal{A}}$ and so also $\text{cl}(\mathcal{A}) \subset \text{cl}(\tilde{\mathcal{A}})$. Therefore,

$$\text{cl}(\mathcal{A}) = \text{cl}(\tilde{\mathcal{A}});$$

hence

$$\partial\mathcal{A} \cap (\text{cl}(\tilde{\mathcal{A}}))^c = \partial\mathcal{A} \cap (\text{cl}(\mathcal{A}))^c = \emptyset,$$

and we have

$$\partial_1 \backslash \tilde{\partial}_1 = \partial R_1 \cap \partial \mathcal{A} \cap \tilde{\mathcal{A}} = R_1 \cap \tilde{\mathcal{A}} \tag{A.7}$$

since R_1 is closed.

Finally we recall that $\partial_0 = \{P_0\}$ and $R_0 = P_0 - \Lambda^*$ (cf. Proposition 3.6, (ii)); hence

$$\forall r > 0: \quad |B(P_0, r) \cap R_0| = \frac{\pi r^2}{4};$$

that is,

$$P_0 \notin \text{int}(\text{cl}(\mathcal{A})) = \tilde{\mathcal{A}},$$

but $\{P_0\} = \partial_0 \subset \partial \mathcal{A}$, so it must be $P_0 \in \tilde{\partial}$. □

Proof of Lemma 5.6. Let $\Omega' \subset \text{cl}(\Omega') \subset \Omega$ be fixed; let $\gamma \in C_0^\infty(\Omega)$ be as in the statement of the lemma. Then, (5.8) implies

$$a(w, \gamma(\tilde{v} - \gamma w)) \geq \int_\Omega f_{x_i} \gamma(\tilde{v} - \gamma w)\, dx \tag{A.8}$$

with $\eta = \gamma^2$ and

$$\tilde{v} := \gamma v \quad \text{with } v \in \mathbb{K}(\Omega),$$

so that $\tilde{v} \in \mathbb{K}_0(\Omega)$. Now we calculate

$$\begin{aligned} a(\gamma w, \tilde{v} - \gamma w) =& \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} (\gamma w)_{x_i} (\tilde{v} - \gamma w)_{x_j}\, dx \\ &- \int_\Omega \sum_{i=1}^{2} g_i (\gamma w)_{x_i} (\tilde{v} - \gamma w)\, dx \\ &+ \int_\Omega \rho \gamma w (\tilde{v} - \gamma w)\, dx \\ =& \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} \gamma w_{x_i} (\tilde{v} - \gamma w)_{x_j}\, dx \\ &- \int_\Omega \sum_{i=1}^{2} g_i \gamma w_{x_i} (\tilde{v} - \gamma w)\, dx \\ &+ \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} \gamma_{x_i} w (\tilde{v} - \gamma w)_{x_j}\, dx \\ &- \int_\Omega \sum_{i=1}^{2} g_i \gamma_{x_i} w (\tilde{v} - \gamma w)\, dx + \int_\Omega \rho \gamma w (\tilde{v} - \gamma w)\, dx. \end{aligned} \tag{A.9}$$

On the other hand,

$$\begin{aligned} a(w, \gamma(\tilde{v} - \gamma w)) =& \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} w_{x_i} \{\gamma(\tilde{v} - \gamma w)_{x_j} + \gamma_{x_j} (\tilde{v} - \gamma w)\}\, dx \\ &- \int_\Omega \sum_{i=1}^{2} g_i w_{x_i} \gamma (\tilde{v} - \gamma w)\, dx + \int_\Omega \rho \gamma w (\tilde{v} - \gamma w)\, dx; \end{aligned} \tag{A.10}$$

hence (A.9) and (A.10) imply

$$
\begin{aligned}
a(\gamma w, \tilde{v} - \gamma w) =& a(w, \gamma(\tilde{v} - \gamma w)) - \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} w_{x_i} \gamma_{x_j} (\tilde{v} - \gamma w)\, dx \\
& - \int_\Omega \rho w \gamma (\tilde{v} - \gamma w)\, dx + \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} \gamma_{x_i} w (\tilde{v} - \gamma w)_{x_j}\, dx \\
& - \int_\Omega \sum_{i=1}^{2} g_i \gamma_{x_i} w (\tilde{v} - \gamma w)\, dx + \int_\Omega \rho \gamma w (\tilde{v} - \gamma w)\, dx,
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
a(\gamma w, \tilde{v} - \gamma w) =& \, a(w, \gamma(\tilde{v} - \gamma w)) - \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} w_{x_i} \gamma_{x_j} (\tilde{v} - \gamma w)\, dx \\
& - \int_\Omega \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} \{\gamma_{x_i x_j} w + \gamma_{x_i} w_{x_j}\} (\tilde{v} - \gamma w)\, dx \\
& - \int_\Omega \sum_{i=1}^{2} g_i \gamma_{x_i} w (\tilde{v} - \gamma w)\, dx \\
=& \, a(w, \gamma(\tilde{v} - \gamma w)) \\
& - \int_\Omega \left\{ \frac{1}{2} \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} \gamma_{x_i x_j} w + \sum_{i,j=1}^{2} (\sigma\sigma^*)_{ij} \gamma_{x_i} w_{x_j} \right\} (\tilde{v} - \gamma w)\, dx \\
& - \int_\Omega \sum_{i=1}^{2} g_i \gamma_{x_i} w (\tilde{v} - \gamma w)\, dx.
\end{aligned}
\tag{A.11}
$$

Now from (A.8) and (A.11) we obtain (5.11) for $\tilde{v} \in \mathbb{K}_0(\Omega)$ of the form $\tilde{v} = \gamma v$ with $v \in \mathbb{K}(\Omega)$. Finally, let v be any element in $\mathbb{K}_0(\Omega)$, let $\Omega' = \operatorname{supp} v$, then (5.11) holds for v since $v = \gamma v$ and $\mathbb{K}_0(\Omega) \subset \mathbb{K}(\Omega)$. □

REFERENCES

[1] H. Brezis and D. Kinderlehrer, *The smoothness of solutions to nonlinear variational inequalities*, Indiana Univ. Math. J., 23 (1974), pp. 831–844.

[2] L. A. Caffarelli, *The regularity of free boundaries in higher dimensions*, Acta Math., 139 (1977), pp. 155–184.

[3] M. B. Chiarolla, *Geometric approach to monotone stochastic control*, Ph.D. dissertation, University of British Columbia, Vancouver, British Columbia, Canada, 1992.

[4] M. B. Chiarolla and U. G. Haussmann, *The optimal control of the cheap monotone follower*, Stochastics, to appear.

[5] P.-L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[6] A. Friedman, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[7] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[8] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, 1984.

[9] I. Karatzas and S. E. Shreve, *Connection between optimal stopping and singular stochastic control* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.

[10] P. L. Lions and A. S. Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1986), pp. 511–537.

[11] J.-L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.
[12] M. MIRANDA, *Superfici cartesiane generalizzate ed insiemi di perimetro localmente finito sui prodotti cartesiani*, Ann. Sc. Norm. Sup. Pisa, 18 (1964), pp. 515–542.
[13] H. M. SONER AND S. E. SHREVE, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.
[14] L. TONELLI, *Sul differenziale dell'arco di curve*, Atti Accad. Lincei, 25 (1916), pp. 207–213.
[15] S. A. WILLIAMS, P.-L. CHOW, AND J.-L. MENALDI, *Regularity of the free boundary in singular stochastic control*, J. Differential Equations, to appear.
[16] W. P. ZIEMER, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 728–743, May 1994

007

GENERALIZED SOLUTIONS OF THE HAMILTON–JACOBI EQUATION OF STOCHASTIC CONTROL*

ULRICH G. HAUSSMANN†

Abstract. A second-order generalized derivative based on Brownian motion is introduced. Using this derivative, an Itô-type formula is derived for functions $f(t,x)$, which are continuously differentiable in x with Lipschitz derivative and are Lipschitz continuous in t. It is then shown that the value function of a stochastic control problem is a "generalized" solution of a second-order Hamilton–Jacobi equation. Such solutions are analogous to the Clarke generalized solutions of first-order Hamilton–Jacobi equations. Finally, it is shown that any "generalized" solution is a viscosity subsolution and a viscosity solution is a "generalized" solution.

Key words. stochastic control, Hamilton–Jacobi equation (H–J), generalized solution, viscosity solution, Itô's formula

AMS subject classifications. 93E20, 49L25, 49J52

1. Introduction.

We consider the control problem:

$$X_s^{t,x,u} = x + \int_t^s b(r, X_r^{t,x,u}, u_r)\,dr + \int_t^s \sigma(r, X_r^{t,x,u}, u_r)\,dw_r \tag{1}$$

with cost

$$J^{t,x}(u) = \mathbb{E}\left\{\int_t^T f(r, X_r^{t,x,u}, u_r)\,dr + f_o(X_T^{t,x,u})\right\}. \tag{2}$$

Here $\{w_r : r \geq 0\}$ is a standard Brownian motion on a probability space $(\Omega, \mathcal{F}, P)$ with filtration $\{\mathcal{F}_t\}_{0\leq t\leq T}$ and $u \in \mathcal{U}$ is a control. The corresponding value function is

$$v(t,x) = \inf_{u\in\mathcal{U}} J^{t,x}(u), \tag{3}$$

where $\mathcal{U}$ is the set of U-valued progressively measurable stochastic process on $[0,T)$. If v is smooth, it is the unique (in the class of functions satisfying $v(T,x) = f_o(x)$) classical solution of the Hamilton–Jacobi equation (H–J), i.e.,

$$-v_t(t,x) + H(t,x,-v_x(t,x),-v_{x,x}(t,x)) = 0,$$

where the subscripts denote partial derivatives and where H is the Hamiltonian for the problem and is defined as

$$H(t,x,p,P) \stackrel{\text{def}}{=} \sup_{u\in U}\{\text{trace}[P\tfrac{1}{2}\sigma(t,x,u)\sigma^T(t,x,u)] + p\cdot b(t,x,u) - f(t,x,u)\}.$$

This situation occurs rarely, but it is true that v is always a weak (in the sense of distributions) solution, although the latter are not unique. On the other hand, if v is continuous, then it is the unique viscosity solution of H–J. In deterministic control theory (i.e., $\sigma = 0$) where H–J is a first order equation the same situation prevails, but another notion of solution has been found to be useful, i.e., the generalized solution of Clarke. The relationship between viscosity solutions and generalized solutions in the first order case has been analysed by Frankowska [3]. In this article we define a generalized solution of the second order H–J; we show that the

* Received by the editors March 19, 1992; accepted for publication (in revised form) December 9, 1992. This work was supported by Natural Sciences and Engineering Research Council of Canada grant 8051.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, V6T 1Z2 Canada.

value function is such a solution and we look at the relationship between generalized solutions and viscosity solutions.

The standing assumptions are as follows:

1. b, σ are bounded, continuous on $[0,T] \times \mathbf{R}^d \times U$ and Lipschitz continuous in (t,x) uniformly in u, $\sigma(t,x,u)$ is a $d \times m$ dimensional matrix;

2. f is continuous, Lipschitz continuous in (t,x) uniformly in u and has at most polynomial growth in x;

3. f_o is continuous;

4. U is a compact metric space.

In §2 we introduce and discuss the second-order generalized derivative; in the following section we extend the Itô formula to functions with bounded generalized second derivatives for processes that are degenerate diffusions. This result is used in the sequel but may be of independent interest. In §4 we define the notion of a generalized solution of (H–J) and we show that the value function is a generalized solution. In the last section we discuss the relationship between the generalized solutions and viscosity solutions. An appendix containing some remarks about generalized second-order derivatives concludes the paper.

2. Generalized derivatives. The first-order generalized directional derivative of a function f at $x \in \mathbf{R}^d$ in the direction $v \in \mathbf{R}^d$, $f^o(x;v)$, is defined (at least if f is Lipschitz) by Clarke [1] as

$$f^o(x;v) \stackrel{\text{def}}{=} \limsup_{\substack{y \to x \\ t \to 0+}} \frac{f(y+tv) - f(y)}{t},$$

and the generalized gradient at $x, \partial f(x)$, is the closed convex set in the dual space $(\mathbf{R}^d)^*$ whose support functional is $v \to f^o(x;v)$. The set is nonempty, compact if f is Lipschitz continuous. The fact that $f^o(x;\cdot)$ is a support functional follows from the subadditivity and positive homogeneity of the map. The second-order generalized directional derivative at x has been defined by Cominetti and Correa [2] as a functional on $\mathbf{R}^d \times \mathbf{R}^d$, i.e.,

$$f^{oo}(x;u,v) \stackrel{\text{def}}{=} \limsup_{\substack{y \to x \\ s,t \to 0}} \frac{f(y+su+tv) - f(y+su) - f(y+tv) + f(y)}{st}, \tag{4}$$

which is positive homogeneous and subadditive in both u and v separately and is symmetric. One can then define a generalized Hessian as the set with support functional f^{oo}; in the Appendix we do this and we relate the result to other definitions appearing in the literature.

For stochastic control, however, where second-order H–J equations arise, it is more convenient to define a second-order derivative and a second-order differential somewhat differently. Let $\mathcal{S}^d$ be the symmetric $d \times d$ dimensional matrices and let $\mathcal{P}^d$ be the cone of nonnegative semi-definite elements of $\mathcal{S}^d$. Let $\mathcal{O} \subset \mathbf{R}^d$ be an open set and let $BM(\mathcal{O})$ be the set of locally bounded measurable real-valued functions defined on $\mathcal{O}$. Here we are considering actual functions, not equivalence classes as is the case with $L^\infty(\mathcal{O})$. For a function $f \in BM(\mathcal{O})$ we define a differential of f at x as a functional on $\mathcal{P}^d \times \mathbf{R}^d$ as follows.

DEFINITION 2.1. *For $a \in \mathcal{P}^d$ and $b \in \mathbf{R}^d$, the* upper generalized Gaussian derivative *of f at x in the direction b with covariance a is*

$$f^G(x;b,a) = \limsup_{\substack{y \to x \\ h \to 0+}} \frac{\mathbb{E}\{(\phi f)(y + bh + \theta w_h) - f(y)\}}{h},$$

where ϕ is any infinitely differentiable function of compact support that is equal to 1 in a neighbourhood of x, where θ is any $d \times m$ dimensional matrix such that $\theta\theta^T = 2a$ and where w is a standard m dimensional Brownian motion. $\mathbb{E}$ stands for expectation.

In [4] we defined a function called $f^G(x;a)$; it is in fact $2f^G(x;0,a)$. We use many of the results proved for f^G; the proofs for the case $b \neq 0$ are straightforward. It can be shown that $f^G(x;b,a)$ is independent of the choice of ϕ, θ, w, [4, Prop. 3.4]. Moreover as in [4, Prop. 4.2], $f^G(x;b,a)$ is positive homogeneous and subadditive on $\mathbf{R}^d \times \mathcal{P}^d$. Observing that $\mathcal{S}^d = \mathcal{P}^d - \mathcal{P}^d$, we can extend $f^G(x;b,a)$ to $\mathbf{R}^d \times \mathcal{S}^d$ as follows:

$$\bar{f}^G(x;b,a) \stackrel{\text{def}}{=} \inf\left\{ \sum_i f^G(x;b_i^+,a_i^+) + \sum_j (-f)^G(x;b_j^-,a_j^-) : a = \sum_i a_i^+ - \sum_j a_j^-, b = \sum_i b_i^+ - \sum_j b_j^-, a_i^+ \in \mathcal{P}^d, a_j^- \in \mathcal{P}^d \right\}. \tag{5}$$

We will show shortly that $\bar{f}^G(x;b,a) = f^G(x;b,a)$ for $a \in \mathcal{P}^d$. To this end, for $a \in \mathcal{P}^d$ and $b \in \mathbf{R}^d$, we define

$$\begin{aligned} f_G(x;b,a) &\stackrel{\text{def}}{=} \liminf_{\substack{y \to x \\ h \to 0+}} \frac{E\{(\phi f)(y + bh + \theta w_h) - f(y)\}}{h} \\ &= -(-f)^G(x;b,a), \end{aligned} \tag{6}$$

where ϕ, θ, w are as in Definition 2.1.

LEMMA 2.1. *For $a \in \mathcal{S}^d, b \in \mathbf{R}^d$,*

$$\bar{f}^G(x;b,a) = \inf\{f^G(x;b^+,a^+) - f_G(x;b^-,a^-) : a = a^+ - a^-, b = b^+ - b^-, a^{\pm} \in \mathcal{P}^d\}. \tag{7}$$

Proof. With $(a,b) = \sum_i (a_i^+, b_i^+) - \sum_j (a_j^-, b_j^-)$, $a^+ = \sum_i a_i^+$, $a^- = \sum_j a_j^-$, $b^+ = \sum_i b_i^+$, $b^- = \sum_j b_j^-$, the subadditivity of f^G implies

$$\begin{aligned} \sum_i f^G(x;b_i^+,a_i^+) + \sum_j (-f)^G(x;b_j^-,a_j^-) &\geq f^G(x;b^+,a^+) + (-f)^G(x;b^-,a^-) \\ &= f^G(x;b^+,a^+) - f_G(x;b^-,a^-) \\ &\geq \text{right-hand side of (7)}. \end{aligned}$$

It follows that the left side of (7) is at least as large as the right side.

The reverse inequality follows readily if we consider the sums in (5) to contain only one term and use (6). □

COROLLARY 2.1. *$\bar{f}^G(x;b,a) = f^G(x;b,a)$ for $(b,a) \in \mathbf{R}^d \times \mathcal{P}^d$.*

Proof. Corollary 4.3 of [4] implies that for $a, a^- \in \mathcal{P}^d$,

$$\begin{aligned} f^G(x;b+b^-,a+a^-) - f_G(x;b^-,a^-) &\geq [f^G(x;b,a) + f_G(x;b^-,a^-)] - f_G(x;b^-,a^-) \\ &= f^G(x;b,a). \end{aligned}$$

Hence with $a = a^+ - a^- \in \mathcal{P}^d$ and $b = b^+ - b^-$, it follows from the Lemma that $\bar{f}^G(x;b,a) \geq f^G(x;b,a)$. As the reverse inequality is obvious, we are done. □

From now on we shall suppress the overbar and take $f^G(x;\cdot,\cdot)$ to be defined on $\mathbf{R}^d \times \mathcal{S}^d$. We can similarly extend $f_G(x;\cdot,\cdot)$ to $\mathbf{R}^d \times \mathcal{S}^d$ by replacing inf by sup and $(\pm f)^G$ by $(\pm f)_G$ in (5). Now we obtain

$$f^G(x;-b,-a) = (-f)^G(x;b,a) = -f_G(x;b,a). \tag{8}$$

We write $\langle a, b\rangle$ for $a \cdot b$ if $a, b \in \mathbf{R}^d$ and for trace(ab) if $a, b \in \mathcal{S}^d$.

DEFINITION 2.2. *The* generalized second-order derivative *of f at x is*

$$\partial^2 f(x) = \{(\beta, \alpha) \in \mathbf{R}^d \times \mathcal{S}^d : \langle \alpha, a\rangle + \langle \beta, b\rangle \le f^G(x; b, a)\, \forall\, (b, a) \in \mathbf{R}^d \times \mathcal{S}^d\}.$$

LEMMA 2.2.

(9)
$$\begin{aligned}\partial^2 f(x) &= \{(\beta, \alpha) \in \mathbf{R}^d \times \mathcal{S}^d : \langle \alpha, a\rangle + \langle \beta, b\rangle \ge f_G(x; b, a)\, \forall\, (b, a) \in \mathbf{R}^d \times \mathcal{S}^d\} \\ &= \{(\beta, \alpha) \in \mathbf{R}^d \times \mathcal{S}^d : f_G(x; b, a) \le \langle \alpha, a\rangle + \langle \beta, b\rangle \le f^G(x; b, a)\, \forall\, (b, a) \in \mathbf{R}^d \times \mathcal{P}^d\}.\end{aligned}$$

Proof. The first equality is a simple application of (8). From this and the definition of $\partial^2 f(x)$ it follows that $\partial^2 f(x)$ is contained in the right side of (9).

Conversely, for (β, α) in the right side of (9), if we decompose $(b, a) \in \mathbf{R}^d \times \mathcal{S}^d$ into

$$(b, a) = (b^+, a^+) - (b^-, a^-), \qquad a^{\pm} \in \mathcal{P}^d,$$

then

$$\begin{aligned}&\langle \alpha, a\rangle + \langle \beta, b\rangle \\ &\quad = \langle \alpha, a^+ - a^-\rangle + \langle \beta, b^+ - b^-\rangle \\ &\quad = [\langle \alpha, a^+\rangle + \langle \beta, b^+\rangle] - [\langle \alpha, a^-\rangle + \langle \beta, b^-\rangle] \\ &\quad \le f^G(x; b^+, a^+) - f_G(x; b^-, a^-),\end{aligned}$$

so the result follows by Lemma 2.1. □

The following decomposition of $\partial^2 f$ is useful. Let us write $\partial^2_H f(x)$ for the Hessian defined by the author in [4], i.e., if we define $f_G(x; a) = 2f_G(x; 0, a)$, then

$$\partial^2_H f(x) = \{\alpha \in \mathcal{S}^d : f_G(x; a) \le \langle \alpha, a\rangle \le f^G(x; a)\, \forall\, a \in \mathcal{P}^d\}.$$

LEMMA 2.3. $\partial^2 f(x) \subset \partial f(x) \times \partial^2_H f(x)$ *with equality if f is strictly differentiable at x, i.e., if $\partial f(x)$ is a singleton.*

Proof. If $(\beta, \alpha) \in \partial^2 f(x)$, then for all $b \in \mathbf{R}^d$ (setting $a = 0$ in the definition)

$$\langle b, \beta\rangle \le f^G(x; b, 0) = f^o(x; b),$$

so that $\beta \in \partial f(x)$. Similarly $\alpha \in \partial^2_H f(x)$. Hence $\partial^2 f(x) \subset \partial f(x) \times \partial^2_H f(x)$.

On the other hand if $\partial f(x) = \{\beta\}$, then $f^G(x; b, a) = f^G(x; 0, a) + \langle b, \beta\rangle$. The result follows. □

Remark. The situation is a little simpler if we assume more regularity. For $\mathcal{O}$ an open set we say that $f \in C^{1,1}(\mathcal{O})$ if f is continuously differentiable on $\mathcal{O}$ with Lipschitz continuous derivatives. This is equivalent to saying that f is in the Sobolev space $W^{2,\infty}(\mathcal{O})$ and implies that the Hessian $f_{x,x}$ exists almost everywhere and is symmetric (since generalized derivatives are symmetric almost everywhere). In this case Lemma 2.3 tells us that $\partial^2 f(x) = \{\nabla f(x)\} \times \partial^2_H f(x)$. Define the Clarke Hessian as $\partial^2_C f(x) \overset{\text{def}}{=} \text{co}\{\lim_i f_{x,x}(x_i) : x_i \to x, x_i \in \text{dom } f_{x,x}\}$. Here co stands for the convex hull. In [4] we showed that for $f \in C^{1,1}(\mathcal{O})$, $\partial^2_H f(x)$ is the conic hull of $\partial^2_C f(x)$, i.e., for $\alpha \in \partial^2_H f(x)$ there exist $\alpha_+, \alpha_- \in \partial^2_C f(x)$ and $p_+, p_- \in \mathcal{P}^d$ such that $\alpha = \alpha_+ + p_+ = \alpha_- - p_-$.

Let us point out that the enlargement to the conic hull is of no importance in the theory of the H–J equation since we will always be concerned with inequalities of the form

$$c + \langle a, \alpha\rangle \le (\ge)\, 0,$$

where c is a scalar and $a \in \mathcal{P}^d$ (this is the ellipticity condition). If such an inequality holds for all $\alpha \in \mathcal{Q}$, then it also holds for all $\alpha \in \mathrm{cc}\mathcal{Q}$, the conic hull of $\mathcal{Q}$. Indeed if, for example, the inequality is $\leq$ and $\alpha \in \mathrm{cc}\mathcal{Q}$, then

$$c + \langle a, \alpha \rangle = c + \langle a, \alpha_- \rangle - \langle a, p_- \rangle \leq -\langle a, p_- \rangle \leq 0.$$

Conversely if the inequality holds for $\alpha \in \mathrm{cc}\mathcal{Q}$, it also holds for $\alpha \in \mathcal{Q}$ since $\mathcal{Q} \subset \mathrm{cc}\mathcal{Q}$.

Hence for $f \in C^{1,1}(\mathcal{O})$, we can take $\partial^2 f(x) = \{\nabla f(x)\} \times \partial^2_C f(x)$.

3. Itô's lemma. For the moment let us consider an uncontrolled Itô process X_t satisfying

$$dX_t = b_t\, dt + \sigma_t\, dw_t,$$

and a function $f : [0,T] \times \mathbf{R}^d \to \mathbf{R}$. We assume that a and b are progressively measurable and $\mathbb{E}\int_0^T |b_t|\, dt < \infty$ and $\mathbb{E}\int_0^T |\sigma_t|^2\, dt < \infty$. It is a well-known result of Itô's that if $f \in C^{1,2}((0,T) \times \mathbf{R}^d)$, then

$$df(t, X_t) = \mathcal{L}f(t, X_t)\, dt + f_x(t, X_t)^T \sigma_t\, dw_t, \quad \text{a.s.}, \tag{10}$$

with

$$\mathcal{L}f(t,x) = f_t(t,x) + \langle a_t, f_{x,x}(t,x) \rangle + b_t \cdot f_x(t,x).$$

Here $a_t = \frac{1}{2}\sigma_t \sigma_t^T$. If f is less regular, a similar result was established by Krylov [7]: Assume that X is as above with additionally $a_t - \lambda I \in \mathcal{P}^d$ for all t, almost all ω, and for some $\lambda > 0$. Suppose that $f \in W^{1,2,d+1}((0,T) \times \mathbf{R}^d)$, the Sobolev space of functions whose derivatives in the sense of distributions of order up to 1 in t and up to 2 in x are in $L^{d+1}((0,T) \times \mathbf{R}^d)$, and suppose also that f is continuous on $[0,T] \times \mathbf{R}^d$ and that the first-order distributional derivatives of f in x are in $L^{2(d+1)}((0,T) \times \mathbf{R}^d)$. Then (10) still holds. If a is degenerate, i.e., $\mathcal{L}$ is not uniformly elliptic, then (10) must, in general, be replaced by an inequality that holds in the mean, [7]. If however, $d = 1$, then the result is correct even in the degenerate case, [6, p. 219]. We shall show that (10) holds in arbitrary dimension for $f \in W^{1,2,\infty}((0,T) \times \mathbf{R}^d)$ if the equation is interpreted appropriately using the generalized Hessian.

We must first define the corresponding generalized second-order derivative. In principle we would use $\partial^2 f(y)$ with $y = (t,x)$, but we observe that in $\mathcal{L}f$ no second-order derivatives in t appear so that we can restrict our attention to a subspace of $\mathcal{S}^{d+1}$, i.e., matrices $a' \in \mathcal{S}^{d+1}$ such that

$$a' = \begin{pmatrix} 0 & 0 \\ 0 & a \end{pmatrix}.$$

If $f(t,x)$ is now a function on $\mathbf{R} \times \mathbf{R}^d = \mathbf{R}^{d+1}$, then for $a \in \mathcal{P}^d$, $b' = (b_o, b)^T \in \mathbf{R}^{d+1}$ we set

$$f^G(t,x;b',a) = \limsup_{\substack{(s,y)\to(t,x) \\ h \to 0+}} \frac{\mathbb{E}\{\phi f(s + b_o h, y + bh + \theta w_h) - f(s,y)\}}{h}$$

and $\partial^2 f(t,x)$ analogously. If $b_o \neq 0$, we can always normalize so that $b_o = 1$ without changing $\partial^2 f(t,x)$. Indeed let us write β' for (β_o, β) and take λ to be a scalar. If $\lambda > 0$, then

$$\langle \alpha, a \rangle + \langle \beta', \lambda b' \rangle \leq f^G(t,x;\lambda b', a)$$

if and only if

$$\langle \alpha, \lambda^{-1}a\rangle + \langle \beta', b'\rangle \le f^G(t,x;b',\lambda^{-1}a).$$

Now if $\lambda = -\mu < 0$, then the "if and only if" becomes (cf. (8))

$$\begin{aligned}\langle \alpha, -\mu^{-1}a\rangle + \langle \beta', b'\rangle &\ge -f^G(t,x;-b',\mu^{-1}a)\\ &= -(-f)^G(t,x;b',-\mu^{-1}a)\\ &= f_G(t,x;b',-\mu^{-1}a).\end{aligned}$$

According to (9), we will not change $\partial^2 f(t,x)$ if we take $\lambda = (b_o)^{-1}$. Hence we can always take $b' = (1,b)^T$. Let $\mathcal{O} \in \mathbf{R}\times\mathbf{R}^d$ be open. Let $f \in BM(\mathcal{O})$.

DEFINITION 3.1. *For* $(b,a) \in \mathbf{R}^d \times \mathcal{P}^d$,

$$\begin{aligned} f^G(t,x;b,a) &= \limsup_{\substack{(s,y)\to(t,x)\\ h\to 0+}} \frac{\mathbb{E}\{\phi f(s+h, y+bh+\theta w_h) - f(s,y)\}}{h},\\ f_G(t,x;a,b) &= \liminf_{\substack{(s,y)\to(t,x)\\ h\to 0+}} \frac{\mathbb{E}\{\phi f(s+h, y+bh+\theta w_h) - f(s,y)\}}{h}. \end{aligned} \tag{11}$$

We can extend these definitions to $a \in \mathcal{S}^d$ as usual. To emphasize the fact that we are only working with a first derivative in t, we now change notation from $\partial^2 f$.

DEFINITION 3.2.

$$\begin{aligned}\partial^{1,2} f(t,x) &= \{(\beta_o,\beta,\alpha) \in \mathbf{R}\times\mathbf{R}^d\times\mathcal{S}^d :\\ &\qquad \langle \alpha, a\rangle + \langle \beta, b\rangle + \beta_o \le f^G(t,x;b,a)\,\forall\,(b,a) \in \mathbf{R}^d\times\mathcal{S}^d\}\\ &= \{(\beta_o,\beta,\alpha) \in \mathbf{R}\times\mathbf{R}^d\times\mathcal{S}^d :\\ &\qquad f_G(t,x;a,b) \le \langle \alpha, a\rangle + \langle \beta, b\rangle + \beta_o \le f^G(t,x;b,a)\,\forall\,(b,a) \in \mathbf{R}^d\times\mathcal{P}^d\}.\end{aligned}$$

THEOREM 3.1. *Assume that* $f \in W^{1,2,\infty}((0,T)\times\mathbf{R}^d)$ *and that* σ, b *are bounded, predictable. Then there exists a predictable process* $(\beta_o(t,\omega),\beta(t,\omega),\alpha(t,\omega))$ *such that*

$$(\beta_o(t,\omega),\beta(t,\omega),\alpha(t,\omega)) \in \partial^{1,2} f(t, X_t(\omega))$$

and

$$\begin{aligned} f(t,X_t) = f(0,X_0) &+ \int_0^t [\beta_o(s,\omega) + \langle b_s, \beta(s,\omega)\rangle + \langle a_s, \alpha(s,\omega)\rangle]\,ds\\ &+ \int_0^t \beta(s,\omega)^T \sigma_s\,dw_s \qquad \text{a.s.}\end{aligned} \tag{12}$$

Moreover, there exists a set A *of full Lebesgue measure such that the usual derivatives* $(f_t(t,x), f_x(t,x), f_{x,x}(t,x))$ *exist for* $(t,x) \in A$ *and*

$$(\beta_o(t,\omega),\beta(t,\omega),\alpha(t,\omega)) = (f_t(t,X_t(\omega)), f_x(t,X_t(\omega)), f_{x,x}(t,X_t(\omega)))$$

whenever $(t,X_t(\omega)) \in A$.

Proof. The last part of the theorem is obvious since functions in $W^{1,2,\infty}$ are Lipschitz in t, and have Lipschitz continuous (in x) first-order partial derivatives with respect to x. So, in fact, $\beta(t,\omega) = f_x(t,X_t(\omega))$. Let X^ϵ satisfy

$$dX_t^\epsilon = b_t\,dt + \sigma_t\,dw_t + \sqrt{\epsilon}\,dw_t'$$

for an independent Brownian motion w'_t. Let

$$g_t^\epsilon = f(t, X_t^\epsilon) - f(0, X_0^\epsilon) - \int_0^t f_x(t, X_r^\epsilon)^T \sigma_r \, dw_r - \sqrt{\epsilon} \int_0^t f_x(t, X_r^\epsilon)^T \, dw'_r.$$

It is well known that $\lim_{\epsilon \to 0} \sup_{0 \le t \le T} |X_t^\epsilon - X_t| = 0$, and so

$$\begin{aligned} \lim_{\epsilon \to 0} g_t^\epsilon &= g_t \\ &\stackrel{\text{def}}{=} f(t, X_t) - f(0, X_0) - \int_0^t f_x(t, X_r)^T \sigma_r \, dw_r. \end{aligned}$$

Moreover, Krylov's result implies that

$$g_t^\epsilon - g_s^\epsilon = \int_s^t [\langle f_{x,x}(r, X_r^\epsilon), a_r \rangle + \langle f_x(r, X_r^\epsilon), b_r \rangle + f_t(r, X_r^\epsilon) + \tfrac{1}{2}\epsilon \langle f_{x,x}(r, X_r^\epsilon), I \rangle] \, dr \quad \text{a.s.},$$

and hence

$$|g_t - g_s| \le K \|f\|_{1,2,\infty} |t - s| \quad \text{a.s.}$$

for some constant K. Thus $dg_t(\omega)/dt \stackrel{\text{def}}{=} g'(t, \omega)$ exists almost everywhere, almost surely. If we set $g'_n(t, \omega) = n[g_t(\omega) - g_{t-1/n}(\omega)]$, then g'_n is adapted and continuous, hence predictable, and $g'_n \to g'$ almost everywhere, almost surely, so g' is predictable (we assume that the underlying filtration is complete). Define

$$F(t, \omega) = \{(\beta_o, \beta, \alpha) \in \partial^{1,2} f(t, X_t(\omega)) : g'(t, \omega) = \beta_o + \langle \beta, b_t(\omega) \rangle + \langle \alpha, a_t(\omega) \rangle\}.$$

We will show that there is a measurable (i.e., predictable) selection of F on $[0, T] \times \Omega$; this will establish the theorem.

We begin by showing that $F(t, \omega) \neq \emptyset$. Let

$$I(t, \omega) \stackrel{\text{def}}{=} \{\beta_o + \langle \beta, b_t(\omega) \rangle + \langle \alpha, a_t(\omega) \rangle : (\beta_o, \beta, \alpha) \in \partial^{1,2} f(t, X_t(\omega))\}.$$

Since $\partial^{1,2} f(t, X_t(\omega))$ is convex and compact, then so is $I(t, \omega)$, i.e., it is an interval $[A(t, \omega), B(t, \omega)]$. We define $A^\epsilon(t, \omega), B^\epsilon(t, \omega)$ similarly using X_t^ϵ. Since $(t, x) \to \partial^{1,2} f(t, x)$ is upper semi-continuous, then

$$\liminf_{\substack{t \to s \\ \epsilon \to 0+}} A^\epsilon(t, \omega) \ge A(s, \omega), \qquad \limsup_{\substack{t \to s \\ \epsilon \to 0+}} B^\epsilon(t, \omega) \le B(s, \omega).$$

Since

$$g_t^\epsilon(\omega) - g_s^\epsilon(\omega) \le \int_s^t B^\epsilon(r, \omega) \, dr + \tfrac{1}{2}\epsilon \|f\|_{1,2,\infty} |t - s|,$$

then for all $\delta > 0$, there exists ϵ_δ such that if $|t - s| < \epsilon_\delta$ and $\epsilon < \epsilon_\delta$, then

$$g_t^\epsilon(\omega) - g_s^\epsilon(\omega) \le [B(s, \omega) + \delta + k\epsilon] |t - s|$$

for some constant k. Now let $\epsilon \to 0$ to conclude that $g'(t, \omega) \le B(t, \omega) + \delta$. Let $\delta \to 0$ and do the corresponding estimate for $A(t, \omega)$ to conclude that $g'(t, \omega) \in I(t, \omega)$ and hence $F(t, \omega) \neq \emptyset$.

Next we observe that $F(t,\omega)$ is compact since $\partial^{1,2}f(t,X(\omega))$ is, since

$$\pi(t,\omega) \stackrel{\text{def}}{=} \{(\beta_o,\beta,\alpha) : g'(t,\omega) = \beta_o + \langle\beta, b_t(\omega)\rangle + \langle\alpha, a_t(\omega)\rangle\}$$

is closed and $F(t,\omega) = \pi(t,\omega) \cap \partial^{1,2}f(t,X_t(\omega))$.

It remains to show that the multifunction F is measurable. Observe that π is a plane so we have a formula for the distance ρ of an arbitrary but fixed point to the plane. As the mapping $(t,\omega) \to \rho$ is predictable, then π is measurable (see [8, Thm. 4.2]). Since $F(t,\omega) = \pi(t,\omega) \cap \partial^{1,2}f(t,X_t(\omega))$, then F is measurable provided $\partial^{1,2}f$ is (see [8, p. 863]). It is easy to see that upper semi-continuity of $\partial^{1,2}f$ as defined in [1] and established in [4] implies upper semi-continuity in the sense of [8], i.e., for any closed set K, $\{(t,x) : \partial^{1,2}f(t,x) \cap K \neq \emptyset\}$ is closed, which obviously implies measurability.

The existence of a measurable selection of F now follows from a standard theorem, [8, Thm. 4.1]. □

4. Generalized solutions. We consider the control problem

$$X_s^{t,x,u} = x + \int_0^s b(t+r, X_t^{t,x,u}, u_r)\,dr + \int_0^s \sigma(t+r, X_r^{t,x,u}, u_r)\,dw_r \tag{13}$$

with cost

$$J^{t,x}(u) = \mathbb{E}\left\{\int_0^{T-t} f(t+r, X_r^{t,x,u}, u_r)\,dr + f_o(X_{T-t}^{t,x,u})\right\}. \tag{14}$$

The corresponding value function is

$$v(t,x) = \inf_{u\in\mathcal{U}} J^{t,x}(u), \tag{15}$$

where $\mathcal{U}$ is the set of U-valued, progressively measurable stochastic process on $[0,T)$. It follows from a generalized dynamic programming argument, [5, Props. 5.9 and 5.11], that

$$\Gamma_s^{t,x,u} \stackrel{\text{def}}{=} \int_0^s f(t+r, X_r^{t,x,u}, u_r)\,dr + v(t+s, X_s^{t,x,u}) \tag{16}$$

is a submartingale for any $u \in \mathcal{U}$ and is a martingale if and only if u is optimal. If v is smooth, this implies that it is the unique classical solution of the H–J equation, i.e.,

$$-v_t(t,x) + H(t,x,-v_x(t,x),-v_{x,x}(t,x)) = 0,$$

where the subscripts denote partial derivatives and where H is the Hamiltonian for the problem and is defined on $[0,T] \times \mathbf{R}^d \times \mathbf{R}^d \times \mathcal{S}^d$ as

$$H(t,x,p,P) \stackrel{\text{def}}{=} \sup_{u\in U}\{\langle P, a(t,x,u)\rangle + \langle p, b(t,x,u)\rangle - f(t,x,u)\}.$$

Note that $a = \frac{1}{2}\sigma\sigma^T$. This situation occurs rarely, but it is true that v is always a weak (in the sense of distributions) solution, although the latter are not unique. We shall show that v satisfies the H–J equation in a generalized sense, at least if v possesses some regularity.

DEFINITION 4.1. *A function $v \in BM(\mathcal{O})$ is a* generalized solution *of* (H–J) *in $\mathcal{O}$ if, for each $(t,x) \in \mathcal{O}$,*

$$\inf_{(\beta_o,\beta,\alpha)\in\partial^{1,2}v(t,x)}\{\beta_o - H(t,x,-\beta,-\alpha)\} = 0. \tag{17}$$

Let us suppose that $v \in W^{1,2,\infty}_{\text{loc}}(\mathcal{O})$ so that v_t and $v_{x,x}$ exist almost everywhere. It follows from [4, Thm. 5.6] that

$$\begin{aligned}\partial^{1,2}v(t,x) = \text{ cc co}\{\lim_i (v_t(t_i,x_i), v_x(t,x), v_{x,x}(t_i,x_i)) : \\ (t_i,x_i) \to (t,x), (t_i,x_i) \in \text{dom}(v_t, v_{x,x})\},\end{aligned} \tag{18}$$

where the operation of taking the conic hull applies to the component in $\mathcal{S}^d$, i.e., for $M \subset \mathbf{R}^{d+1} \times \mathcal{S}^d$, $\text{cc}M = (M + \mathcal{P}^d_o) \cap (M - \mathcal{P}^d_o)$ where $\mathcal{P}^d_o = \{0\} \times \mathcal{P}^d$. Hence $\partial^{1,2}v(t,x) \neq \emptyset$. We now have the main result of this section.

THEOREM 4.1. *Assume that $v \in W^{1,2,\infty}_{\text{loc}}((0,T) \times \mathbf{R}^d)$ is the value function. Then v is a generalized solution of* (H–J) *on* $(0,T) \times \mathbf{R}^d$.

Proof. We shall first show that the left side of (17) is at least 0. Let us take u constant in (13); then

$$\begin{aligned}X^{s,y,u}_h &= y + \int_0^h b(s+r, X^{s,y,u}_r, u)\,dr + \int_0^h \sigma(s+r, X^{s,y,u}_r, u)\,dw_r \\ &= y + hb(t,x,u) + \sigma(t,x,u)w_h + \rho_1 + \rho_2,\end{aligned}$$

where

$$\rho_1 = \int_0^h [b(s+r, X^{s,y,u}_r, u) - b(t,x,u)]\,dr$$

and

$$\rho_2 = \int_0^h [\sigma(s+r, X^{s,y,u}_r, u) - \sigma(t,x,u)]\,dw_r.$$

Observe that the Lipschitz continuity in x of b and standard estimates show that

$$h^{-1}\rho_1 = o(1), \tag{19}$$

where $o(1)$ is a random function $g(s,y)$ such that

$$\lim_{\substack{(s,y)\to(t,x)\\ h\to 0+}} \mathbb{E}|g(s,y)|^2 = 0.$$

Similarly

$$h^{-1/2}\rho_2 = o(1). \tag{20}$$

Now

$$\begin{aligned}v(s+h, X^{s,y,u}_h) - v(s,y) &= [v(s+h, y + hb(t,x,u) + \sigma(t,x,u)w_h + \rho_1 + \rho_2) \\ &\quad - v(s+h, y + hb(t,x,u) + \sigma(t,x,u)w_h)] \\ &\quad + [v(s+h, y + hb(t,x,u) + \sigma(t,x,u)w_h) - v(s,y)] \\ &= \Delta v + [v(s+h, y + hb(t,x,u) + \sigma(t,x,u)w_h) - v(s,y)].\end{aligned} \tag{21}$$

But v_x is locally Lipschitz continuous, so

$$\begin{aligned}\Delta v &= [v_x(s+h, y + hb(t,x,u) + \sigma(t,x,u)w_h + \xi(\rho_1+\rho_2))](\rho_1+\rho_2) \\ &= v_x(s+h,y)(\rho_1+\rho_2) + [v_x(s+h, y + hb(t,x,u) + \sigma(t,x,u)w_h \\ &\qquad + \xi(\rho_1+\rho_2)) - v_x(s+h,y)](\rho_1+\rho_2),\end{aligned}$$

where ξ is a random variable assuming values in $(0, 1)$. Set

$$v_x(s+h, y+hb(t,x,u)+\sigma(t,x,u)w_h+\xi(\rho_1+\rho_2)) - v_x(s+h,y) = \Delta_1 v.$$

Then

$$\begin{aligned}\Delta_1 v &\le K|hb(t,x,u)+\sigma(t,x,u)w_h+\xi(\rho_1+\rho_2)| \\ &\le K_1\{|h|+|w_h|+|\rho_1|+|\rho_2|\}\end{aligned}$$

and hence

$$\lim_{\substack{(s,y)\to(t,x)\\ h\to 0+}} h^{-1}\mathbb{E}\Delta v = 0. \tag{22}$$

If now we set $b = b(t,x,u)$ and $a = \frac{1}{2}\sigma(t,x,u)\sigma(t,x,u)^T$, then it follows from (16), (11), (21), and (22) that

$$\begin{aligned}v_G(t,x;b,a) &= \liminf_{\substack{(s,y)\to(t,x)\\ h\to 0+}} h^{-1}\mathbb{E}\{\phi v(s+h, X_h^{s,y,u}) - v(s,y)\} \\ &= \liminf_{\substack{(s,y)\to(t,x)\\ h\to 0+}} h^{-1}\mathbb{E}\{v(s+h, X_h^{s,y,u}) - v(s,y)\} \\ &\ge \liminf_{\substack{(s,y)\to(t,x)\\ h\to 0+}} h^{-1}\mathbb{E}\int_0^h -f(s+r, X_r^{s,y,u})\,dr \\ &= -f(t,x,u)\end{aligned}$$

(the ϕ can be omitted since v has at most polynomial growth). Hence for $(\beta_o, \beta, \alpha) \in \partial^{1,2}v(t,x)$,

$$f(t,x,u) + \langle \alpha, a(t,x,u)\rangle + \langle \beta, b(t,x,u)\rangle + \beta_o \ge 0$$

so that

$$\beta_o - H(t,x,-\beta,-\alpha) \ge 0.$$

Let us now show that the left side of (4.1) is at most zero. Let (t,x) be a point such that $v_t(t,x), v_{x,x}(t,x)$ exist, i.e., $(t,x) \in \mathrm{dom}(v_t, v_{x,x})$.

Since we are not assuming convexity of U, we consider now the relaxed control problem, i.e., we take the controls to be probability measures on U, and if μ is such a control, we replace $f(t,x,u)$ by

$$f(t,x,u) = \int_U f(t,x,u)\,\mu(du)$$

and similarly for b. On the other hand, $\sigma(t,x,\mu)$ is a square root of

$$2a(t,x,\mu) = 2\int_U a(t,x,u)\,\mu(du).$$

If a is singular for some (t,x,u), then we may have to extend the probability space to carry an independent Brownian motion in addition to X. Neither the value function v nor the Hamiltonian H change under this enlargement. Let μ^* be an optimal Markov control. It exists by the results of [5]. We write $X^{t,x}_\cdot$ for the corresponding state, i.e.,

$$X_s^{t,x} = x + \int_0^s b(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\,dr + \int_0^s \sigma(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\,dw_r.$$

Then, cf. (16), Γ_s^{t,x,μ^*} is a martingale and so Theorem 3.1 implies there is a $\partial^{1,2}v(t+s, X_s^{t,x})$ valued process $(\beta_o(s), \beta(s), \alpha(s))$ such that

$$\int_0^s [f(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x})) + \beta_o(r) + \langle \beta(r), b(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\rangle + \langle \alpha(r), a(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\rangle]\, dr = 0 \qquad \text{a.s.}$$

We wish to differentiate the above with respect to s at $s = 0$ so we consider

(23)

$$h^{-1}\int_0^h [f(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x})) + \beta_o(r) + \langle \beta(r), b(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\rangle + \langle \alpha(r), a(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\rangle]\, dr = 0 \qquad \text{a.s.}$$

A typical term in (23) is

$$\begin{aligned} & h^{-1}\int_0^h \langle \alpha(r), a(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\rangle\, dr \\ & = h^{-1}\int_0^h \langle v_{x,x}(t,x), a(t+r, X_r^{t,x}, \mu^*(t+r, X_r^{t,x}))\rangle\, dr + o(1) \\ & = h^{-1}\int_0^h \langle v_{x,x}(t,x), \int_U a(t,x,u)\mu^*(t+r, X_r^{t,x})(du)\rangle\, dr + o(1) \\ & = \langle v_{x,x}(t,x), a(t,x,\mu_h)\rangle + o(1) \end{aligned} \tag{24}$$

by the upper semi-continuity of $\partial^{1,2}v$ and the Lipschitz continuity of a. Here

$$\mu_h(A) = h^{-1}\int_0^h \mu^*(t+r, X_r^{t,x})(A)\, dr,$$

again a probability measure on U and $o(1) \to 0$ almost surely as $h \to 0$. Since U is compact, then there exists a sequence $h_n \to 0$ such that μ_{h_n} converges weakly to some probability measure μ_o on U. The terms involving f and b are treated similarly. Hence taking the limit $h \to 0$ along a suitable subsequence in (23) leads to

$$v_t(t,x) - \{\langle -v_{x,x}(t,x), a(t,x,\mu_o)\rangle - \langle -v_x(t,x), b(t,x,\mu_o)\rangle - f(t,x,\mu_o)\} = 0, \tag{25}$$

and hence

$$v_t(t,x) - H(t,x,-v_x,-v_{x,x}) \le 0.$$

This holds for $(t,x) \in \operatorname{dom}(v_t, v_{x,x})$. For arbitrary (t,x) let (t_i, x_i) be a sequence in $\operatorname{dom}(v_t, v_{x,x})$ converging to (t,x) such that $(v_t(t_i,x_i), v_x(t_i,x_i), v_{x,x}(t_i,x_i))$ converges; it will converge to an element of $\partial^{1,2}v(t,x)$ according to (18). It follows that

$$\inf_{(\beta_o,\beta,\alpha)\in\partial^{1,2}v(t,x)} \{\beta_o - H(t,x,-\beta,-\alpha)\} \le 0$$

and the proof is complete. □

5. Viscosity solutions. Let us now investigate the relation between generalized solutions and viscosity solutions of H–J. Recall that the viscosity solution is unique and is, in fact, the value function. It follows from Theorem 4.1 that the viscosity solution of H–J is a generalized solution. We would like, however, to investigate this relationship without recourse to the value function. We now call H–J any equation of the form

$$v_t(t,x) - H(t,x,-v_x(t,x),-v_{x,x}(t,x)) = 0,$$

with H continuous, convex in its last two arguments, and elliptic in the sense that $H(t,x,p,P) \leq H(t,x,p,Q)$ if $P \leq Q$. We begin with some definitions related to viscosity solutions. Let f be continuous in a neighbourhood of x.

DEFINITION 5.1. *The* superdifferential *of f at $x \in \boldsymbol{R}^d$ is the set*

$$D^2_+ f(x) = \left\{ (\beta,\alpha) \in \mathbf{R}^d \times \mathcal{S}^d : \limsup_{y \to x} \frac{f(y) - f(x) - \beta \cdot (y-x) - \frac{1}{2}(y-x)^T \alpha (y-x)}{\|y-x\|^2} \leq 0 \right\}.$$

The subdifferential *of f at $x \in \boldsymbol{R}^d$ is the set*

$$D^2_- f(x) = \left\{ (\beta,\alpha) \in \mathbf{R}^d \times \mathcal{S}^d : \liminf_{y \to x} \frac{f(y) - f(x) - \beta \cdot (y-x) - \frac{1}{2}(y-x)^T \alpha (y-x)}{\|y-x\|^2} \geq 0 \right\}.$$

It follows that these sets are closed, convex, and unbounded if not empty. In fact if $(\beta,\alpha) \in D^2_+ f(x)$, then $(\beta,\alpha) + \mathcal{P}^d_o \subset D^2_+ f(x)$. Recall that $\mathcal{P}^d_o = \{0\} \times \mathcal{P}^d$. Moreover

$$D^2_- f(x) = -D^2_+(-f)(x). \tag{26}$$

We point out that it is customary to let $\alpha \in \mathbf{R}^{d \times d}$ in the definition of the sub- and superdifferentials; however, since such α are only considered in inner products with symmetric matrices both in the definition and applications, it is no great loss to symmetrize them. To relate these derivatives to $\partial^2 f(t,x)$, we replace $\mathbf{R}^d$ by $\mathbf{R}^{d+1}$ and $\mathcal{S}^d$ by $\mathcal{S}^{d+1}$ in the above definitions. We write

$$\alpha' = \begin{pmatrix} \alpha_o & \alpha_1 \\ \alpha_1^T & \alpha \end{pmatrix}, \qquad \beta' = \begin{pmatrix} \beta_o \\ \beta \end{pmatrix},$$

with $\alpha \in \mathcal{S}^d, \beta \in \mathbf{R}^d$, and α_o, β_o scalars. Consider parabolic equations as degenerate elliptic equations; as these involve second-order derivatives only with respect to x, let us define $D^{1,2}_+ f$ and $D^{1,2}_- f$ by projecting out the unnecessary components in $D^2_+ f$ and $D^2_- f$, i.e.,

$$D^{1,2}_+ f(t,x) = \{(\beta_o,\beta,\alpha) \in \mathbf{R} \times \mathbf{R}^d \times \mathcal{S}^d : (\beta',\alpha') \in D^2_+ f(t,x)\},$$
$$D^{1,2}_- f(t,x) = \{(\beta_o,\beta,\alpha) \in \mathbf{R} \times \mathbf{R}^d \times \mathcal{S}^d : (\beta',\alpha') \in D^2_- f(t,x)\}.$$

We observe that these differentials are subsets of the corresponding differentials $D^{1,2,\pm}_{t+,x} f(t,x)$ defined by Zhou [9] as

$$D^{1,2,+}_{t+,x} f(t,x) = \Bigg\{ (\beta_o,\beta,\alpha) \in \mathbf{R} \times \mathbf{R}^d \times \mathcal{S}^d : \limsup_{\substack{y \to x \\ h \to 0+}} \frac{f(t+h,y) - f(t,x) - \beta_o h - \beta \cdot (y-x) - \frac{1}{2}(y-x)^T \alpha (y-x)}{h + \|y-x\|^2} \leq 0 \Bigg\},$$

and similarly $D^{1,2,-}_{t+,x} f(t,x)$. Note that if $(\beta_o^+, \beta^+, \alpha^+) \in D^{1,2}_+ f(t,x)$ and $(\beta_o^-, \beta^-, \alpha^-) \in D^{1,2}_- f(t,x)$, then $\beta_o^+ = \beta_o^-, \beta^+ = \beta^-$, and $\alpha^+ \geq \alpha^-$, whereas if $D^{1,2}_\pm f(t,x)$ is replaced by $D^{1,2,\pm}_{t+,x} f(t,x)$, then in the above we replace $\beta_o^+ = \beta_o^-$ by $\beta_o^+ \geq \beta_o^-$. Let $\mathcal{O} \subset \mathbf{R}^{d+1}$ be open.

THEOREM 5.1. *If f is continuous at $(t,x) \in \mathcal{O}$, then*

$$D^{1,2}_+ f(t,x) \subset \partial^{1,2} f(t,x) + \mathcal{P}^d_o,$$
$$D^{1,2}_- f(t,x) \subset \partial^{1,2} f(t,x) - \mathcal{P}^d_o.$$

Proof. We only prove the second inclusion as the first then follows from (26) and the fact that $-\partial^{1,2}(-f) = \partial^{1,2} f$. Consider $(\beta_o, \beta, \alpha) \in D^{1,2}_- f(t,x)$. Given $\epsilon > 0$, there exists $\delta > 0$ such that

$$\begin{aligned} f(s,y) - f(t,x) - \beta_o(s-t) - \beta \cdot (y-x) - \tfrac{1}{2}(y-x)^T \alpha (y-x) \\ \geq -\epsilon\{|s-t|^2 + \|y-x\|^2\} + (y-x)^T \alpha_1 (s-t) + \tfrac{1}{2}\alpha_o (s-t)^2 \end{aligned} \tag{27}$$

for $\|(s,y) - (t,x)\| < \delta$. For $a = \frac{1}{2}\theta\theta^T \in \mathcal{P}^d$, $b \in \mathbf{R}^d$ let $s = t + h$, $y = x + bh + \theta w_h$ for a suitable Brownian motion. If h is sufficiently small, then by Chebychev's inequality $\|(s,y) - (t,x)\| < \delta$ except on a set of probability measure less than Kh^2, where K is a constant that may depend on ϵ and a but not h. Let ϕ be a smooth function of compact support, equal to one on $\|(s,y) - (t,x)\| < \delta$. Since f is bounded near (t,x), then from (27) it follows that

$$\begin{aligned} &\mathbb{E}\phi f(t+h, x+bh+\theta w_h) - f(t,x) \\ &\quad \geq \beta_o h + \langle \beta, b \rangle h + \langle \alpha, a \rangle h - \epsilon\{(1 + \|b\|^2)h^2 + 2\|a\|h\} - O(h^2) \end{aligned}$$

and hence, after dividing by h, letting $h \to 0$ and then $\epsilon \to 0$,

$$f^G(t,x;b,a) \geq \beta_o + \langle \beta, b \rangle + \langle \alpha, a \rangle.$$

Since the support function at $(1,b,a)$ of $\partial^{1,2} f(t,x) - \mathcal{P}^d_o \stackrel{\text{def}}{=} \mathcal{A}$ is

$$\begin{aligned} \sup_{(\beta_o,\beta,\alpha)\in\mathcal{A}} \{\beta_o + \langle b, \beta \rangle + \langle a, \alpha \rangle\} &= f^G(t,x;b,a) - \inf\{\langle a, p \rangle : p \in \mathcal{P}^d\} \\ &= \begin{cases} f^G(t,x;b,a) & \text{if } a \in \mathcal{P}^d, \\ \infty & \text{otherwise,} \end{cases} \end{aligned}$$

we can conclude that $(\beta_o, \beta, \alpha) \in \partial^{1,2} f(t,x) - \mathcal{P}^d_o$. □

DEFINITION 5.2. *A continuous function v is a* viscosity subsolution *of* H–J *if, for every* (t,x),

$$\inf_{(\beta_o,\beta,\alpha)\in D^{1,2}_+ v(t,x)} \{\beta_o - H(t,x,-\beta,-\alpha)\} \geq 0;$$

it is a viscosity supersolution *if*

$$\sup_{(\beta_o,\beta,\alpha)\in D^{1,2}_- v(t,x)} \{\beta_o - H(t,x,-\beta,-\alpha)\} \leq 0;$$

and it is a viscosity solution *if it is both a viscosity sub- and supersolution.*

From Theorem 5.1 and the fact that for $p \in \mathcal{P}^d$,

$$H(t,x,-\beta,-\alpha) \geq H(t,x,-\beta,-\alpha-p),$$

it follows that *a generalized solution is a viscosity subsolution.* Unlike the deterministic case [3], it is not clear what additional hypothesis is required to ensure that it is a supersolution. We can, however, show that a viscosity solution is a generalized solution.

THEOREM 5.2. *If* $v \in W^{1,2,\infty}_{\text{loc}}(\mathcal{O})$ *is a viscosity solution of* H–J *on* $\mathcal{O}$, *then it is a generalized solution on* $\mathcal{O}$.

Proof. We follow Frankowska [3]. Let $(t_i, x_i) \to (t, x)$ with $(t_i, x_i) \in \operatorname{dom}(v_t, v_{x,x})$ and $(v_t(t_i, x_i), v_{x,x}(t_i, x_i)) \to (\beta_o, \alpha)$ for some (β_o, α). As a viscosity solution is an almost everywhere solution, then

$$v_t(t_i, x_i) - H(t_i, x_i, -v_x(t_i, x_i), -v_{x,x}(t_i, x_i)) = 0.$$

The continuity of H implies that

$$\beta_o - H(t, x, -v_x(t, x), -\alpha) = 0.$$

The convexity of H implies that the left-hand side above is concave. Then for $(\beta_o, \beta, \alpha) \in \operatorname{co}\{\lim_{(t_i,x_i)\to(t,x)}(v_t(t_i, x_i), v_x(t_i, x_i), v_{x,x}(t_i, x_i))\}$,

$$\beta_o - H(t, x, -\beta, -\alpha) \geq 0.$$

As pointed out previously, the same inequality now holds on the conic hull of the set and hence on $\partial^2 v(t, x)$ and v is a generalized solution. □

6. Appendix. To derive a generalized Hessian as a closed convex set with support functional f^{oo} (cf. (4) for the definition), it is easiest to use the setting of tensors. Recall that if $\mathcal{B}(\mathbf{R}^d)$ is the space of continuous bilinear functionals on $\mathbf{R}^d$ and if $(\mathbf{R}^d \otimes \mathbf{R}^d)^*$ is the dual of the tensor product $\mathbf{R}^d \otimes \mathbf{R}^d$, then $\mathcal{B}(\mathbf{R}^d)$ is isomorphic to $(\mathbf{R}^d \otimes \mathbf{R}^d)^*$. We shall identify these two isomorphic spaces. Now f^{oo} can be extended to a positive homogeneous subadditive functional defined on $\mathbf{R}^d \otimes \mathbf{R}^d$ by

$$f^{oo}_T(x; a) \stackrel{\text{def}}{=} \inf\left\{\sum_i f^{oo}(x; u_i, v_i) : a = \sum_i u_i \otimes v_i\right\}.$$

We can define the generalized Hessian of f at x as a set in $\mathcal{B}(\mathbf{R}^d)$ as follows:

$$\partial^2_T f(x) \stackrel{\text{def}}{=} \{\alpha \in \mathcal{B}(\mathbf{R}^d) : \langle \alpha, a\rangle \leq f^{oo}_T(x; a), a \in \mathbf{R}^d \otimes \mathbf{R}^d\},$$

where $\langle \cdot\,, \cdot \rangle$ denotes duality between $\mathbf{R}^d \otimes \mathbf{R}^d$ and $(\mathbf{R}^d \otimes \mathbf{R}^d)^*$ or $\mathcal{B}(\mathbf{R}^d)$. Cominetti and Correa [2] define a generalized Hessian, $\partial^2_{CC} f(x)$, somewhat differently using fans and prefans. It can be shown that, in their terminology, $\partial^2_T f(x)$ is the generator of the largest linearly generated prefan contained in $\partial^2_{CC} f(x)$.

Of course we may want the Hessian to contain only symmetric elements. In that case we simply work on a suitable subspace of $\mathbf{R}^d \otimes \mathbf{R}^d$. Let $\mathcal{S}^d$ denote the symmetric elements in $\mathcal{B}(\mathbf{R}^d)$, i.e.,

$$\mathcal{S}^d = \{f \in \mathcal{B}(\mathbf{R}^d) : f(x, y) = f(y, x)\}$$

(we identify it with the symmetric $d \times d$ matrices). The $\mathcal{S}^d$ is isomorphic to a subspace S^* of "symmetric" elements of $(\mathbf{R}^d \otimes \mathbf{R}^d)^*$. For $a = \sum_i x_i \otimes y_i$, we set $a^* = \sum_i y_i \otimes x_i$ and we set

$$\begin{aligned}\mathbf{R}^d \otimes_s \mathbf{R}^d &= \left\{\sum_i x_i \otimes x_i - \sum_j y_j \otimes y_j\right\} \\ &= \{a \in \mathbf{R}^d \otimes \mathbf{R}^d : a^* = a\}.\end{aligned}$$

Then polar decomposition and symmetry imply that the elements of S^* are determined by their action on $\mathbf{R}^d \otimes_s \mathbf{R}^d$ and we can take S^* to be $(\mathbf{R}^d \otimes_s \mathbf{R}^d)^*$. Ignoring the isomorphism from now on, we identify $\mathcal{S}^d$ with $(\mathbf{R}^d \otimes_s \mathbf{R}^d)^*$. Then for $a \in \mathbf{R}^d \otimes_s \mathbf{R}^d$ we define

$$f_S^{oo}(x;a) \stackrel{\text{def}}{=} \inf\left\{\sum_i f^{oo}(x;u_i,u_i) + \sum_j (-f)^{oo}(x;v_j,v_j) : a = \sum_i u_i \otimes u_i - \sum_j v_j \otimes v_j\right\}$$

and

$$\partial_S^2 f(x) \stackrel{\text{def}}{=} \{\alpha \in \mathcal{S}^d : \langle \alpha, a\rangle \le f_S^{oo}(x;a), a \in \mathbf{R}^d \otimes_s \mathbf{R}^d\}.$$

Observe that the definition of $f_S^{oo}(x;a)$ looks much like that of $\bar{f}^G(s;0,a)$, but we point out that in (5) $a_\cdot^\pm$ need not be rank one matrices!

LEMMA 6.1. (i) $\partial_S^2 f(x) \supset \partial_T^2 f(x) \cap \mathcal{S}^d$.

(ii) *If* $\partial_T^2(f)(x) \neq \emptyset$, *then* $\partial_S^2 f(x) \neq \emptyset$.

Proof. Since $(-f)^{oo}(x;u,v) = f^{oo}(x;-u,v)$, then $f_T^{oo}(x;a) \le f_S^{oo}(x;a)$ on $\mathbf{R}^d \otimes_s \mathbf{R}^d$. Now (i) follows. For $\alpha \in \mathcal{B}(\mathbf{R}^d) = (\mathbf{R}^d \otimes \mathbf{R}^d)^*$ we define α^* by $\alpha^*(u,v) \stackrel{\text{def}}{=} \alpha(v,u)$. Then $\langle \alpha^*, a\rangle = \langle \alpha, a^*\rangle$. The symmetry of f^{oo} implies that $f_T^{oo}(x;a^*) = f_T^{oo}(x;a)$ so that if $\alpha \in \partial_T^2 f(x)$, then $\alpha^* \in \partial_T^2 f(x)$ and hence $\frac{1}{2}(\alpha + \alpha^*) \in \partial_T^2 f(x) \cap \mathcal{S}^d$. The second result now follows by (i). □

We conjecture that equality holds in (i) above under reasonable hypotheses. The above definitions can of course all be given with $\mathbf{R}^d$ replaced by a topological vector space. At this point we can compare our Hessian, $\partial_H^2 f$, defined using a stochastic setting, and various other Hessians.

LEMMA 6.2. $f^G(x;0,a) \le f_S^{oo}(x;a)$ *and* $\partial_H^2 f(x) \subset \partial_S^2 f(x)$.

Proof. The second result follows trivially from the first so let us prove this result. Given $\theta \in \mathbf{R}^d$ and $\epsilon > 0$, there exists $\delta > 0$ such that if $|y-x| < \delta, |t| < \delta$, then

$$f(y+2\theta t) - 2f(y+\theta t) + f(y) \le t^2[f^{oo}(x;\theta,\theta) + \epsilon]. \tag{28}$$

For a scalar Brownian motion w and $h > 0$, let $t = w_h/2$ and choose ϕ smooth with compact support such that $\phi = 1$ on $\{y : |y-x| < \delta(|\theta|+1)\}$. Then for $|y-x| < \delta$, inequality (28) implies

$$\begin{aligned}\mathbf{1}_{\{|w_h|<\delta\}}[(\phi f)(y+\theta w_h) - f(y)] \le\ & 2\mathbf{1}_{\{|w_h|<\delta\}}[(\phi f)(y+\tfrac{1}{2}\theta w_h) - f(y)] \\ & + \tfrac{1}{4}\mathbf{1}_{\{|w_h|<\delta\}} w_h^2 (f^{oo}(x;\theta,\theta) + \epsilon).\end{aligned} \tag{29}$$

Since ϕf is bounded then, with $\rho = 1$ or $\frac{1}{2}$, $\mathbb{E}\mathbf{1}_{\{|w_h|\ge\delta\}}[(\phi f)(y+\rho\theta w_h) - f(y)] = O(h^2)$ by Chebychev's inequality, and $\mathbb{E}\mathbf{1}_{\{|w_h|\ge\delta\}} w_h^2 = O(h^{3/2})$. Now the inequality (29) implies

$$\begin{aligned}\mathbb{E}[(\phi f)(y+\theta w_h) - f(y)] \le\ & 2\mathbb{E}[(\phi f)(y+\tfrac{1}{2}\theta w_h) - f(y)] \\ & + \tfrac{1}{4}\mathbb{E} w_h^2 (f^{oo}(x;\theta,\theta) + \epsilon) + o(h).\end{aligned}$$

Hence $f^G(x;0,\frac{1}{2}\theta\theta^T) \le \frac{1}{2} f^{oo}(x;\theta,\theta)$ and so $f^G(x;0,a) \le f_S^{oo}(x;a)$ for $a \in \mathbf{R}^d \otimes_s \mathbf{R}^d$. □

Remark. We continue here the remark made at the end of §2. For $f \in C^{1,1}(\mathcal{O})$, $\partial_{CC}^2 f(x)$ is linearly generated by $\partial_C^2 f(x)$ [2]. Since it is also generated by $\partial_T^2 f(x)$, then $\partial_T^2 f(x) = \partial_C^2 f(x)$, and the conic hull of these sets is $\partial_H^2 f(x)$. It follows that in this setting

$$\partial_T^2 f(x) \subset \operatorname{cc}\partial_T^2 f(x) = \partial_H^2 f(x) \subset \partial_S^2 f(x),$$

and we do not have equality in Lemma 6.1(i).

REFERENCES

[1] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[2] R. COMINETTI AND R. CORREA, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.
[3] H. FRANKOWSKA, *Hamilton–Jacobi equation: viscosity solutions and generalized gradients*, J. Math. Anal. Appl., 141 (1989), pp. 21–26.
[4] U. G. HAUSSMANN, *A probabilistic approach to the generalized Hessian*, Math. Oper. Res., 17 (1992), pp. 411–443.
[5] U. G. HAUSSMANN AND J.-P. LEPELTIER, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.
[6] I. KARATZAS AND S. E. SHREVE, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[7] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.
[8] D. H. WAGNER, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859–906.
[9] X. ZHOU, *A unified treatment of maximum principle and dynamic programming in optimal stochastic controls*, Stochastics Stochastics Rep., 36 (1991), pp. 137–161.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 744–789, May 1994

008

ON THE NONLINEAR DYNAMICS OF FAST FILTERING ALGORITHMS*

CHRISTOPHER I. BYRNES†, ANDERS LINDQUIST‡, AND YISHAO ZHOU§

Abstract. The main purpose of this paper is to address a fundamental open problem in linear filtering and estimation, namely, what is the steady-state or asymptotic behavior of the Kalman filter, or the Kalman gain, when the observed stationary stochastic process is not generated by a finite-dimensional stochastic system, or when it is generated by a stochastic system having higher-dimensional unmodeled dynamics. For example, some time ago Kalman pointed out that the usual positivity conditions assumed in the classical situation are not in fact necessary for the Kalman filter to converge. Using a "fast filtering" algorithm, which incorporates the statistics of the observation process as initial conditions for a dynamical system, this question is analyzed in terms of the phase portrait of a "universal" nonlinear dynamical system. This point of view has additional advantages as well, since it enables one to use the theory of dynamical systems to study the sensitivity of the Kalman filter to (small) changes in initial conditions; e.g., to changes in the statistics of the underlying process. This is especially important since these statistics are often either approximated or estimated. In this paper, for a scalar observation process, necessary and sufficient conditions for the Kalman filter to converge are derived using methods from stochastic systems and from nonlinear dynamics—especially the use of stable, unstable, and center manifolds. It is also shown that, in nonconvergent cases, there exist periodic points of every period p, $p \geq 3$ that are arbitrarily close to initial conditions having unbounded orbits, rigorously demonstrating that the Kalman filter can also be "sensitive to initial conditions."

Key words. Kalman filtering, fast filtering algorithms, Riccati equations, nonlinear dynamics, dynamical systems, power method, Lagrange-Grassmannian manifolds

AMS subject classifications. 93E11, 93B27, 58F40

1. Introduction.

Given a scalar stationary stochastic process $\{y_0, y_1, y_2, \ldots\}$ that is the output of a linear, finite-dimensional stochastic system driven by white noise, it is well known that the minimum variance estimate $\hat{x}_t$ of the current state x_t of the system is generated by the Kalman filter. Indeed, the Kalman filter is a model of the unforced stochastic system driven by a term consisting of the current output estimation error amplified by the so-called "Kalman gain" k_t, which itself can be determined "off-line" by solving a matrix Riccati equation. In this case, the steady-state behavior of both the Riccati equation and the Kalman filter is well understood. The purpose of this paper is to address a fundamental open problem concerning filtering and estimation, namely, what is the steady-state or asymptotic behavior of the Kalman filter, or the Kalman gain, when the stochastic process $\{y_t\}$ is not generated by a stochastic system, or when it is generated by a stochastic system having higher-dimensional, unmodeled dynamics? This question has been raised, for example, by Kalman, who pointed out that the positivity constraints associated with the existence of a stochastic system realizing $\{y_t\}$ might not be necessary for the Kalman filter to converge, a fact rigorously established for first-order systems in [5]

* Received by the editors July 22, 1991; accepted for publication (in revised form) September 25, 1992. This research was supported in part by grants from the Air Force Office of Scientific Research, the National Science Foundation, the Swedish Board for Technical Development, the Göran Gustafsson Foundation, and Southwestern Bell.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130 (byrnes@cec1.wustl.edu).

‡ Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden (alq@math.kth.se).

§ Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden (yishao@math.kth.se).

and for two-dimensional systems in [6]. Indeed, in [5] a complete phase portrait of the Kalman gain and the Kalman filter, as a dynamical system, was derived for first-order systems.

The basis for this analysis of the Kalman filtering as a dynamical system was the formulation [25] of "fast filtering" algorithms two decades ago. Instead of determining the n-vector k_t by first solving a matrix Riccati equation for a symmetric matrix P_t involving $n(n+1)/2$ variables, the fast filtering algorithm involves solving only a system of $2n$ equations, which consist of a dynamical system propagating k_t and an "adjoint" vector k_t^*. Moreover, as first shown in [27] and crucial for our dynamical systems analysis of the Kalman filter, this dynamical system can be reformulated so that the statistics of the process $\{y_t\}$ enter into the fast filtering algorithm only as initial conditions. Thus, we can analyze the asymptotic behavior of the Kalman filter for different statistics in terms of the phase portrait of the fast filtering algorithm. This is in sharp contrast to analysis of the Riccati equation as a dynamical system, since different statistics lead to different Riccati equations and, in fact, not to different initial conditions.

This point of view has additional advantages as well, since it enables us to study the sensitivity of the Kalman filter to (small) changes in initial conditions; e.g., to changes in the underlying system $\{y_t\}$ or its statistics. This is especially important since the statistics of the underlying process are often either approximated or estimated. In this direction, for the first-order case, necessary and sufficient conditions for asymptotic convergence of k_t were discovered [5], verifying the expectation that the Kalman filter would indeed converge for a much larger set of initial conditions or "initial statistics" than the classical theory predicts. In the complement of this set of (convergent) initial conditions, it was shown that there existed infinitely many periodic points of each period p, $p \geq 3$. Moreover, arbitrarily close to each of these periodic initial conditions are initial conditions for trajectories that are unbounded. For this reason, in the complement of the set of convergent initial conditions the Kalman filter is sensitive to initial conditions.

In this paper, for nth-order filtering problems we derive a systems theoretic necessary and sufficient condition on the process $\{y_t\}$ for the sequence of Kalman gains, k_t, to converge to a classical limit. En route to this result, we must develop a good understanding of the phase portrait of the fast filtering algorithm as a nonlinear dynamical system, including the determination (via spectral factorization) of a complete set of analytic invariant integrals. This, in turn, requires the extension of the several classical and more recent results concerning positive real transfer functions, positive semidefinite Toeplitz forms, and spectral factorization to situations where the relevant positivity conditions are not necessarily satisfied. Indeed, one of the main themes of this paper is that several important results classically conceived in terms of certain positivity conditions actually hold in a more universal context. While our main interest in this phenomenon lies in characterizing when the Kalman filter converges to a classical limit, this theme is of course quite old. For example, Hurwitz's derivation [20] of the Routh–Hurwitz criterion actually computed the difference between left-half plane and right-half plane zeros (or poles) as the signature of a Hankel matrix, while the Routh–Hurwitz conditions are simply the inequalities reflecting the positivity of this Hankel matrix. A more recent, and more relevant, example is the relaxation of the positive real conditions in circuit synthesis in the development of modern realization theory, based on rationality of transfer functions, or on rank conditions on Hankel matrices.

The paper is organized as follows. In §2, we set notation and recall some preliminary results needed throughout the paper. We begin §3 by reviewing part of the important relationship between shaping filters and Toeplitz forms, both for positive real transfer functions and in general. This relationship then enables us to extend an elegant parameterization, discovered by Kimura [22] and by Georgiou [16], of positive real transfer functions in terms of Szegö polynomials to a parameterization of all rational transfer functions. Just as the Kimura–Georgiou parameterization plays an important role in the covariance extension problem, this generalized parameterization plays an essential role in analyzing the global asymptotic behavior of the Kalman filter. This generalized Kimura–Georgiou parameterization is in fact a bona fide (birational) change of coordinates, as we show in §4. We then express the fast filtering algorithm and (what turns out to be a complete set of) its analytic invariant integrals in this new coordinate system. We begin §5 with a brief introduction to stable, unstable, and center manifold theory and its application to local stability analysis. After calculating the dimensions of these invariant manifolds at an equilibrium of the fast filtering algorithm, we show that the level sets of the invariant integrals defined above locally define smooth submanifolds near the equilibria, and we identify the invariant manifolds in terms of these invariants.

In §§6 and 7 we turn to the problem of global convergence of the fast filtering algorithm. In terms of the basic invariant integrals, it is easy to determine a system theoretic necessary condition for an initial condition to generate a trajectory of the fast filtering algorithm, which converges to a classical limit. This condition, derived from a spectral factorization argument, is simply that a certain pseudopolynomial be sign-definite on the unit circle. Moreover, the local stability analysis carried out in §5 shows that, for initial conditions sufficiently near an equilibrium, this necessary condition is also sufficient locally. Our main result, Theorem 7.1, asserts that this is also true in the large: except for a thin set of points that escape in finite time (and that can be explicitly characterized), a necessary and sufficient condition for global convergence of the Kalman gain k_t to a limit k_∞ is sign definiteness of the corresponding pseudopolynomial. The proof is based on a well-known interpretation of fast filtering algorithms and an equivalent Riccati equation as a dynamical system evolving on a Lagrangian Grassmannian. We conclude the paper in §8 with a series of examples and simulations for first- and second-order systems.

2. Preliminaries. Let $v(z)$ be a proper rational function of degree n with a minimal realization

$$v(z) = \tfrac{1}{2} + h'(zI - F)^{-1}g \tag{2.1}$$

(where $F \in \mathbb{R}^{n\times n}, g, h \in \mathbb{R}^n$ and prime denotes transpose) and consider the corresponding matrix Riccati equation

$$P_{t+1} = FP_tF' + (g - FP_th)(1 - h'P_th)^{-1}(g - FP_th)' \tag{2.2}$$

having an orbit of symmetric matrices $\{P_1, P_2, P_3, \ldots\}$ for each symmetric $P_0 \in \mathbb{R}^{n\times n}$.

If $v(z)$ is *positive real*, i.e.,

$$v(z) \text{ analytic on } |z| \geq 1, \tag{2.3a}$$

$$v(z) + v(1/z) > 0 \quad \text{on } |z| = 1, \tag{2.3b}$$

then

$$\Phi(e^{i\omega}) = v(e^{i\omega}) + v(e^{-i\omega}) \tag{2.4}$$

is the spectral density of a stationary stochastic process $\{y_t; t \in \mathbb{Z}\}$ that can be represented (in uncountably many ways) by a minimal stochastic realization

$$\begin{aligned} x_{t+1} &= Fx_t + v_t, \\ y_t &= h'x_t + w_t \end{aligned} \tag{2.5}$$

of y, i.e., a stochastic system with $E\{x_{t+1}y_t\} = g$ obtained by passing white noise $\{v_t, w_t\}$ through a shaping filter, the transfer function of which

$$w(z) = h'(zI - F)^{-1}B + d' \tag{2.6}$$

(where B is a matrix and d a vector of appropriate dimensions) is a minimal stable spectral factor of Φ, i.e., $w(z)w(1/z)' = \Phi(z)$. In general $w(z)$ is a row vector valued rational function. If, in particular, $w(z)$ is a scalar and both its numerator and denominator polynomials are stable (all zeros in the open unit disc), we say that $w(z)$ is a *minimum phase spectral factor*. All such realizations (2.6) have the same *Kalman filter*

$$\hat{x}_{t+1} = F\hat{x}_t + k_t(y_t - h'\hat{x}_t); \qquad \hat{x}_0 = 0 \tag{2.7}$$

(where $\hat{x}_t$ is the linear minimum-variance estimate of x_t given $\{y_0, y_1, \dots, y_{t-1}\}$), and the gain

$$k_t = (1 - h'P_th)^{-1}(g - FP_th) \tag{2.8}$$

is determined by solving the corresponding matrix Riccati equation (2.2) with initial condition

$$P_0 = 0. \tag{2.9}$$

It is well known that, under these conditions, P_t tends monotonically to the stable equilibrium of (2.2) [13], [26]. The question addressed in this paper is what happens to the solution (2.2) when the parameters have been chosen such that $v(z)$ is no longer positive real.

Without loss of generality, we henceforth take (F, g, h) in the observer canonical form

$$F = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 \\ -a_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_{n-1} & 0 & 0 & \cdots & 1 \\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad g = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}, \quad h = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{2.10}$$

in terms of which we may write $F = J - ah'$, where a is the column vector $(a_1, a_2, \dots, a_n)'$ and J is the obvious shift matrix. Consequently, the Riccati equation is determined by the $2n$ parameters (a, g) and there are also the coefficients of the rational function $v(z)$, i.e.,

$$v(z) = \frac{1}{2} + \frac{g_1z^{n-1} + g_2z^{n-2} + \cdots + g_n}{z^n + a_1z^{n-1} + \cdots + a_n}. \tag{2.11}$$

For simplicity we write

$$v(z) = \frac{1}{2} + \frac{g(z)}{a(z)} = \frac{1}{2}\frac{b(z)}{a(z)}, \tag{2.12}$$

where $b(z) := a(z) + 2g(z)$ is a monic polynomial of degree n. It is easy to see that, if $v(z)$ is positive real, then

$$D(z, z^{-1}) = \tfrac{1}{2}[a(z)b(1/z) + a(1/z)b(z)] > 0 \quad \text{on } |z| = 1, \tag{2.13}$$

$$a(z) \text{ has all its zeros in } |z| < 1, \tag{2.14}$$

$$b(z) \text{ has all its zeros in } |z| < 1. \tag{2.15}$$

Conversely, if (2.13) plus either (2.14) or (2.15) hold, then $v(z)$ is positive real.

To determine the Kalman filter we can, instead of the Riccati equation (2.2), use the algorithm

$$a(t+1) = \frac{1}{1 - g_1(t)}[a(t) + (I - J)g(t)]; \qquad a(0) = a \tag{2.16a}$$

$$g(t+1) = \frac{1}{1 - g_1(t)^2}[-g_1(t)a(t) + (J - g_1(t)I)g(t)]; \quad g(0) = g \tag{2.16b}$$

consisting of $2n$ nonlinear first-order difference equations in terms of which

$$k_t = a(t) + g(t) - a. \tag{2.17}$$

This algorithm is a version, appearing in [27], of the fast Kalman filtering algorithm introduced in [25]. (Also see [5] where these matters are reviewed; in the notation of this paper $a(t) = q_t - q_t^*$.) Suppose $r_t := \prod_{k=0}^{t-1}[1 - g_1(k)^2]$ and the monic polynomials $a_t(z)$ and $b_t(z) := a_t(z) + 2g_t(z)$ are formed from $a(t)$ and $b(t) := a(t) + 2g(t)$ as above, then it is shown in [27] that the equality

$$r_t[a_t(z)b_t(1/z) + a_t(1/z)b_t(z)] = 2D(z, z^{-1}) \tag{2.18}$$

is preserved along the trajectory of (2.16). It is also shown in [27] that $a_t(z)$ and $b_t(z)$ have all their zeros in the unit disc $|z| < 1$. Consequently, if $v(z)$ is positive real, then so is

$$v_t(z) = \frac{1}{2}\frac{b_t(z)}{a_t(z)} = \frac{1}{2} + \frac{g_t(z)}{a_t(z)} \tag{2.19}$$

for each $t = 1, 2, 3, \ldots$, so that each $(a(t), g(t))$ is an admissible pair of parameters for the Kalman filtering problem, corresponding to stochastic systems.

3. Systems theoretic enhancements of some classical positivity results. One of the main results of this paper is to establish and analyze the fact that filtering algorithms do converge for parameter values that do not correspond to a bona fide stochastic system and, hence, that do not satisfy the relevant positivity conditions. These positivity conditions can be expressed in terms of a transfer function being positive real or a Toeplitz matrix being positive definite, as well as a number of other conditions involving familiar objects from classical analysis and systems theory. Our main result is just one manifestation of the fact that several classical and more recent results containing positivity and positive real functions have, of course, natural enhancements to statements concerning broader classes of nonsingular matrices and systems. In this section we develop this theme in the context of several particular results that we will find to be very useful in the remaining sections.

It is well known that a function $v(z)$ with the Laurent expansion

$$v(z) = \tfrac{1}{2} + c_1 z^{-1} + c_2 z^{-2} + c_3 z^{-3} + \cdots \tag{3.1}$$

around $z = \infty$ is positive real if and only if the Toeplitz matrices

$$T_t = \begin{bmatrix} 1 & c_1 & \cdots & c_t \\ c_1 & 1 & \cdots & c_{t-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_t & c_{t-1} & \cdots & 1 \end{bmatrix} \tag{3.2}$$

are positive definite for all $t = 1, 2, 3, \ldots$.

A simpler test of positive realness due to Schur [32] can be described in terms of the Szegö polynomials $\{\varphi_0(z), \varphi_1(z), \varphi_2(z), \ldots\}$, a sequence of monic polynomials

$$\varphi_t(z) = z^t + \varphi_{t1} z^{t-1} + \cdots + \varphi_{tt} \tag{3.3}$$

that are orthogonal on the unit circle. Similarly, we define the reversed polynomial $\varphi_t^*(z)$ as

$$\varphi_t^*(z) = \varphi_{tt} z^t + \varphi_{t,t-1} z^{t-1} + \cdots + 1. \tag{3.4}$$

The Szegö polynomials are then determined from the sequence $\{c_1, c_2, c_3, \ldots\}$ through the polynomial recursions

$$\begin{aligned} \varphi_{t+1}(z) &= z\varphi_t(z) - \gamma_t \varphi_t^*(z); \qquad \varphi_0(z) = 1, \\ \varphi_{t+1}^*(z) &= \varphi_t^*(z) - \gamma_t z \varphi_t(z); \qquad \varphi_0^*(z) = 1, \end{aligned} \tag{3.5}$$

where $\{\gamma_0, \gamma_1, \gamma_2, \ldots\}$ are the *Schur parameters*

$$\gamma_t = \frac{1}{r_t} \sum_{k=0}^{t} \varphi_{t,t-k} c_{k+1} \tag{3.6}$$

and $\{r_0, r_1, r_2, \ldots\}$ are given by the recursion

$$r_{t+1} = (1 - \gamma_t^2) r_t; \qquad r_0 = 1, \tag{3.7}$$

the algorithm terminating if $|\gamma_t|$ becomes one. Indeed, it has been shown by Schur that

$$T_t > 0 \Leftrightarrow |\gamma_k| < 1 \quad \text{for } k = 0, 1, 2, \ldots, t-1. \tag{3.8}$$

It is also classical that the function (3.1) has an infinite *Schur parameter sequence* $\{\gamma_0, \gamma_1, \gamma_2, \ldots\}$ if and only if $|\gamma_t|$ never becomes one—otherwise, the Schur parameter sequence is finite, ending with a term of modulus one—and that for each $t = 1, 2, 3, \ldots$, there is a one-to-one correspondence between the set of all subsequences $\{c_1, c_2, \ldots, c_t\}$ such that T_k is nonsingular for $k = 1, 2, \ldots, t$ and the set of all subsequences $\{\gamma_0, \gamma_1, \ldots, \gamma_{t-1}\}$ such that $|\gamma_k| \neq 1$ for $k = 0, 1, 2, \ldots, t-2$.

That these claims also hold for nonpositive data follows from the following well-known enhancement of the positive result (3.8).

PROPOSITION 3.1. $\det T_t = \prod_{k=0}^{t} r_k$.

As a second illustration of this theme, Kimura [22] and Georgiou [16] have independently shown that to any positive real function (3.1) with the first n c-coefficients prescribed, or alternatively with $\gamma := (\gamma_0, \gamma_1, \ldots, \gamma_{n-1})'$ fixed, there is a unique vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)'$ of real numbers such that

$$v(z) = \frac{1}{2} \frac{\psi_n(z) + \alpha_1 \psi_{n-1}(z) + \cdots + \alpha_n \psi_0(z)}{\varphi_n(z) + \alpha_1 \varphi_{n-1}(z) + \cdots + \alpha_n \varphi_0(z)}, \tag{3.9}$$

where $\{\psi_0, \psi_1, \psi_2, \dots\}$ are the Szegö polynomials obtained by exchanging the Schur parameters $\{\gamma_t\}$ by $\{-\gamma_t\}$. This is a useful parameterization for the *rational covariance extension problem* [21], but, as we now demonstrate, (3.9) is actually a general interpolation formula that holds regardless of whether $v(z)$ is positive real, provided that the algorithm does not terminate for $t < n$. In fact, it follows that there is a one-to-one correspondence between the open dense set in $\mathbb{R}^{2n}$ of $2n$ parameters (α, γ) for which none of the elements of the vector $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_{n-1})' \in \mathbb{R}^n$ has modulus one, and the corresponding open dense set of $(a, g) \in \mathbb{R}^{2n}$.

THEOREM 3.2. *Let $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_{n-1})'$ be an arbitrary vector in $\mathbb{R}^n$ such that $\gamma_k^2 \neq 1$ for $k = 0, 1, \dots, n-2$, let $\{\varphi_k(z), \psi_k(z); k = 0, 1, \dots, n-1\}$ be the corresponding polynomials generated by* (3.5), *and set $c_1 := \gamma_0$ and*

$$c_{k+1} := r_k \gamma_k - \sum_{j=0}^{k-1} \varphi_{k,k-j}\, c_{j+1} \tag{3.10}$$

for $k = 1, 2, \dots, n-1$, where $r_1, r_2, \dots, r_n$ are defined by (3.7). *Let $a(z)$ and $b(z)$ be arbitrary monic polynomials of degree n such that*

$$\frac{b(z)}{2a(z)} = \frac{1}{2} + c_1 z^{-1} + c_2 z^{-2} + \cdots + c_n z^{-n} + \cdots. \tag{3.11}$$

Then there is a unique $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)' \in \mathbb{R}^n$ such that

$$a(z) = \varphi_n(z) + \alpha_1 \varphi_{n-1}(z) + \cdots + \alpha_n, \tag{3.12a}$$

$$b(z) = \psi_n(z) + \alpha_1 \psi_{n-1}(z) + \cdots + \alpha_n. \tag{3.12b}$$

The proof of Theorem 3.2 is based on the following lemma.

LEMMA 3.3. *Let the polynomials $\{\varphi_k(z), \psi_k(z);\ k = 0, 1, \dots, n-1\}$ and the sequence $\{c_1, c_2, \dots, c_n\}$ be as defined in Theorem* 3.2. *Then*

$$\Psi_{n+1} = C_{n+1} \Phi_{n+1}, \tag{3.13}$$

where Φ, Ψ, and C are the nonsingular $(n+1) \times (n+1)$-matrices

$$\Phi_{n+1} = \begin{bmatrix} 1 & & & & \\ \varphi_{n1} & 1 & & & \\ \varphi_{n2} & \varphi_{n-1,1} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \varphi_{nn} & \varphi_{n-1,n-1} & \varphi_{n-2,n-2} & \cdots & 1 \end{bmatrix}, \tag{3.14a}$$

$$\Psi_{n+1} = \begin{bmatrix} 1 & & & & \\ \psi_{n1} & 1 & & & \\ \psi_{n2} & \psi_{n-1,1} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \psi_{nn} & \psi_{n-1,n-1} & \psi_{n-2,n-2} & \cdots & 1 \end{bmatrix}, \tag{3.14b}$$

$$C_{n+1} = \begin{bmatrix} 1 & & & & \\ 2c_1 & 1 & & & \\ 2c_2 & 2c_1 & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ 2c_n & 2c_{n-1} & 2c_{n-2} & \cdots & 1 \end{bmatrix}. \tag{3.14c}$$

Proof. We want to prove that

$$\psi_{tk} = 2c_k + 2c_{k-1}\varphi_{t1} + 2c_{k-2}\varphi_{t2} + \cdots + 2c_1\varphi_{t,k-1} + \varphi_{tk}$$

for all $t \geq k$, or, equivalently,

$$\rho_{tk} = c_k + c_{k-1}\varphi_{t1} + c_{k-2}\varphi_{t2} + \cdots + c_1\varphi_{t,k-1} \tag{3.15}$$

for all $t = k, k+1, \ldots, n$, where $\{\rho_{tk}\}$ are the coefficients of the polynomials

$$\rho_t(z) = \tfrac{1}{2}[\psi_t(z) - \varphi_t(z)]. \tag{3.16}$$

Then, the recursions in φ_t and ψ_t imply that

$$\rho_{t+1}(z) = z\rho_t(z) + \gamma_t\pi_t^*(z), \tag{3.17}$$

where $\pi_t^*(z)$ is the reversed polynomial of

$$\pi_t(z) = \tfrac{1}{2}[\psi_t(z) + \varphi_t(z)], \tag{3.18}$$

i.e., $\pi_t^*(z) := z^n\pi_t(1/z)$. We also recall from the literature [24], [1], [17] that the coefficients of $\{\varphi_t\}$ satisfy the *normal equations*

$$\begin{bmatrix} 1 & c_1 & \cdots & c_{t-1} \\ c_1 & 1 & \cdots & c_{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{t-2} & c_{t-3} & \cdots & c_1 \\ c_{t-1} & c_{t-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \varphi_{tt} \\ \varphi_{t,t-1} \\ \vdots \\ \varphi_{t1} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ r_t \end{bmatrix} \tag{3.19}$$

having the Toeplitz matrix T_t as its coefficient matrix. As we have pointed out above, T_t is nonsingular if our basic assumption that $|\gamma_k| \neq 1$ for all $k = 0, 1, 2, \ldots, t-1$ holds [1]. It follows from (3.5) that $\varphi_{tt} = -\gamma_{t-1}$, and consequently $\psi_{tt} = \gamma_{t-1}$, and thus $\rho_{tt} = \varphi_{tt}$. Therefore, we see from the first equation (3.19) that (3.17) holds for $t = k$. We now proceed by induction. Suppose (3.15) holds for $t = s$, where $k \leq s \leq n-1$. We want to prove that it holds for $t = s+1$. To this end, note that for $t = s$ one of the normal equations reads

$$c_k + c_{k-1}\varphi_{s1} + c_{k-2}\varphi_{s2} + \cdots + c_1\varphi_{s,k-1} + \varphi_{sk} + c_1\varphi_{s,k+1} + \cdots + c_{s-k}\varphi_{ss} = 0.$$

However, in view of the induction hypothesis, this can be written

$$\rho_{sk} + \varphi_{sk} + c_1\varphi_{s,k+1} + \cdots + c_{s-k}\varphi_{ss} = 0$$

and therefore, since $\pi_{sk} = \rho_{sk} + \varphi_{sk}$,

$$\pi_{sk} = -c_1\varphi_{s,k+1} - \cdots - c_{s-k}\varphi_{ss}. \tag{3.20}$$

Now, identifying coefficients in the polynomial recursion (3.18), we obtain

$$\rho_{s+1,k} = \rho_{sk} + \gamma_s\pi_{s,s+1-k},$$

which, after inserting (3.20) and applying the induction hypothesis, takes the form

$$\rho_{s+1,k} = c_k + c_{k-1}(\varphi_{s1} - \gamma_s\varphi_{ss}) + \cdots + c_1(\varphi_{s,k-1} - \gamma_s\varphi_{s,s+2-k}).$$

However, it follows from (3.5) that $\varphi_{s+1,k} = \varphi_{sk} - \gamma_s \varphi_{s,s+1-k}$, and therefore (3.15) holds for $t = s+1$ as required. Hence the lemma follows by induction. □

Proof of Theorem 3.2. Since $\{\varphi_t\}$ and $\{\psi_t\}$ are families of monic polynomials of increasing degree t, there are $\alpha, \beta \in \mathbb{R}^n$ such that

$$a(z) = \varphi_n(z) + \alpha_1 \varphi_{n-1}(z) + \cdots + \alpha_n,$$
$$b(z) = \psi_n(z) + \beta_1 \psi_{n-1}(z) + \cdots + \beta_n.$$

Then (3.11) yields

$$\psi_n(z) + \beta_1 \psi_{n-1}(z) + \cdots + \beta_n$$
$$= [\varphi_n(z) + \alpha_1 \varphi_{n-1}(z) + \cdots + \alpha_n][1 + 2c_1 z^{-1} + 2c_2 z^{-2} + \cdots].$$

Therefore, identifying coefficients of nonnegative powers of z, we have

$$\Psi_{n+1} \begin{bmatrix} 1 \\ \beta \end{bmatrix} = C_{n+1} \Phi_{n+1} \begin{bmatrix} 1 \\ \alpha \end{bmatrix},$$

which, by Lemma 3.3, implies that $\beta = \alpha$. □

COROLLARY 3.4. *Consider the maps* $\gamma := (\gamma_0, \gamma_1, \ldots, \gamma_{n-1}) \to \Phi_{n+1}(\gamma)$ *and* $\gamma \to \Psi_{n+1}(\gamma)$ *defined through* (3.5), *the corresponding recursion for* $\{\psi_t\}$, (3.14a) *and* (3.14b). *Then* $\Psi_{n+1}(\gamma) = \Phi_{n+1}(-\gamma)$ *and* $\Phi_{n+1}(0) = \Psi_{n+1}(0) = I_{n+1}$. *Moreover,*

$$\begin{bmatrix} 1 \\ a \end{bmatrix} = \Phi_{n+1}(\gamma) \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \qquad \begin{bmatrix} 1 \\ b \end{bmatrix} = \Psi_{n+1}(\gamma) \begin{bmatrix} 1 \\ \alpha \end{bmatrix}. \tag{3.21}$$

4. The fast filtering algorithm and its invariant integrals. One of the principal goals of this section is to express the fast filtering algorithm (2.16) in a more convenient way in terms of the parameters (α, γ) entering in the generalization, Theorem 3.2, of the Kimura–Georgiou parameterization of positive real systems. As a preliminary step, we first show that this parameterization constitutes in fact a bona fide change of coordinates. In the language of classical algebraic geometry, the map defined by (3.12) is a birational isomorphism [33]. More explicitly, consider the set

$$U_\gamma = \{(\alpha, \gamma) \in \mathbb{R}^{2n} \mid \gamma_i^2 \neq 1, i = 0, 1, \ldots, n-2\}.$$

Also, by virtue of (3.11), the generalized "correlation" coefficients $c_1, c_2, \ldots, c_n$ are functions of (a, b) so that we may define the open, dense set

$$V_c = \{(a, b) \in \mathbb{R}^{2n} \mid \det T_i \neq 0, i = 1, 2, \ldots, n-1\}.$$

We show that that the polynomial map $\mathcal{F}$ is a bijection of U_γ with V_c having a rational inverse so that $\mathcal{F}$ is indeed a birational isomorphism.

PROPOSITION 4.1. *The map* $\mathcal{F}$, *defined by* (3.12), *sending* $(\alpha, \gamma) \in \mathbb{R}^{2n}$ *to* $(a, b) \in \mathbb{R}^{2n}$ *is a polynomial map given by*

$$a = \varphi_n(\gamma) + \Phi_n(\gamma)\alpha,$$
$$b = \psi_n(\gamma) + \Psi_n(\gamma)\alpha, \tag{4.1}$$

where $\varphi_n := (\varphi_{n1}, \varphi_{n2}, \ldots, \varphi_{nn})'$ *and* $\psi_n := (\psi_{n1}, \psi_{n2}, \ldots, \psi_{nn})'$, *and* $\Phi_n(\gamma)$ *and* $\Psi_n(\gamma)$ *are given by* (3.14). *Moreover,* $\mathcal{F} : U_\gamma \to V_c$ *is a bijection with a rational inverse* $\mathcal{F}^{-1}$.

Proof. On V_c the map $\mathcal{F}$ has an inverse $\mathcal{F}^{-1}$ defined in the following way: (a,b) defines through (3.11) a sequence $\{c_1,c_2,\dots,c_n\}$ that, by (3.6), corresponds to a vector $\gamma=(\gamma_0,\gamma_1,\dots,\gamma_{n-1})'$, from which in turn the polynomials $\{\varphi_k(z);k=0,1,2,\dots,n-1\}$ can be defined. Then $\alpha\in\mathbb{R}^n$ is uniquely determined by (3.12a). Finally, (4.1) follows from (3.21). □

Recall from [27] or from [5] that if $\{\gamma_0,\gamma_1,\gamma_2,\dots\}$ is the (infinite or finite) Schur parameter sequence of (2.1), as defined in §3, then

$$\gamma_t=g_1(t)\qquad t=0,1,2,\dots, \tag{4.2}$$

where g_1 is generated by the fast filtering algorithm (2.16). A key observation now is that (2.16) is a time-invariant dynamical system in parameter space. In particular, let us stress the following simple but important observation.

LEMMA 4.2. *Let $v(z)$ be defined as in §2 and let $\{\gamma_0,\gamma_1,\gamma_2,\dots\}$ be its (infinite or finite) Schur parameter sequence. Then, for each $t=0,1,2,\dots$, as long as the algorithm (2.16) does not escape, $v_t(z)$ defined by (2.19) has the Schur parameter sequence $\{\gamma_t,\gamma_{t+1},\gamma_{t+2},\dots\}$.*

Proof. The fast filtering algorithm (2.16) is a time-invariant dynamical system, and unless it has escaped it will therefore trivially generate via (4.2) the sequence $\{\gamma_t,\gamma_{t+1},\gamma_{t+2},\dots\}$ if initialized at $(a(t),g(t))$ corresponding to $v_t(z)$. □

Corollary 3.4 allows us to change coordinates in the fast algorithm, expressing it instead in terms of (α,γ), where

$$\alpha=\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n\end{bmatrix},\qquad \gamma=\begin{bmatrix}\gamma_0\\ \gamma_1\\ \vdots\\ \gamma_{n-1}\end{bmatrix}, \tag{4.3}$$

as long as $(\alpha,\gamma)\in U_\gamma$, i.e., as long as $\{\gamma_0,\gamma_1,\dots,\gamma_{n-1}\}$ is the initial subsequence of a Schur parameter sequence.

THEOREM 4.3. *Let the rational function $v(z)$ defined by (2.1) have a Schur parameter sequence such that $\gamma_k^2\neq 1$ for $k=0,1,\dots,n-1$. Then the fast filtering algorithm takes the form*

$$\alpha(t+1)=A(\gamma(t))\alpha(t),\qquad \alpha(0)=\alpha \tag{4.4a}$$

$$\gamma(t+1)=G(\alpha(t+1))\gamma(t),\qquad \gamma(0)=\gamma \tag{4.4b}$$

in the coordinates of the generalized Kimura–Georgiou parameterization, where the maps $A,G:\mathbb{R}^n\to\mathbb{R}^{n\times n}$ are defined as

$$A(\gamma)=\begin{bmatrix}\frac{1}{1-\gamma_{n-1}^2} & \frac{\gamma_{n-1}\gamma_{n-2}}{(1-\gamma_{n-1}^2)(1-\gamma_{n-2}^2)} & \cdots & \frac{\gamma_{n-1}\gamma_0}{(1-\gamma_{n-1}^2)\cdots(1-\gamma_0^2)}\\ 0 & \frac{1}{1-\gamma_{n-2}^2} & \cdots & \frac{\gamma_{n-2}\gamma_0}{(1-\gamma_{n-2}^2)\cdots(1-\gamma_0^2)}\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{1}{1-\gamma_0^2}\end{bmatrix} \tag{4.5}$$

and

$$G(\alpha)=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \cdots & -\alpha_1\end{bmatrix}. \tag{4.6}$$

More precisely, if (α, γ) *are the parameters of* $v(z)$ *in the representation* (3.9), *then* $(\alpha(t), \gamma(t))$ *are the parameters of* $v_t(z)$, *as defined in* (2.19), *for each point in the finite or infinite orbit of* (α, γ). *Moreover, if* $\{\gamma_0, \gamma_1, \gamma_2, \dots\}$ *is the sequence of Schur parameters of* $v(z)$, *then the sequence* $\{\gamma_t, \gamma_{t+1}, \gamma_{t+2}, \dots\}$ *obtained by deleting the first* t *elements is the Schur parameters sequence of* $v_t(z)$. *In fact,*

$$\gamma_k(t) = \gamma_{t+k} \tag{4.7}$$

and therefore the Schur parameters are updated according to the recursion

$$\gamma_{t+n} = -\alpha_1(t+1)\gamma_{t+n-1} - \alpha_2(t+1)\gamma_{t+n-2} - \cdots - \alpha_n(t+1)\gamma_t. \tag{4.8}$$

Finally, the gain sequence $\{k_0, k_1, k_2, \dots, \}$ *of the Kalman filter is given by*

$$k_t = \Pi_n(\gamma(t))\alpha(t) + \pi_n(\gamma(t)) - \Phi_n(\gamma)\alpha - \varphi_n(\gamma), \tag{4.9}$$

where $\varphi_n(\gamma)$ *and* $\psi_n(\gamma)$ *are* n*-vectors of coefficients of* $\varphi(z)$ *and* $\psi(z)$, *and*

$$\Pi_n(\gamma) = \tfrac{1}{2}[\Phi_n(\gamma) + \Psi_n(\gamma)],$$
$$\pi_n(\gamma) = \tfrac{1}{2}[\varphi_n(\gamma) + \psi_n(\gamma)].$$

For a proof, we refer the reader to the Appendix.

In §5 we will show that the dynamical system (4.4) evolves on an invariant manifold X_D defined by the preserved pseudopolynomial (2.18), which we write in the form

$$D(z, z^{-1}) = d(z) + d(1/z), \tag{4.10a}$$

where

$$d(z) = \tfrac{1}{2}d_0 + d_1 z + d_2 z^2 + \cdots + d_n z^n. \tag{4.10b}$$

The symmetric pseudopolynomial D is determined by the initial condition (α, γ) in a manner described by the following lemma, the proof of which is given in the Appendix.

LEMMA 4.4. *Let* $D(z, z^{-1})$ *be the pseudopolynomial* (4.10) *corresponding to the initial condition* (α, γ). *Then*

$$d_0 = \alpha_n^2 + r_1\alpha_{n-1}^2 + \cdots + r_n, \tag{4.11}$$

where $r_1, r_2, \dots, r_n$ *are defined by* (3.7), *and* $d_i := d_i^{(n)}(\alpha, \gamma)$ *for* $i = 1, 2, \dots, n$, *where* $d_i^{(n)}$ *is determined recursively by*

$$\begin{aligned}
&d_1^{(1)}(\alpha_1, \gamma_0) = \alpha_1;\\
&d_i^{(k)}(\alpha_1, \dots, \alpha_k, \gamma_0, \dots, \gamma_{k-1}) = (1-\gamma_0^2)d_i^{(k-1)}(\alpha_1, \dots, \alpha_{k-1}, \gamma_1, \dots, \gamma_{k-1})\\
&\qquad +\alpha_k \textstyle\sum_{j=1}^{k} \alpha_{k-j}\pi_{j,j-i}, \quad \text{for } i = 1, 2, \dots, k-1;\\
&d_k^{(k)}(\alpha_1, \dots, \alpha_k, \gamma_0, \dots, \gamma_{k-1}) = \alpha_k,
\end{aligned}$$

where $\{\pi_{jl}\}$ *are the coefficients of the polynomials*

$$\pi_j(z) = z^j + \pi_{j1}z^{j-1} + \cdots + \pi_{jj}$$

generated by the polynomial recursion

$$\begin{aligned} &\pi_{t+1}(z) = (1+z)\pi_t(z) + (\gamma_t\gamma_{t-1} - 1)z\pi_{t-1}(z), \\ &\pi_0 = 1, \qquad \pi_1(z) = z, \end{aligned} \tag{4.12}$$

and $\pi_{ji} = 0$ *for* $i > j$. *Moreover, if* $\gamma_k^2 \neq 1$ *for* $k = 0, 1, \ldots, n-1$, *then at least one of the coefficients* $d_0, d_1, \ldots, d_n$ *of the pseudopolynomial* $D(z, z^{-1})$ *is nonzero.*

Comparing coefficients of $(z^i + z^{-i})$ in (2.18) we see that

$$r_t d_i(\alpha(t), \gamma(t)) = d_i(\alpha(0), \gamma(0)) \qquad i = 0, 1, 2, \ldots, n \tag{4.13}$$

for all $t \in \mathbb{Z}$ along the trajectory of the dynamical system (4.4). Hence the $n+1$ functions $d_i(\alpha, \gamma)$, $i = 0, 1, \ldots, n-1$ defined in Lemma 4.4 are invariant under the evolution of (4.4) up to a (common) scaling factor; i.e., these $(n+1)$ functions are projectively invariant. We can obtain n invariant quantities, either by viewing the pseudopolynomial, in terms of homogeneous coordinates, as a point in $\mathbb{P}^n$ (see [33]), or equivalently by dividing each of the $(n+1)$ equations in (4.13) by any one of the $(n+1)$ functions that is nonzero (by Lemma 4.4, there is always one), obtaining rational functions having values independent of r_t and hence depending only on (α, γ). That is, we can view the pseudopolynomial D either as determining $(n+1)$ projectively invariant functions $T_1, \ldots, T_{n+1}$ or as determining a map $\overline{T}$ to $\mathbb{P}^n$:

$$\begin{array}{ccc} \mathbb{R}^{2n} & \xrightarrow{T} & \mathbb{R}^{n+1} - \{0\} \\ & \overline{T} \searrow & \downarrow \Pi \\ & & \mathbb{P}^n, \end{array}$$

where $T = (T_1, \ldots, T_{n+1})$ and $\overline{T} = \Pi \circ T$ where

$$\Pi(x_1, \ldots, x_{n+1}) = [x_1, \ldots, x_{n+1}].$$

In this way we might expect (2.18) to define an n-fold X_D in $\mathbb{R}^{2n}$. Indeed, this analytic set will be a smooth n-manifold at a point (α, γ) provided $\operatorname{Jac} T\big|_{(\alpha,\gamma)}$ has an n-dimensional kernel. We return to this question in §§5 and 6 after having introduced some additional analytic tools.

From Theorem 4.3 and Lemma 4.4 it is clear that the fast filtering algorithm has a quasi-nested structure in the sense that whenever $\alpha_n = \alpha_{n-1} = \cdots = \alpha_{k+1} = 0$ but $\alpha_k \neq 0$, the dynamical system (4.4) and the invariant set X_D reduce to the k-dimensional case with the Schur parameter sequence shifted $n-k$ steps. (This is related to the occurrence of invariant directions [3] in the corresponding matrix Riccati equation (2.2) as pointed out in [27] and further elaborated on [31].) As explained in [25], [27], [26], [5], the fast algorithm is intimately connected to the Szegö orthogonal polynomial recursion (3.5), which in fact was the basic tool in the original derivation [25]. In view of this and the analysis above we would expect that it would also be connected to the Schur algorithm. Indeed, this has been shown in a recent paper [8].

5. Invariant manifolds and local convergence for the fast filtering algorithms. We now turn to a stability analysis of the equilibria of the fast filtering algorithms, expressed in the form (4.4), i.e.,

$$\begin{pmatrix} \alpha \\ \gamma \end{pmatrix}_{t+1} = f\begin{pmatrix} \alpha_t \\ \gamma_t \end{pmatrix}, \tag{5.1a}$$

where

$$f\begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = \begin{bmatrix} A(\gamma)\alpha \\ G(A(\gamma)\alpha)\gamma \end{bmatrix}. \tag{5.1b}$$

For the stability analysis of the fast filtering algorithms, we need the geometric concepts of stable, unstable, and center manifolds, which play a role for nonlinear systems similar to the role played by generalized eigenspaces for the stability analysis of linear systems. Because this role is so important in determining stability, especially in the critical case, we precede our analysis of the local stability of the fast filtering algorithms with an introductory discussion of local invariant manifolds for nonlinear systems. As supplementary references we recommend, among other texts, Guckenheimer and Holmes [18], and Marsden and McCracken [28].

At an equilibrium $(\alpha_\infty, \gamma_\infty)$ of (5.1),

$$\begin{bmatrix} \alpha - \alpha_\infty \\ \gamma - \gamma_\infty \end{bmatrix}_{t+1} = C \begin{bmatrix} \alpha - \alpha_\infty \\ \gamma - \gamma_\infty \end{bmatrix}_t + O(\|\alpha - \alpha_\infty\|^2 + \|\gamma - \gamma_\infty\|^2)$$

determines to first order a linear system

$$\begin{bmatrix} \bar{\alpha} \\ \bar{\gamma} \end{bmatrix}_{t+1} = C \begin{bmatrix} \bar{\alpha} \\ \bar{\gamma} \end{bmatrix}_t,$$

where $\bar{\alpha} = \alpha - \alpha_\infty$, $\bar{\gamma} = \gamma - \gamma_\infty$. Denote by s the number of eigenvalues of the matrix C having modulus less than one, counting roots of the characteristic polynomial with their algebraic multiplicities. Similarly, denote by u the number of eigenvalues having modulus greater than one and by c the number of eigenvalues having modulus one. It is well known that if $u \geq 1$, then (5.1) is unstable, so we suppose for the moment that $u = 0$. In this case, if $c = 0$, then $(\alpha_\infty, \gamma_\infty)$ is an asymptotically stable equilibrium for the system (5.1), with all solutions converging geometrically to the equilibrium. The critical case $c \neq 0$ is more subtle, even for linear systems where Lyapunov stability is determined by the geometric multiplicities of the eigenvalues lying on the unit circle.

Remarkably, the linear case can in fact be analyzed geometrically in a manner that can be adapted to the critical nonlinear case, mutatis mutandis. Denote by V^s the sum of the generalized eigenspaces corresponding to eigenvalues inside the unit disk, by V^u the sum of the generalized eigenspaces corresponding to eigenvalues outside the unit disk, and by V^c the sum of the generalized eigenspaces corresponding to eigenvalues lying on the unit circle. Then, we have

$$\dim V^s = s, \quad \dim V^u = u, \quad \text{and} \dim V^c = c.$$

In particular, there is a direct sum decomposition of the state space consisting of three invariant subspaces

$$\mathbb{R}^{2n} = V^s \oplus V^u \oplus V^c.$$

Moreover, the evolution of the entire linear system is a superposition of the three motions on the constituent invariant subspaces: the asymptotically stable motion on V^s, the asymptotically expanding motion on V^u, and the motion on V^c, which is determined by the dimension of the Jordan blocks corresponding to the eigenvalues of unit modulus. For example, if $u = 0$ as assumed above, it can be easily verified that any trajectory of the full linear system converges geometrically to a trajectory

lying on V^c. Therefore, if $u = 0$, the (asymptotic) stability or instability of the full linear system is determined by the (asymptotic) stability or instability of the reduced dynamics on V^c.

In the nonlinear case, the geometric situation is similar. It is now classical that the nonlinear analogue of V^s can be locally defined as the set W^s of initial conditions that converge to the equilibrium at a geometric rate. The set W^s, referred to as the *stable manifold*, is known to be locally invariant and to be locally a smooth submanifold of the state space, having dimension s. A similar characterization of the set of geometrically expanding points can be given, leading to the unstable manifold W^u, which is locally defined as an invariant, smooth submanifold of dimension u. In this context, it is easy to see that if $u = 0$ and if $c = 0$, then W^s is a neighborhood of the equilibrium and therefore the equilibrium is locally asymptotically stable. An analysis of the critical case, $u = 0$ but $c \neq 0$, is facilitated by the existence of a center manifold W^c that plays a role analogous to the role played for linear systems by V^c. The existence of a center manifold has been established only relatively recently in part due to the absence of an explicit characterization of W^c as a set, a fact that also partially explains the fact that center manifolds need not be unique. This existence result is only part of the fundamental "center manifold theorem," which we now describe in more detail.

CENTER MANIFOLD THEOREM ([18], [28]).

(i) Existence. *Suppose* (5.1) *is a* C^{k+1} *system with an equilibrium* $(\alpha_\infty, \gamma_\infty)$ *for which* $\dim V^c = c$. *Then, in a sufficiently small neighborhood of the equilibrium there exists a* C^k*-submanifold* W^c *of dimension* c, *which is locally invariant and for which the tangent space to* W^c *at* $(\alpha_\infty, \gamma_\infty)$ *is* V^c.

(ii) Principle of asymptotic phase. *Suppose further that* $u = 0$ *for the equilibrium* $(\alpha_\infty, \gamma_\infty)$. *Then, for each initial condition sufficiently close to the equilibrium there is an initial condition on* W^c *for which the error between the corresponding trajectories asymptotically decreases geometrically.*

We will use this theorem for convergence analysis of the fast filtering algorithm (5.1).

LEMMA 5.1. *The point* (α, γ) *is an equilibrium of the fast filtering algorithm* (5.1) *if and only if* $\gamma = 0$. *The Jacobian of* f *at the equilibrium* $(\alpha, 0)$ *is given by*

$$\operatorname{Jac} f\Big|_{(\alpha,0)} = \begin{bmatrix} I & 0 \\ 0 & G(\alpha) \end{bmatrix},$$

where $G(\alpha)$ *is defined as in* (4.6).

Proof. Since $A(0) = I$, $(\alpha, 0)$ is clearly an equilibrium for each $\alpha \in \mathbb{R}^n$. It remains to show that each equilibrium is of this form. To this end, let (α, γ) satisfy

$$\alpha = A(\gamma)\alpha, \tag{5.2a}$$

$$\gamma = G(\alpha)\gamma. \tag{5.2b}$$

The last of equations (5.2a) reads

$$\alpha_n = \frac{\alpha_n}{1 - \gamma_0^2},$$

which requires that either α_n or γ_0 is zero, i.e., $\alpha_n \gamma_0 = 0$. In view of this, the second to last equation becomes

$$\alpha_{n-1} = \frac{\alpha_{n-1}}{1 - \gamma_1^2},$$

which implies that $\alpha_{n-1}\gamma_1 = 0$. Proceeding in this manner we see that

$$\alpha_{n-i}\gamma_i = 0; \quad i = 0, 1, 2, \ldots, n-1$$

and therefore the last of equations (5.2b) yields $\gamma_{n-1} = 0$. Then solving (5.2b) successively from the bottom yields $\gamma = 0$ as required. Since $A(0) = I$ and $\partial A/\partial\gamma(0) = 0$, the Jacobian of f is as stated in the lemma. □

In particular, this lemma shows that whenever α corresponds to a Schur polynomial,

$$\alpha(z) = z^n + \alpha_1 z^{n-1} + \cdots + \alpha_n, \tag{5.3}$$

the stable manifold of the fast filtering algorithm at $(\alpha, 0)$ is n-dimensional. The next result significantly refines this observation. In particular, we will characterize the stable manifold explicitly. As a preliminary, we denote by V_λ the generalized (complex) eigenspace of $G(\alpha)$ corresponding to an eigenvalue λ of $G(\alpha)$; i.e., a root of (5.3), and we define

$$\begin{aligned} s(\alpha) &= \dim_{\mathbf{C}} \textstyle\sum_{|\lambda|<1} V_\lambda, \\ u(\alpha) &= \dim_{\mathbf{C}} \textstyle\sum_{|\lambda|>1} V_\lambda, \\ c(\alpha) &= \dim_{\mathbf{C}} \textstyle\sum_{|\lambda|=1} V_\lambda. \end{aligned}$$

In particular, $(\alpha, 0)$ is hyperbolic if and only if $s(\alpha) + u(\alpha) = n$ or, equivalently, $c(\alpha) = 0$.

THEOREM 5.2. *The dimensions of the stable manifold and unstable manifold at* $(\alpha_\infty, 0)$ *are* $s(\alpha_\infty)$ *and* $u(\alpha_\infty)$, *respectively. The dimension of a center manifold is always* $n + c(\alpha_\infty)$. *In fact, any center manifold contains an open neighborhood of* $(\alpha_\infty, 0)$ *in the* n*-dimensional equilibrium manifold*

$$E = \{(\alpha, 0) : \alpha \in \mathbb{R}^n\}.$$

Moreover, if $c(\alpha_\infty) = 0$, *then the center manifold is unique and locally coincides with* E. *In this case, the equilibrium* $(\alpha_\infty, 0)$ *is Lyapunov stable if and only if* $u(\alpha_\infty) = 0$, *in which case the stable manifold is* n*-dimensional.*

Proof. Since any center manifold M must contain all local attractors in some neighborhood U of $(\alpha_\infty, 0)$, $M \cap U \supset E \cap U$. If $c(\alpha_\infty) = 0$, then by a dimension argument $M \cap U = E \cap U$ and hence $M \cap U$ is unique. In this case, by the center manifold theorem, the overall system will be Lyapunov stable when $u(\alpha_\infty) = 0$ and trajectories initialized at points (α, γ) sufficiently close to $E \cap A$, where $A = \{(\alpha, 0) :$ (5.3) is a Schur polynomial$\}$, will approach $(\alpha_\infty, 0)$ determined by (4.13) with a convergence rate

$$|\alpha_t| < C \cdot \mu^t \left\| \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} \right\|,$$

where $\mu = \max_{|\lambda|<1} |\lambda|$, and λ is an eigenvalue of $G(\alpha_\infty)$. □

Finally, suppose $(\alpha_\infty, 0)$ is an equilibrium corresponding to a Schur polynomial (5.3) so that $(\alpha_\infty, 0)$ has an n-dimensional stable manifold, $W^s(\alpha_\infty, 0)$. Let (α, γ) be an initial condition lying on $W^s(\alpha_\infty, 0)$. We have noted that the equality (2.18) will hold on the orbit $\{(\alpha_t, \gamma_t); t = 0, 1, \ldots\}$ and hence must hold for $(\alpha_\infty, 0)$. From this observation we can obtain the n-invariants (4.13) in a simple form, by computing the

right-hand side of (2.18) in the limit as a solution of a spectral factorization problem, namely,

$$2r_\infty \alpha_\infty(z)\alpha_\infty(1/z) = 2D(z, z^{-1}), \tag{5.4}$$

where r_∞ is the limit of r_t as $t \to \infty$.

COROLLARY 5.3 (see [5]). *A necessary condition for an initial condition (α, γ) to generate a convergent trajectory is that the pseudo-polynomial $D(z, z^{-1})$ in (2.18) be sign definite.*

If the invariant set X_D introduced in §4 contains an equilibrium point, then $d_0 \neq 0$ in Lemma 4.4, and we may describe X_D in terms of the functions $\mathbb{R}^{2n} \to \mathbb{R}^n$ as

$$h_i(\alpha, \gamma) = 2\frac{d_i(\alpha, \gamma)}{d_0(\alpha, \gamma)}, \qquad i = 1, 2, \ldots, n, \tag{5.5}$$

where $d_i(\alpha, \gamma)$, $i = 0, 1, 2, \ldots, n$, are as defined in Lemma 4.4.

THEOREM 5.4. *Suppose that (α, γ) generates a convergent trajectory for the dynamical system (4.4), and let $D(z, z^{-1})$ be the corresponding pseudopolynomial (4.10). Then, at each point of the trajectory,*

$$h_i(\alpha(t), \gamma(t)) = \kappa_i, \qquad i = 1, 2, \ldots, n, \tag{5.6}$$

where $\kappa_1, \kappa_2, \ldots, \kappa_n$ are constants which can be determined from the initial condition (α, γ). In fact, if $\alpha_n \neq 0$, then $\kappa_n \neq 0$ and

$$d(z) = \alpha_n \left[z^n + \frac{\kappa_{n-1}}{\kappa_n} z^{n-1} + \cdots + \frac{1}{\kappa_n} \right], \tag{5.7}$$

and, if $\alpha_n = \cdots = \alpha_{k+1} = 0$ but $\alpha_k \neq 0$, then $\kappa_n = \cdots = \kappa_{k+1} = 0$, $\kappa_k \neq 0$ and

$$d(z) = r_{n-k}\alpha_k \left[z^k + \frac{\kappa_{k-1}}{\kappa_k} z^{k-1} + \cdots + \frac{1}{\kappa_k} \right]. \tag{5.8}$$

Conversely, any point (α, γ) such that

$$h_i(\alpha, \gamma) = \kappa_i \qquad i = 1, 2, \ldots, n \tag{5.9}$$

has a (finite or infinite) orbit satisfying (5.6) and the same pseudopolynomial (4.10) modulo multiplication by a nonzero constant.

Proof. According to Lemma 5.1, the equilibrium has the form $(\alpha_\infty, 0)$, and, since there is no finite escape, $r_t \neq 0$ for all $t \in \mathbb{Z}$. Consequently, in view of (A-16) and (4.11),

$$d_0(\alpha(t), \gamma(t)) = \frac{r_\infty}{r_t} d_0(\alpha_\infty, 0), \tag{5.10}$$

where, by (4.11),

$$d_0(\alpha, 0) = \alpha_n^2 + \alpha_{n-1}^2 + \cdots + \alpha_1^2 + 1$$

is nonzero. Moreover, $r_\infty \neq 0$. In fact, if $r_\infty = 0$, (5.4) implies that $D(z, z^{-1}) \equiv 0$, which contradicts Lemma 4.4. Hence (5.10) is nonzero and the rational functions (5.5) are finite on the whole trajectory. Moreover, for all $t \in \mathbb{Z}$,

$$h_i(\alpha(t), \gamma(t)) = h_i(\alpha(0), \gamma(0)) \qquad i = 0, 1, 2, \ldots, n.$$

Setting $\kappa_i := h_i(\alpha(0), \gamma(0))$, $i = 1, 2, \ldots, n$, we obtain (5.6). Next, note that $d_i = \frac{1}{2} d_0 \kappa_i$ and $d_n = \alpha_n$. Therefore, if $\alpha_n \neq 0$, then $\kappa_n \neq 0$ and $\frac{1}{2} d_0 = \alpha_n / \kappa_n$, and consequently (5.7) follows. If $\alpha_n = \cdots = \alpha_{k+1} = 0$ but $\alpha_k \neq 0$, then, by Lemma 4.4, $d_i = 0$ for $i = k+1, \ldots, n$ and $d_k = r_{n-k} \alpha_k \neq 0$. Hence, $\kappa_n = \cdots = \kappa_{k+1} = 0$, $\kappa_k \neq 0$ and $\frac{1}{2} d_0 = r_{n-k} \alpha_k / \kappa_k$, and therefore (5.8) follows. Consequently, any $(\alpha, \gamma) \in \mathbb{R}^{2n}$ satisfying (5.9) has a pseudopolynomial that differs from $D(z, z^{-1})$ by at most the nonzero constant, α_n or $r_{n-k} \alpha_k$, whichever case applies, and therefore the points on its orbit satisfy (5.6). □

In view of Theorem 5.4 it is reasonable to let the invariant set (5.9) be denoted X_D, with D interpreted as a point in projective space $\mathbb{P}^n$, as explained at the end of §4. We would like to determine at what points (α, γ) the invariant set X_D is an n-fold, i.e., for which (α, γ) the tangent space $T_{(\alpha,\gamma)} X_D$, or, which is the same, the kernel of the Jacobian of f at (α, γ), has dimension n.

To investigate this point, let us return to the pseudopolynomial relation (2.18) defining the integrals (5.6). To this end, let

$$S(a)v = a(z)v(1/z) + a(1/z)v(z) \tag{5.11}$$

define an operator $S(a) : V_n \to \mathcal{D}_n$ from the vector space V_n of polynomials having degree less than or equal to n into the vector space $\mathcal{D}_n$ of symmetric pseudopolynomials of degree at most n. Such an operator can be defined for each polynomial $a(z)$ of degree n. Relation (2.13) defining $D(z, z^{-1})$ in terms of $a(z)$ and $b(z)$ can then be written

$$S(a)b = 2D \tag{5.12}$$

and we may ask under what conditions this linear system may be solved for b in terms of D and a. It is well known that the answer to this question depends on the location of the zeros of $a(z)$. We say that $(\lambda, 1/\lambda)$ is a *pair of reciprocal roots* (of multiplicity μ) of a pseudo-polynomial $D(z, z^{-1})$ provided that both λ and $1/\lambda$ are roots (of multiplicity μ). According to this definition *each* root (of multiplicity μ) at $\lambda = 1$ or $\lambda = -1$ determines a pair, $(1, 1)$ or $(-1, -1)$, of reciprocal roots (of multiplicity μ).

LEMMA 5.5. *Let ρ be the number of reciprocal roots of $a(z)$ counted with multiplicity. Then*

$$\dim \ker S(a) = \rho. \tag{5.13}$$

Proof. This follows easily from the unit circle version of Orlando's formula [14]. Also see [10], noting that the Jury matrix of $a(z)$ is a matrix representation of $S(a)$. □

We may now write the invariance relation (2.18) in the form

$$r_t S(a_t) b_t = 2D. \tag{5.14}$$

The next lemma establishes notation for the subsequent analysis. Denote by H_n the hyperplane in V_n of monic polynomials. We note that for $\beta \in H_n$ the tangent space $T_\beta H_n$ to H_n at β is canonically isomorphic with V_{n-1}.

LEMMA 5.6. *Let (α, γ) be a point in the invariant algebraic set X_D, defined by* (5.9) *in Theorem* 5.4, *with the property that $\gamma_k^2 \neq 1$ for $k = 0, 1, \ldots, n-1$, and let*

(a, b) be given by (3.12). *Then the tangent space $T_{(\alpha,\gamma)}X_D$ of X_D at (α, γ) has the same dimension as the tangent space of*

$$\phi(a, b, r) = 2D \tag{5.15}$$

at $(a, b, 1)$, where $\phi : H_n \times H_n \times \mathbb{R} \to \mathcal{D}_n$ is defined by

$$\phi(a, b, r) = rS(a)b. \tag{5.16}$$

Proof. The lemma follows immediately from the fact that (5.9) is obtained from (5.15) by merely eliminating the variable r, which is nonzero since all $\gamma_k^2 \neq 1$, and changing coordinates under the bijection $\mathcal{F}$ of Corollary 3.4. □

It is not hard to characterize those tangent vectors that are annihilated by the Jacobian of ϕ at $(a, b, 1)$ and hence span the tangent space of (5.15) at $(a, b, 1)$.

LEMMA 5.7. *At any point $(a, b, 1)$, the kernel of the Jacobian of ϕ consists of those tangent vectors $(u, v, q_0) \in V_{n-1} \times V_{n-1} \times \mathbb{R}$ satisfying*

$$S(a)q + S(b)p = 0, \tag{5.17}$$

where

$$p(z) := u(z), \qquad q(z) := q_0 b(z) + v(z). \tag{5.18}$$

In other words, the kernel of the Jacobian of ϕ can be identified with pairs $(p, q) \in V_{n-1} \times V_n$, i.e., those polynomials of the form

$$\begin{aligned} p(z) &= p_1 z^{n-1} + \cdots + p_n, \\ q(z) &= q_0 z^n + q_1 z^{n-1} + \cdots + q_n, \end{aligned} \tag{5.19}$$

which satisfy the "variational equation"

$$a(z)q(1/z) + a(1/z)q(z) + b(z)p(1/z) + b(1/z)p(z) = 0. \tag{5.17}'$$

Proof. Consider the tangent vector $(a + \varepsilon u, b + \varepsilon v, 1 + \varepsilon q_0)$ at the point $(a, b, 1)$ where $u \in V_{n-1}, v \in V_{n-1}$ and $q_0 \in \mathbb{R}$. We compute the directional derivative of ϕ in the direction (u, v, q_0) as the limit of a Newton quotient

$$\frac{1}{\varepsilon}[\phi(a + \varepsilon u, b + \varepsilon v, 1 + \varepsilon q_0) - \phi(a, b)] \tag{5.20}$$

as $\varepsilon \to 0$, yielding (5.17)′. □

LEMMA 5.8. *Suppose α_∞ corresponds to a polynomial $\alpha(z)$, via* (5.3), *which has no pair of reciprocal roots. Then the invariant algebraic set X_D is a smooth submanifold of dimension n at the equilibrium $(\alpha_\infty, 0)$.*

Proof. When $\gamma = 0$ we have $a(z) = b(z) = \alpha(z)$, so that the variational equation reduces to

$$S(a)[p + q] = 0. \tag{5.21}$$

Since $a(z)$ has no pair of reciprocal roots, by Lemma 5.5, $\ker S(a) = 0$ and therefore we must have

$$p(z) = -q(z). \tag{5.22}$$

Note, in particular, that $q_0 = 0$, i.e., the tangent vectors belong to the $2n$-dimensional space with coordinates (a, b) or (α, γ) as in Corollary 3.4. Since (5.22) defines a subspace of tangent vectors having dimension n, by Lemma 5.6, (5.9) locally defines a smooth submanifold in a neighborhood of $(\alpha_\infty, 0)$ by the implicit function theorem. □

THEOREM 5.9. *Let $(\alpha_\infty, 0)$ be an equilibrium. If $c(\alpha_\infty) = u(\alpha_\infty) = 0$, then the stable manifold through $(\alpha_\infty, 0)$ coincides with an open subset of the invariant n-fold X_D determined from (5.9). Moreover, any point (α, γ) on X_D corresponding to a positive real function $v(z) := \frac{1}{2}(b(z)/a(z))$ will lie on this stable manifold and the minimum phase spectral factor of the spectral density $v(z) + v(1/z)$ will be*

$$w(z) = r_\infty^{1/2} \frac{\alpha_\infty(z)}{a(z)}. \tag{5.23}$$

Proof. Since $\alpha_\infty(z)$ is a Schur polynomial, (5.9) locally defines a smooth submanifold at $(\alpha_\infty, 0)$ by Lemma 5.8, with tangent space given by (5.22). We claim that (5.22) also characterizes, in the (a, b)-coordinates, those tangent vectors (p, q) that are vertical in the (α, γ)-coordinates at point $(\alpha_\infty, 0)$. In fact, the map $\mathcal{F}$ of Corollary 3.4 sends the "vertical vector" $(0, \gamma)$ to

$$(a, b) = (\varphi_n(\gamma), \psi_n(\gamma)), \tag{5.24}$$

where here φ_n and ψ_n are the n-vectors of coefficients in the Szegö polynomials $\varphi_n(z)$ and $\psi_n(z)$ as functions of γ. The vertical vectors $(0, \gamma)$ at the point $(\alpha_\infty, 0)$ corresponds to the tangent vectors (p, q) of (5.24), i.e., the vectors of the form

$$\left(\frac{\partial \varphi_n}{\partial \gamma_i}(0), \frac{\partial \psi_n}{\partial \gamma_i}(0) \right), \qquad i = 1, 2, \dots, n.$$

But, according to Corollary 3.4, $\varphi_n(\gamma) = \psi_n(-\gamma)$ so that

$$\frac{\partial \varphi_n}{\partial \gamma_i}(0) = -\frac{\partial \psi_n}{\partial \gamma_i}(0),$$

and hence $p = -q$ as claimed.

Now recall that the vertical vectors at $(\alpha_\infty, 0)$ are precisely the vectors lying in the sum of the generalized eigenspaces for the Jacobian, corresponding to asymptotically stable eigenvalues, i.e., in the tangent space to the stable manifold at $(\alpha_\infty, 0)$. In summary, the invariant set X_D is an n-dimensional smooth submanifold near the equilibrium $(\alpha_\infty, 0)$ that it contains, provided $(\alpha_\infty, 0)$ is an asymptotically stable equilibrium. In this case, the tangent space to this submanifold at the equilibrium coincides with the tangent space to the stable manifold $W^s(\alpha_\infty, 0)$. In particular, an initial condition lying on X_D corresponding to positive real functions (2.1) will converge geometrically to $(\alpha_\infty, 0)$, in harmony with classically known convergence properties of the Kalman filter. By uniqueness of $W^s(\alpha_\infty, 0)$ we see that it coincides, in a neighborhood of $(\alpha_\infty, 0)$, with the invariant set defined by (5.9). Finally, from (2.13) and (5.4), we see that the minimum phase spectral factor $w(z)$ corresponding to $v(z)$ is given by (5.23). □

Remark 1. As a consequence of Theorem 5.9, we can see that the set of rational integral invariants $\{h_1, h_2, \dots, h_n\}$, defined in (5.5), is complete. That is, there is no analytic (or meromorphic) invariant function h for which the differential dh is linearly

independent of the differentials dh_i, $i = 1, 2, \ldots, n$, at some point (α, γ). Indeed, if for some point $(\bar{\alpha}, \bar{\gamma})$

$$dh(\bar{\alpha}, \bar{\gamma}) \notin \text{span}\,\{dh_i(\bar{\alpha}, \bar{\gamma})\}_{i=1}^n,$$

then for all (α, γ) in some open dense set U we must have

$$dh(\alpha, \gamma) \notin \text{span}\,\{dh_i(\alpha, \gamma)\}_{i=1}^n \tag{5.25}$$

by a standard analyticity argument. Now consider the region $\mathcal{P}_n$ of all $(\alpha, \gamma) \in \mathbb{R}^{2n}$ satisfying the positive real conditions (2.13)–(2.15) with $a(z)$ and $b(z)$ given by (4.1). For any initial condition $(\alpha(0), \gamma(0)) \in \mathcal{P}_n$, we must have

$$h(\alpha(t), \gamma(t)) = h(\alpha_\infty, 0)$$

so that, on $\mathcal{P}_n$, h is determined by its restriction $\bar{h}$ to the equilibrium set

$$E_s = \{(\alpha_\infty, 0) \in \mathbb{R}^{2n} | \alpha_\infty(z) \text{ a Schur polynomial}\}.$$

On the other hand, α_∞ can be computed from $(\alpha(0), \gamma(0))$ as a rational function of the h_i. More explicitly, $h_i(\alpha(0), \gamma(0))$ determine, up to a scalar multiple, the pseudopolynomial $D(z, z^{-1})$ via (5.4) and (A-16). From $D(z, z^{-1})$, which is positive on the unit circle, we determine (independently of the scalar multiple) the stable polynomial $\alpha_\infty(z)$, and hence α_∞, via (5.4). Therefore, on all of $\mathcal{P}_n$ we have

$$h(\alpha(0), \gamma(0)) = \bar{h}(\alpha_\infty, 0) = \bar{h}(F(h_1(\alpha(0), \gamma(0)), \ldots, h_n(\alpha(0), \gamma(0)), 0), \tag{5.26}$$

where F is the rational function defined by (5.4), (4.10), and (5.4). In particular, on $\mathcal{P}_n$ we must have

$$dh \in \text{span}\,\{dh_i\},$$

contrary to the assertion that $U \cap \mathcal{P}_n$ be open and dense in $\mathcal{P}_n$.

So far, we have recovered (cf. Corollary 5.3) a necessary condition for an initial condition (α, γ) to generate an asymptotically convergent trajectory under the dynamics of the fast filtering algorithm; namely the pseudopolynomial $D(z, z^{-1})$ determined by (α, γ) must be sign-definite. In the case where $n = 1$, it has been demonstrated in [5] using somewhat specialized, detailed analysis that, apart from initial data that can escape in finite time, this condition is also sufficient for global convergence. In the case where $n = 2$, this has also been shown [6], although for $n \geq 2$ it is possible to have asymptotic convergence to equilibria that have a lower-dimensional stable manifold (see Theorem 5.2) as well as to the unique Lyapunov stable equilibrium, corresponding to a stable factor of the pseudopolynomial $D(z, z^{-1})$, as occurs in classical filtering. The final result in this section enables us to determine at what points (α, γ) the invariant n-fold X_D defined by (5.9) is a smooth manifold. As it turns out, the singular points correspond to certain systems having a lower-dimensional realization, which also are initial data converging to unstable equilibria.

THEOREM 5.10. *Consider the n-fold X_D defined by* (5.9) *with a corresponding pseudopolynomial $D(z, z^{-1})$. A point (α, γ) is a singular point of X_D if and only if $a(z)$ and $b(z)$, defined in terms of (α, γ) by* (3.12), *have a common pair of reciprocal roots.*

To prove this result, we derive a general formula for the dimension of the tangent space $T_{(\alpha,\gamma)}X_D$ from which our assertion will follow by the implicit function theorem.

LEMMA 5.11. $\dim T_{(\alpha,\gamma)}X_D = n + \sigma$, *where* σ *is the number of common pairs of reciprocal roots of the polynomials* $a(z)$ *and* $b(z)$ *given by* (3.12).

Proof of Theorem 5.10. According to Lemma 5.6 and Lemma 5.7, the tangent space $T_{(\alpha,\gamma)}X_D$ to X_D at (α,γ) has the same dimension as the vector space W of all solutions (p,q) to the variational equation (5.17), i.e., W is the subspace

$$W = \{(p,q) \mid S(a)q + S(b)p = 0\} \tag{5.27}$$

of $V_{n-1} \times V_n$, where $a(z)$ and $b(z)$ are given by (3.12). Now consider the map $\text{proj}_1 : W \to V_{n-1}$ defined via

$$\text{proj}_1(p,q) = p. \tag{5.28}$$

In particular,

$$\dim W = \dim \ker(\text{proj}_1) + \dim \text{range}(\text{proj}_1), \tag{5.29}$$

where

$$\ker(\text{proj}_1) = \{(0,q) \mid S(a)q = 0\} \simeq \ker S(a) \tag{5.30}$$

and

$$\text{range}(\text{proj}_1) = \{p \in V_{n-1} | S(b)p \in \text{range}\, S(a)\}. \tag{5.31}$$

Now recall from Lemma 5.5 that

$$\dim \ker S(a) = \rho_1, \quad \dim \ker S(b) = \rho_2, \tag{5.32}$$

where ρ_1 and ρ_2 are the number of pairs of reciprocal roots $(\lambda, 1/\lambda)$ of $a(z)$ and $b(z)$, respectively, counted with multiplicity. Matters being so, we can also characterize the range of $S(a)$ (and that of $S(b)$) in the vector space of symmetric pseudopolynomials $\mathcal{D}_n$.

Explicitly, if $(\lambda_1, 1/\lambda_1), \dots, (\lambda_{\rho_1}, 1/\lambda_{\rho_1})$ are the ρ_1 pairs of reciprocal roots (counted with multiplicity) of $a(z)$, then the range of $S(a)$ is the codimension ρ_1 subspace of symmetric pseudopolynomials that vanish at $\{\lambda_1, \dots, \lambda_{\rho_1}, 1/\lambda_1, \dots, 1/\lambda_{\rho_1}\}$. The notation

$$\dim \ker(\text{proj}_1) = \rho_1 \tag{5.33}$$

fixes the first term of (5.29). To determine the second, observe that the map $S(b)$ sends any $p \in \text{range}\, \text{proj}_1$ into the subspace

$$U := \text{range}\, S(a) \cap \text{range}\, S(b), \tag{5.34}$$

which consists of all pseudopolynomials in $\mathcal{D}_n$ with $\rho_1 + \rho_2 - \sigma$ pairs of reciprocal zeros fixed, where σ is the number of common such pairs of $a(z)$ and $b(z)$. The space U has codimension $\rho_1 + \rho_2 - \sigma$ in V_n, i.e.,

$$\dim U = n + 1 - \rho_1 - \rho_2 + \sigma. \tag{5.35}$$

Therefore, since $\dim \ker S(b) = \rho_2$, the dimensions of the subspace

$$Z := \{p \in V_n | S(b)p \in \text{range}\, S(a)\} \tag{5.36}$$

is $n+1-\rho_1+\sigma$, and, consequently,

$$\dim \text{range proj}_1 = n-\rho_1+\sigma, \tag{5.37}$$

i.e., one less than $\dim Z$, provided that there is a $p \in Z$ that does not belong to V_{n-1}. However, this is the case because

$$S(b)a = S(a)b \in \mathcal{D}_n, \tag{5.38}$$

i.e., $a(z) \in Z$. Combining (5.29), (5.33), and (5.37) we then see that $\dim W = n+\sigma$. □

6. Fast filtering algorithm, Riccati equations, and Lagrangian Grassmannians. Our goal is to prove a global convergence theorem for the fast filtering algorithm (4.4), or, equivalently, (2.16). As is to be expected, the convergence of (4.4) is intimately connected to the convergence of the matrix Riccati equation (2.2). It is no restriction to assume that $\alpha_n \neq 0$. In fact, as noted above, if $\alpha_n = 0$ we can reduce the dimension of the dynamical system (4.4), replacing n by $k < n$, so that $\alpha_k \neq 0$.

LEMMA 6.1. *Let $\alpha_n \neq 0$. Then, the fast filtering algorithm* (4.4) *tends to a limit* $(\alpha_\infty, 0)$ *if and only if the Riccati equation* (2.2) *with initial condition* $P_0 = 0$ *converges to some equilibrium* P_∞. *Here* P_∞ *satisfies the algebraic Riccati equation*

$$\Lambda(P) := FPF' - P + (g - FPh)(1 - h'Ph)^{-1}(g - FPh)' = 0, \tag{6.1}$$

where the parameters (F, g) *are those corresponding to the initial condition* (α, γ) *of* (4.4), *and*

$$\alpha_\infty = (1 - h'P_\infty h)^{-1}(g - FP_\infty h) + a. \tag{6.2}$$

Proof. The Riccati equation (2.2) can be written as

$$P_{t+1} - P_t = \Lambda(P_t). \tag{6.3}$$

As shown in [25] and pointed out in [5], the structure of the fast filtering algorithm is reflected in the fact that initial condition $P_0 = 0$ renders $\Lambda(P_0) = gg'$ nonnegative definite and rank one, a property that is preserved along the trajectory so that

$$\Lambda(P_t) = r_t g(t) g(t)', \tag{6.4}$$

where

$$r_t = 1 - h'P_t h. \tag{6.5}$$

If the fast filtering algorithm (4.4) converges, then by Lemma 5.1, $(\alpha(t), \gamma(t)) \to (\alpha_\infty, 0)$ for some $\alpha_\infty \in \mathbb{R}^n$, and r_t tends to a limit r_∞ as $t \to \infty$. Hence, according to Corollary 3.4, $a(t) \to \alpha_\infty$ and $b(t) \to \alpha_\infty$ and consequently $g(t) := \frac{1}{2}[b(t) - a(t)] \to 0$. In view of (6.3) and (6.4), this implies that P_t tends to a limit P_∞ as $t \to \infty$. Conversely, suppose that $P_t \to P_\infty$ as $t \to \infty$. Then $\Lambda(P_t) \to 0$, and, by (6.5), $r_t \to r_\infty$. The condition $\alpha_n = a_n + g_n \neq 0$ implies that $r_\infty \neq 0$. In fact, if $r_\infty = 0$, i.e., $h'P_\infty h = 1$, convergence would require that $g = FP_\infty h$ and consequently that $g_n = -a_n h'P_\infty h = -a_n$ (see (2.10)), contradicting the assumption that $\alpha_n \neq 0$. Therefore, by (6.4), $g(t)g(t)'$ and hence g(t) tends to zero, which, in turn, implies that $\gamma_t = g_1(t) \to 0$, i.e., $\gamma(t) \to 0$. Then it follows from (4.4a) that $\alpha(t)$ tends to

a limit α_∞ as $t \to \infty$. Finally, we see from (2.17) and (4.1) that the Kalman gain tends to $k_\infty := \alpha_\infty - a$ as $t \to \infty$. On the other hand, it follows from (2.8) that $k_\infty = r_\infty^{-1}(g - FP_\infty h)$ and hence (6.2) holds. □

LEMMA 6.2. *The statement of Lemma* 6.1 *remains true if* $F := J - ah'$, *defined by* (2.10), *is replaced by* $F := J - bh'$. *If* $\alpha_n \neq 0$, *then at least one of the matrices* $(J - ah')$ *and* $(J - bh')$ *is nonsingular.*

Proof. Exchanging the roles of a and b amounts to changing the initial condition (α, γ) of the dynamical system (4.4) for $(\alpha, -\gamma)$. If (α, γ) has the orbit $\{(\alpha(t), \gamma(t))\}$, then a simple inspection of (4.4) shows that $(\alpha, -\gamma)$ has the orbit $\{(\alpha(t), -\gamma(t))\}$ so that (α, γ) converges if and only if $(\alpha, -\gamma)$ does, both tending to the same limit $(\alpha_\infty, 0)$. This proves the first part of the lemma. To prove the second part, suppose that both $(J - ah')$ and $(J - bh')$ are singular, i.e., $a_n = b_n = 0$. Then $\alpha_n = \frac{1}{2}(a_n + b_n) = 0$, contradicting the assumption that $\alpha_n \neq 0$. □

LEMMA 6.3. *Suppose* F *is nonsingular. Then* $h'F^{-1}g = 1$ *if and only if* $\alpha_n = 0$.

Proof. Let $v(z)$ be as in (2.1). Then

$$v(0) = \tfrac{1}{2} - h'F^{-1}g.$$

On the other hand, since $\gamma_t = -\varphi_{t+1}(0) = \psi_{t+1}(0)$ for $t = 0, 1, 2, \ldots$, and $\varphi_0 = \psi_0 = 1$, (3.9) yields

$$v(0) = -\frac{1}{2}\frac{\gamma_{n-1} + \alpha_1\gamma_{n-2} + \cdots + \alpha_{n-1}\gamma_0 - \alpha_n}{\gamma_{n-1} + \alpha_1\gamma_{n-2} + \cdots + \alpha_{n-1}\gamma_0 + \alpha_n},$$

which equals $-\frac{1}{2}$ if and only if $\alpha_n = 0$. □

In considering the algebraic Riccati equation (6.1) it is important to remember that the situation here is more general than that usually considered in Kalman filtering (where $v(z)$ is positive real) since F may be unstable and $r_\infty := 1 - h'P_\infty h$ may be negative. Here the symmetric matrix P_∞ may have both negative and zero eigenvalues.

Recall now that there is an extensive literature, see, e.g., [2], [29], [30], [35], on the solution of a matrix Riccati equation as a power iteration on the *Lagrangian Grassmannian* manifold, $LG(n, 2n)$ consisting of n-dimensional subspaces $\mathcal{U} \subset \mathbb{R}^{2n}$ which are *Lagrangian* in the sense that

$$x' \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} y = 0 \quad \text{for all } x, y \in \mathcal{U}.$$

In regard to (2.2) this amounts to noting first the well-known fact that the dynamics of the matrix Riccati equation can be described via a linear fractional transformation. Note that, in view of Lemma 6.2 and Lemma 6.3, it is no restriction to assume that F is nonsingular and that the parameter σ, defined in Proposition 6.4, is nonzero in analyzing the convergence of (4.4).

PROPOSITION 6.4. *The matrix Riccati recursion* (2.2) *may be reformulated as*

$$P_{t+1} = (S_{21} + S_{22}P_t)(S_{11} + S_{12}P_t)^{-1}, \tag{6.6}$$

where the $2n \times 2n$ *matrix*

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \tag{6.7}$$

is the symplectic matrix

$$S = \begin{bmatrix} F^{-1} + F^{-1}gh'F^{-1}\sigma^{-1} & F^{-1}gg'\sigma^{-1} \\ -hh'F^{-1}\sigma^{-1} & F' - hg'\sigma^{-1} \end{bmatrix}' \tag{6.8a}$$

with

$$\sigma = 1 - h'F^{-1}g. \tag{6.8b}$$

Proof. A straightforward calculation shows that $\Lambda(P)$, defined by (6.1), may be written

$$\Lambda(P) := APA' - P + APh(1 - h'Ph)^{-1}h'PA' + gg', \tag{6.9}$$

where $A := F - gh'$. Since F is invertible, so is A. In fact,

$$A^{-1} = F^{-1} + F^{-1}gh'F^{-1}\sigma^{-1}. \tag{6.10}$$

Consequently, (6.3) and the fact that

$$(I - hh'P)^{-1} = I + (1 - h'Ph)^{-1}hh'P \tag{6.11}$$

implies that

$$\begin{aligned} P_{t+1} &= gg' + AP_t(I - hh'P_t)^{-1}A' \\ &= [gg'(A')^{-1}(I - hh'P_t) + AP_t](I - hh'P_t)^{-1}A', \end{aligned}$$

which yields (6.6) with $S_{11} = (A')^{-1}$, $S_{12} = -(A')^{-1}hh'$, $S_{21} = gg'(A')^{-1}$, and $S_{22} = A + (1 - \sigma^{-1})gh'$. Inserting (6.10) then yields (6.8). A simple calculation shows that S is symplectic, i.e., that

$$S'\hat{J}S = \hat{J}, \tag{6.12a}$$

where

$$\hat{J} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}. \qquad \square \tag{6.12b}$$

COROLLARY 6.5. *The algebraic Riccati equation* (6.1) *may be written in the alternative form*

$$P = APA' + APh(1 - h'Ph)^{-1}h'PA' + gg', \tag{6.13}$$

where $A := F - gh'$ is invertible, and, in terms of A, the symplectic matrix S takes the form

$$S = \begin{bmatrix} (A')^{-1} & -(A')^{-1}hh' \\ gg'(A')^{-1} & A + (1 - \sigma^{-1})gh' \end{bmatrix}. \tag{6.14}$$

Next, setting

$$P_t = Y_tX_t^{-1} \tag{6.15}$$

and applying Proposition 6.4, we see that the matrix Riccati equation may be viewed as a linear symplectic system

$$Z_{t+1} = SZ_t \tag{6.16}$$

where

$$Z_t = \begin{bmatrix} X_t \\ Y_t \end{bmatrix} \quad \text{and} \quad Z_0 = \begin{bmatrix} I \\ P_0 \end{bmatrix}.$$

In particular, Lemma 6.1 states that the dynamics of the fast filtering algorithm correspond to the initial condition $P_0 = 0$, i.e.,

$$Z_0 = \begin{bmatrix} I \\ 0 \end{bmatrix}. \tag{6.17}$$

Studying the linear system (6.16) on the manifold $LG(n, 2n)$ of Lagrangian subspaces in $\mathbb{R}^{2n}$ instead of (2.2) or (4.4) amounts to a compactification of the phase space in the sense that P_t is also allowed to take infinite values, corresponding to X_t being singular. In particular, this compactification provides a model in which we can analyze high-gain limits, as well as finite escape, of the sequence of Kalman gains. The fact that P_t is symmetric insures that the subspace spanned by the columns of $\left[\begin{smallmatrix} X_t \\ Y_t \end{smallmatrix}\right]$ is Lagrangian.

In view of this, the dynamical behavior of the Riccati equation (2.2), as well as the fast algorithm (2.16) or (4.4), depends on the eigenvalue structure of S, which is connected to the zero structure of the pseudo-polynomial $D(z, z^{-1})$ through the following proposition.

PROPOSITION 6.6. *Let $\alpha_n \neq 0$. Then the eigenvalues of S are identical to the zeros of the pseudopolynomial $D(z, z^{-1})$.*

Proof. Since $\alpha_n \neq 0$, we have $\kappa_n \neq 0$. By a straightforward computation, we see that the characteristic polynomial of S is

$$\begin{aligned} \chi_S(z) = z^{2n} + \tfrac{\kappa_{n-1}}{\kappa_n} z^{2n-1} + \tfrac{\kappa_{n-2}}{\kappa_n} z^{2n-2} + \cdots + \tfrac{\kappa_1}{\kappa_n} z^{n+1} + \tfrac{2}{\kappa_n} z^n \\ + \tfrac{\kappa_1}{\kappa_n} z^{n-1} + \cdots + \tfrac{\kappa_{n-2}}{\kappa_n} z^2 + \tfrac{\kappa_{n-1}}{\kappa_n} z + 1, \end{aligned}$$

where $\kappa_1, \kappa_2, \ldots, \kappa_n$ are integral constants defined in Theorem 5.4. Comparing this with (4.10) and (5.7) we see that

$$D(z, z^{-1}) = \alpha_n z^{-n} \chi_S(z), \tag{6.18}$$

from which the proposition follows. □

This proposition allows us to analyze the dynamics of (2.2) and (4.4) not only for parameters (or, in the case of (4.4), initial conditions) that satisfy the positive real condition, but for general choices of parameters (initial conditions). If $\alpha_n = 0$, Proposition 6.6 should be applied to the dimension-reduced problem mentioned above. Hence, in this case the Riccati equation (2.2) can be replaced by one of smaller dimension, which is actually due to the occurrence of invariant directions [3], as was pointed out in [27] and further developed in [31].

The basic question now is to determine under what conditions (6.16) converges, i.e., under what conditions $S^t \mathcal{Z}_0$ tends to a limit as $t \to \infty$, where $\mathcal{Z}_0$ is the subspace spanned by the columns of Z_0, i.e., $\mathcal{Z}_0 = \operatorname{Im} \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$. Let us first study the set of equilibria of the power iteration $S^t \mathcal{Z}_0$, which must clearly consist of those n-dimensional

subspaces

$$\mathcal{U} = \operatorname{Im} \begin{bmatrix} X \\ Y \end{bmatrix}, \qquad X, Y \; n \times n, \tag{6.19}$$

which are S-invariant. In order that $\mathcal{U}$ should correspond to a (finite) solution of algebraic Riccati equation (6.1), as required by Lemma 6.1, $\mathcal{U}$ must be such that X is nonsingular so that

$$P = YX^{-1} \tag{6.20}$$

can be formed, and, for P to be the limit of the sequence $\{P_t\}$, $\mathcal{U}$ must be Lagrangian so that P is symmetric. The following is a consequence of (h, F) being observable.

LEMMA 6.7. *Let $\alpha_n \neq 0$ and let $\mathcal{U}$, defined by* (6.19), *be Lagrangian. Then X is nonsingular.*

For the proof, we need a result that is a discrete-time version of a result due the Kučera [23]; see [34, p. 379]. Since it is surprisingly more complicated than the continuous-time result, and we need it again below, we state it as a lemma, the proof of which is deferred to the Appendix.

LEMMA 6.8. *Let $\alpha_n \neq 0$ and let $\mathcal{U}$ be an n-dimensional S-invariant Lagrangian subspace. Then, the subspace*

$$\mathcal{W} := \mathcal{U} \cap \operatorname{Im} \begin{bmatrix} I \\ 0 \end{bmatrix}$$

satisfies the invariance condition

(i) $S\mathcal{W} \subset \mathcal{W}$

and, which is equivalent, the reversed invariance condition

(ii) $S^{-1}\mathcal{W} \subset \mathcal{W}$.

The same statements hold for

$$\tilde{\mathcal{W}} := \mathcal{U} \cap \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

Now the proof of Lemma 6.7 follows along the lines of the proof of Shayman's Proposition 1 in [34].

Proof of Lemma 6.7. Suppose $\tilde{\mathcal{W}}$, defined in Lemma 6.8, has dimension k, and that X is singular so that $k > 0$. Then

$$\tilde{\mathcal{W}} = \operatorname{Im} \begin{bmatrix} 0 \\ V \end{bmatrix}$$

for some $n \times k$ matrix V. Since $S\tilde{\mathcal{W}} \subset \tilde{\mathcal{W}}$ (Lemma 6.8) and S is nonsingular (Proposition 6.6), there is a nonsingular $k \times k$ matrix T such that

$$S \begin{bmatrix} 0 \\ V \end{bmatrix} = \begin{bmatrix} 0 \\ V \end{bmatrix} T,$$

i.e., $-(A')^{-1}hh'V = 0$ and $AV + (1 - \sigma^{-1})gh'V = VT$. The first of these equations yields

$$h'V = 0, \tag{6.21a}$$

whereupon the second becomes

$$AV = VT. \tag{6.21b}$$

However, since (h, F) is observable, so is (h, A), for $A = F - gh'$. Therefore (6.21) implies that $V = 0$, contradicting the assumption that X is singular. □

A partial answer to the question of whether the power iteration $S^t\mathcal{Z}_0$ converges can now be given by the following lemma, which generalizes some results due to Parlett and Poole [30]. This requires a few definitions. For any linear operator $A : \mathbb{R}^m \to \mathbb{R}^m$, an A-invariant subspace $\mathcal{U}$ is *dominant* (*codominant*) if the eigenvalues of the restriction $A|\mathcal{U}$ have moduli greater than or equal to (less than or equal to) those of all other eigenvalues of A.

LEMMA 6.9. *Let $A : \mathbb{R}^m \to \mathbb{R}^m$ be a linear operator. If there is a unique p-dimensional dominant A-invariant subspace $\mathcal{U}^-$ and a unique $(m-p)$-dimensional codominant A-invariant subspace $\mathcal{U}^+$, then $A^t\mathcal{X} \to \mathcal{U}^-$ as $t \to \infty$ for each p-dimensional subspace $\mathcal{X}$ such that $\mathcal{X} \cap \mathcal{U}^+ = 0$.*

Proof. If the eigenvalues of A (counted with multiplicity) satisfy

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_m|,$$

then the statement of the lemma follows directly from Theorem 4 in [30]. On the other hand, if

$$|\lambda_1| \geq \cdots \geq |\lambda_{p-q}| > |\lambda_{p-q+1}| = \cdots = |\lambda_{p+r}| > |\lambda_{p+r+1}| \geq \cdots \geq |\lambda_m|,$$

(or there is no eigenvalue larger (smaller) in modulus than λ_p, in which case we set $q = p$ $(r = m - p)$), we define $\hat{\mathcal{U}}^-$ and $\hat{\mathcal{U}}^+$ to be the subspaces spanned by the generalized eigenvectors corresponding to $\{\lambda_1, \ldots, \lambda_{p-q}\}$ and $\{\lambda_{p-q+1}, \ldots, \lambda_m\}$, respectively. Moreover, let $\tilde{\mathcal{U}}^-$ and $\tilde{\mathcal{U}}^+$ be the subspaces spanned by the generalized eigenvectors in $\mathcal{U}^-$, respectively, $\mathcal{U}^+$ corresponding to eigenvalues of modulus $|\lambda_p|$. Then $\tilde{\mathcal{U}}^-$ and $\tilde{\mathcal{U}}^+$ are A-invariant subspaces of $\mathcal{U}^-$ and $\mathcal{U}^+$ of dimensions q and r, respectively. In fact, $\dim(\tilde{\mathcal{U}}^- \cap \tilde{\mathcal{U}}^+) = \min(q, r)$. Now, since $\mathcal{X} \cap \mathcal{U}^+ = 0$, $\dim \mathcal{X} = p$ and $\dim \mathcal{U}^+ = m - p$, $\mathbb{R}^m = \mathcal{X} \oplus \mathcal{U}^+$, where $\oplus$ denotes direct sum. Therefore, since $\mathcal{U}^+ \subset \hat{\mathcal{U}}^+$, there is a subspace $\tilde{\mathcal{X}} \subset \mathcal{X}$ of dimension q such that $\hat{\mathcal{U}}^+ = \tilde{\mathcal{X}} \oplus \mathcal{U}^+$. Let $\hat{\mathcal{X}}$ be any $(p-q)$-dimensional subspace of $\mathcal{X}$ such that $\mathbb{R}^m = \hat{\mathcal{X}} \oplus \hat{\mathcal{U}}^+$. Now, since $\hat{\mathcal{U}}^-$ is the unique dominant $(p-q)$-dimensional A-invariant subspace, and $\hat{\mathcal{U}}^+$ is an invariant complement that, by construction, satisfies $\hat{\mathcal{X}} \cap \hat{\mathcal{U}}^+ = 0$, $A^t\hat{\mathcal{X}} \to \hat{\mathcal{U}}^-$ as $t \to \infty$ by [30, Thm. 4]. Moreover, $\mathcal{X} \cap \mathcal{U}^+ = 0$ implies that $\tilde{\mathcal{X}} \cap \tilde{\mathcal{U}}^+ = 0$. Then, following the argument in the proof of [30, Thm. 7], we see that $A^t\tilde{\mathcal{X}}$ becomes disjoint from the subspace corresponding to the eigenvalues $\{\lambda_{p+r+1}, \ldots, \lambda_m\}$ as $t \to \infty$ as these are smaller in modulus than $|\lambda_p|$. Therefore, since $\tilde{\mathcal{X}} \subset \hat{\mathcal{U}}^+$, the question of convergence of $A^t\tilde{\mathcal{X}}$ is reduced to that of [30, Thm. 6] dealing with the equimodular case. Hence, because $\tilde{\mathcal{U}}^-$ $(\tilde{\mathcal{U}}^+)$ is the unique q-dimensional (r-dimensional) dominant A-invariant subspace of $\hat{\mathcal{U}}^+$ and $\tilde{\mathcal{X}} \cap \tilde{\mathcal{U}}^+ = 0$, $A^t\tilde{\mathcal{X}} \to \tilde{\mathcal{U}}^-$ as $t \to \infty$. Consequently, since $\hat{\mathcal{X}} \oplus \tilde{\mathcal{X}} = \mathcal{X}$ and $\hat{\mathcal{U}}^- \oplus \tilde{\mathcal{U}}^- = \mathcal{U}^-$, $A^t\mathcal{X} \to \mathcal{U}^-$ as $t \to \infty$, as claimed. □

The following lemma shows that the basic assumptions of Lemma 6.9 are fulfilled for the power iteration $S^t\mathcal{Z}_0$, provided $D(z, z^{-1})$ is sign definite, i.e., has no zeros of odd multiplicity on the unit circle (Proposition 6.6). First, let us introduce some notation. Following Parlett and Poole [30] let us order the $2n$ generalized eigenvectors of S first by modulus of the associated eigenvalue with the largest first. Generalized

eigenvectors whose eigenvalues have the same modulus are ordered by exponent, where the *exponent* $e(v)$ of a generalized eigenvector v is defined as

$$e(v) = m - 2g + 1, \tag{6.22}$$

where m is the *multiplicity* of v, i.e., the dimension of the smallest invariant subspace containing it, and g is the *grade* of v, i.e., the dimension of the largest cyclic subspace containing v. Thus let

$$v_1, v_2, \ldots, v_{2n} \tag{6.23}$$

be the generalized eigenvectors ordered in this way, and let

$$\lambda_1, \lambda_2 \ldots, \lambda_{2n} \tag{6.24}$$

be the corresponding eigenvalues (which may be repeated). Then, for each $k = 1, 2, \ldots, 2n$, span $\{v_1, v_2, \ldots, v_k\}$ is a dominant S-invariant subspace.

LEMMA 6.10. *If S has no eigenvalues of odd multiplicity on the unit circle, there is a unique dominant n-dimensional S-invariant subspace $\mathcal{U}_D^-$ and a unique codominant n-dimensional subspace $\mathcal{U}_D^+$. Both are Lagrangian. In particular, $\mathcal{U}_D^-$ is spanned by $\{v_1, v_2, \ldots, v_n\}$ in* (6.23).

Proof. With the generalized eigenvectors of S and its corresponding eigenvalues ordered as in (6.23), (6.24), $\mathcal{U}_D^- := \text{span}\{v_1, v_2, \ldots, v_n\}$ is the *unique* dominant S-invariant n-subspace if either

(i) $|\lambda_n| > |\lambda_{n+1}|$

or

(ii) $|\lambda_n| = |\lambda_{n+1}|$ but $e(v_n) > e(v_{n+1})$;

see [30, p. 404]. Now, recall that S is symplectic so that if λ is an eigenvalue then so is $1/\lambda$. Therefore, if S has no eigenvalues on the unit circle, then case (i) holds, so there is a unique dominant S-invariant n-subspace. If there are eigenvalues on the unit circle, there must be an even number, say $2q$, where $q \leq n$, so that $\{v_1, v_2, \ldots, v_n\}$ contains $n - q$ generalized eigenvectors whose eigenvalues have moduli greater than 1 and q whose eigenvalues lie on the unit circle. If we can show that $e(v_n) > e(v_{n+1})$, case (ii) holds and there is a unique dominant S-invariant n-space, namely $\{v_1, v_2, \ldots, v_n\}$. To this end, let $\mu_1, \mu_2, \ldots, \mu_k$ be the eigenvalues of S on the unit circle (now *not* repeated), and let $m_1, m_2, \ldots, m_k$ be their multiplicities. Then, $\sum_{i=1}^{k} m_i = 2q$. For each $i = 1, \ldots, k$ let $v_i^{(j)}$, $j = 1, 2, \ldots, m_i$, be the (generalized) eigenvectors corresponding to μ_i. The exponent of $v_i^{(j)}$ is

$$e(v_i^{(j)}) = m_i - 2j + 1.$$

Since, by assumption, there are no eigenvalues of odd multiplicity on the unit circle, i.e., m_i is even, $e(v_i^{(j)}) \neq 0$. Therefore, $e(v_i^{(j)})$ is positive for $j = 1, 2, \ldots, m_i/2$ and negative for $j = m_i/2 + 1, \ldots, m_i$, and hence $\mathcal{U}_D^-$ is unique. In the same way, it is seen that $\mathcal{U}_D^+ := \text{span}\{v_{n-q+1}, \ldots, v_n, v_{n+q+1}, \ldots, v_{2n}\}$ is the unique codominant S-invariant n-subspace. The proof that $\mathcal{U}_D^-$ and $\mathcal{U}_D^+$ are Lagrangian can be found in the appendix of [7, Lem. 3.1x]. □

Returning to the fast filtering algorithm (4.4), the following lemma establishes the proper interpretation of the convergence of $S^t \mathcal{Z}_0$ to the dominant S-invariant subspace.

LEMMA 6.11. *Let* $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ *be the eigenvalues (counted with multiplicity) corresponding to the dominant* S*-invariant subspace* $\mathcal{U}_D^-$ *of Lemma* 6.10, *and let* $\mathcal{Z}_0$ *be the* n*-dimensional subspace spanned by the columns of* (6.17). *Then, if* $S^t\mathcal{Z}_0 \to \mathcal{U}_D^-$, *either the trajectory of* (4.4) *escapes to infinity in finite time or* $(\alpha(t), \gamma(t)) \to (\alpha_\infty, 0)$, *where the zeros of the corresponding polynomial*

$$\alpha_\infty(z) = z^n + \alpha_{\infty 1} z^{n-1} + \cdots + \alpha_{\infty n}$$

all lie in the closed unit disc. More precisely,

$$\alpha_\infty(z) = \left(z - \frac{1}{\lambda_1}\right)\left(z - \frac{1}{\lambda_2}\right)\cdots\left(z - \frac{1}{\lambda_n}\right). \tag{6.25}$$

Proof. To say that $S^t\mathcal{Z}_0 \to \mathcal{U}_D^-$ is equivalent to saying that

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} \to \begin{bmatrix} X \\ Y \end{bmatrix} = (v_1, v_2, \dots, v_n)T \tag{6.26}$$

for some nonsingular $n \times n$ matrix T, where, as above,

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = S^t \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

Since $\mathcal{U}_D^-$ is Lagrangian, X is nonsingular (Lemma 6.7). Therefore, if X_t is nonsingular for all $t \in \mathbb{Z}$, the solution $P_t = Y_t X_t^{-1}$ of the matrix Riccati equation (2.2) with initial condition $P_0 = 0$ tends to the limit $P = YX^{-1}$, which is thus a real symmetric solution of the algebraic Riccati equation (6.1). Then, by Lemma 6.1, $(\alpha(t), \gamma(t)) \to (\alpha_\infty, 0)$ where

$$\alpha_\infty = (1 - h'Ph)^{-1}(a + g - JPh). \tag{6.27}$$

If, on the other hand, X_t becomes singular in finite time τ, the Riccati trajectory $P_t = Y_t X_t^{-1}$ escapes to infinity at time τ. To analyze the convergent case, first note that T cancels out in forming $P_t = Y_t X_t^{-1}$ and $P = YX^{-1}$ and therefore we may without restriction assume that $T = I$. Hence

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} \Lambda, \tag{6.28}$$

where Λ is the block diagonal matrix formed by the Jordan blocks corresponding to $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. From this it follows that

$$S_{11} + S_{12}P = X\Lambda X^{-1}. \tag{6.29}$$

Now, substituting S_{11} and S_{12} in (6.29) for their values as defined in (6.8), we have

$$(S_{11}' + PS_{12}')^{-1} = F[F + \sigma^{-1}(g - FPh)h']^{-1}F, \tag{6.30}$$

to which we apply the well-known "matrix inversion lemma"

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1} \tag{6.31}$$

to obtain

$$(S_{11}' + PS_{12}')^{-1} = F - (1 - h'Ph)^{-1}(g - FPh)h' = J - \alpha_\infty h'. \tag{6.32}$$

Therefore, setting $\Gamma := J - \alpha_\infty h'$, (6.29) and (6.32) yield

$$(\Gamma')^{-1}X = X\Lambda, \tag{6.33}$$

i.e., $(\Gamma')^{-1}$ has eigenvalues $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$. Then, $\alpha_\infty(z)$ being the characteristic polynomial of Γ must have the form (6.25), as claimed. Since $|\lambda_i| \geq 1$ for $i = 1, 2, \dots, n$, the zeros of $\alpha_\infty(z)$ are all in the closed unit disc. □

Finally, to establish a global convergence theorem for the fast filtering algorithm (4.4) based on Lemma 6.9, it therefore remains to interpret the condition $\mathcal{U}_D^+ \cap \mathcal{Z}_0 = 0$, where $\mathcal{Z}_0$ is the initial space corresponding to $Z_0 = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$, in terms of the parameters (i.e., the initial conditions) of the algorithm.

LEMMA 6.12. *Let $\alpha_n \neq 0$ and let $\mathcal{U}$ be an n-dimensional Lagrangian S-invariant subspace, and set*

$$\mathcal{Z}_0 := \operatorname{Im} \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

Then, if $a(z)$ and $b(z)$ are coprime, $\mathcal{U} \cap \mathcal{Z}_0 = 0$.

Proof. As above, set $\mathcal{W} := \mathcal{U} \cap \mathcal{Z}_0$ and let U be a full-rank matrix such that

$$\mathcal{W} = \operatorname{Im} \begin{bmatrix} U \\ 0 \end{bmatrix}.$$

Then, since $\mathcal{W}$ is S-invariant (Lemma 6.8), there is a square matrix T such that

$$S \begin{bmatrix} U \\ 0 \end{bmatrix} = \begin{bmatrix} U \\ 0 \end{bmatrix} T.$$

Therefore, in view of (6.14), $(A')^{-1}U = UT$ and $gg'(A')^{-1}U = 0$, from which we see that

$$U'A^{-n}[g, Ag, \dots, A^{n-1}g] = 0. \tag{6.34}$$

Consequently, $U = 0$, i.e., $\mathcal{W} = 0$, if and only if (A, g) is reachable. Since $A = F - gh'$, this is equivalent to (F, g) being reachable. However, since

$$\frac{1}{2}\frac{b(z)}{a(z)} = h'(zI - F)^{-1}g + \frac{1}{2}$$

and (h, F) is observable, (F, g) is reachable if and only if $a(z)$ and $b(z)$ are coprime. □

7. Global convergence of the fast filtering algorithm. We are now in a position to formulate the global convergence theorem. To this end, let $\mathcal{D}$ be the subset of all $(\alpha, \gamma) \in \mathbb{R}^{2n}$ such that $D(z, z^{-1})$ is sign definite on the unit circle, i.e., either nonnegative or nonpositive there. Finally, denote by Ω_e the subset of initial conditions $(\alpha, \gamma) \in \mathbb{R}^{2n}$, which generate trajectories that escape in finite time.

THEOREM 7.1. *For initial conditions $(\alpha, \gamma) \in \mathbb{R}^{2n} - \Omega_e$ there is convergence to an equilibrium under the dynamics of the fast filtering algorithm if and only if the corresponding pseudopolynomial $D(z, z^{-1})$ is sign definite. More precisely, the following statements hold:*

(i) *Ω_e and $\overline{\mathcal{D} \cap \Omega_e}$ have Lebesgue measure zero.*

(ii) $(\alpha,\gamma) \in \mathcal{D} - \mathcal{D} \cap \Omega_e$ *is a necessary and sufficient condition for convergence to an equilibrium.*

(iii) *If* $(\alpha,\gamma) \in \mathcal{D} - \mathcal{D} \cap \Omega_e$, *then* $(\alpha_t,\gamma_t) \to (\alpha_\infty, 0)$ *and the corresponding limit polynomial*

$$\alpha_\infty(z) = z^n + \alpha_{\infty 1} z^{n-1} + \cdots + \alpha_{\infty n}$$

satisfies

$$\alpha_\infty(z) = \tilde{\alpha}_\infty(z)\theta(z),$$

where $\tilde{\alpha}_\infty(z)$ *has all its zeros in the closed unit disc and where*

$$\theta(z) = (a, b),$$

i.e., $\theta(z)$ *is the greatest common divisor of* $a(z)$ *and* $b(z)$.

Moreover, $\tilde{\alpha}_\infty(z)$ *is determined up to a nonzero, scalar multiplicative factor* r_∞ *by the spectral factorization problem*

$$\tilde{a}(z)\tilde{b}(1/z) + \tilde{a}(1/z)\tilde{b}(z) = r_\infty \tilde{\alpha}_\infty(z)\tilde{\alpha}_\infty(1/z), \tag{7.1a}$$

where

$$a(z) = \tilde{a}(z)\theta(z), \qquad b(z) = \tilde{b}(z)\theta(z). \tag{7.1b}$$

Theorem 7.1 not only characterizes those initial conditions that generate a convergent trajectory, but also provides for an explicit determination of the equilibrium to which the corresponding trajectory will converge. Conversely, from this explicit recipe we can also determine which initial conditions will generate a trajectory that converges to a given equilibrium.

COROLLARY 7.2. *In the notation of Theorem* 7.1, *suppose* $\alpha_\infty(z) = \tilde{\alpha}_\infty(z)\theta(z)$ *where* $\tilde{\alpha}_\infty(z)$ *is a Schur polynomial and* $\theta(z)$ *has all of its zeros in* $|z| > 1$. *Then, the global stable "manifold"* $W^s(\alpha_\infty, 0)$ *is given by*

$$W^s(\alpha_\infty, 0) = \{(\alpha,\gamma) \notin \Omega_e\colon \text{(7.1) } \textit{holds with } (a,b) \textit{ given by } (4.1) \textit{ and } (\tilde{a},\tilde{b}) = 1\}.$$

Similarly, the global unstable "manifold" can be parameterized as all coprime pairs $(\bar{a},\bar{b})$ *satisfying*

$$a(z) = \bar{a}(z)\tilde{\alpha}_\infty(z), \qquad b(z) = \bar{b}(z)\tilde{\alpha}_\infty(z),$$

and

$$\bar{a}(z)\bar{b}(1/z) + \bar{a}(1/z)\bar{b}(z) = \bar{r}_\infty \theta(z)\theta(1/z).$$

Finally, a global center manifold $W^c(\alpha_\infty, 0)$ *is given by the equilibrium set* E.

Remark 2. The existence of stable and unstable manifolds as locally invariant immersed manifolds is of course a local result. In harmony with this, Lemma 5.8 gives a result characterizing $W^s(\alpha_\infty, 0)$ as a submanifold near $(\alpha_\infty, 0)$. In contrast, the description of $W^s(\alpha_\infty, 0)$ in the large as given in Corollary 7.2 does allow for singular points. These singular points are characterized in Theorem 5.10.

Proof of Theorem 7.1. We first assume that $\alpha_n \neq 0$ so that the pseudopolynomial $D(z, z^{-1})$ has degree n and the symplectic matrix S is well defined and nonsingular.

Finite escape occurs for precisely the initial conditions

$$Z_0 = S^{-t}\tilde{Z}, \qquad t = 0, 1, 2, \dots, \tag{7.2}$$

for which $\tilde{X}$ is singular. In (α, γ)-space, $\tilde{X}$ being singular corresponds to $(\tilde{\alpha}, \tilde{\gamma})$ belonging to the two hyperplanes $\gamma_{n-1} = \pm 1$. Forming the union of the countably many iterates via (7.2) of these hyperplanes forms a set Ω_e that has measure zero. Here $\mathcal{D} \cap \Omega_e$ also has measure zero. We defer the proof that $\overline{\mathcal{D} \cap \Omega_e}$ is a set of measure zero.

Concerning the second part of the theorem, we have already established that it is necessary that $D(z, z^{-1})$ is sign definite for (4.4) to converge (Corollary 5.3). We prove the converse statement by proving assertion (iii). Suppose that $D(z, z^{-1})$ is sign definite, i.e., it has no zeros of odd multiplicity on the unit circle. Then Proposition 6.6 implies that S has no eigenvalues of odd multiplicity on the unit circle. However, this implies that there are unique dominant and codominant S-invariant n-dimensional subspaces $\mathcal{U}_D^-$ and $\mathcal{U}_D^+$, respectively, which are Lagrangian (Lemma 6.10), so that we can apply Lemma 6.9.

First, suppose that $a(z)$ and $b(z)$ are relatively prime. Then, if $\mathcal{Z}_0$ is the subspace spanned by the columns of $Z_0 = \begin{bmatrix} I \\ 0 \end{bmatrix}$, we have $\mathcal{U}_D^+ \cap \mathcal{Z}_0 = 0$ (Lemma 6.12) so that $S^t\mathcal{Z}_0 \to \mathcal{U}_D^-$ as $t \to \infty$ (Lemma 6.9). But this implies that $(\alpha(t), \gamma(t)) \to (\alpha_\infty, 0)$ (Lemma 6.11) where $\alpha_\infty(z)$ has all its zeros in the closed unit disc, unless there is finite escape.

Next, suppose that $a(z) = \tilde{a}(z)\theta(z)$ and $b(z) = \tilde{b}(z)\theta(z)$, where $\theta(z)$ is a nontrivial monic polynomial and $\tilde{a}(z)$ and $\tilde{b}(z)$ are relatively prime. Then the factor $\theta(z)$ can be canceled in $v(z) = \frac{1}{2}(a(z)/b(z))$, so we may consider the dimension-reduced problem with (a, b) exchanged for $(\tilde{a}, \tilde{b})$. Since $D(z, z^{-1}) = \tilde{D}(z, z^{-1})|\theta(z)|^2$ is sign definite on the unit circle, then so is $\tilde{D}(z, z^{-1})$, and consequently, unless there is finite escape,

$$(\tilde{\alpha}(t), \tilde{\gamma}(t)) \to (\tilde{\alpha}_\infty, 0) \tag{7.3}$$

as $t \to \infty$, where $\tilde{\alpha}_\infty(z)$ has all its zeros in the closed unit disc (Lemma 6.11). On the other hand, it was shown in [27], and further elaborated on in [5], that the fast algorithm (2.16), which is equivalent to (4.4), can be written in the Szegö-like polynomial form

$$\begin{aligned} Q_{t+1}(z) &= Q_t(z) - \gamma_t z Q_t^*(z), \\ Q_{t+1}^*(z) &= z Q_t^*(z) - \gamma_t Q_t(z) \end{aligned} \tag{7.4}$$

(see [5, §2]), which, through the transformation

$$\begin{aligned} a_t(z) &= r_t^{-1}[Q_t(z) - Q_t^*(z)], \\ b_t(z) &= r_t^{-1}[Q_t(z) + Q_t^*(z)], \end{aligned} \tag{7.5}$$

and (3.7), provides us with a polynomial version in (a, b)-coordinates of the dynamical system (4.4). From this we see that, if $a_0(z) := a(z)$ and $b_0(z) := b(z)$ have a nontrivial common factor $\theta(z)$, then

$$\begin{aligned} a_t(z) &= \tilde{a}_t(z)\theta(z), \\ b_t(z) &= \tilde{b}_t(z)\theta(z) \end{aligned} \tag{7.6}$$

for *all* $t = 0, 1, 2, 3, \dots$. Since $\tilde{a}_0(z) = \tilde{a}(z)$ and $\tilde{b}_0(z) = \tilde{b}(z)$, it follows readily from (7.4) and (7.5) that $\{(\tilde{a}_t, \tilde{b}_t)\}$ is a trajectory in (a, b)-coordinates of the reduced system

obtained by cancellation of the factor $\theta(z)$. In view of (7.3) and (A-1), $\tilde{a}_t(z) \to \tilde{\alpha}_\infty(z)$ and $\tilde{b}_t(z) \to \tilde{\alpha}_\infty(z)$ as $t \to \infty$, where $\alpha_\infty(z)$ has all its zeros in the closed unit disc. Therefore, $a_t(z) \to \tilde{\alpha}_\infty(z)\theta(z)$ and $b_t(z) \to \tilde{\alpha}_\infty(z)\theta(z)$ so that $(\alpha_t, \gamma_t) \to (\alpha_\infty, 0)$ where

$$\alpha_\infty(z) = \tilde{\alpha}_\infty(z)\theta(z), \tag{7.7}$$

the zeros of which are located in the closed unit disc if and only if $\theta(z)$, the common factor of $a(z)$ and $b(z)$, has all its zeros in the closed unit disc.

To complete the proof of (i), we now demonstrate that

$$\overline{\mathcal{D} \cap \Omega_e} \subset \mathcal{D} \cap \Omega_e \cup \mathcal{F}_1 \cup \mathcal{F}_2,$$

where $\mathcal{F}_1$ and $\mathcal{F}_2$ are sets of measure zero. Indeed, $\mathcal{F}_1$ is the algebraic set consisting of those pairs (α, γ) for which the corresponding polynomials (a, b) have a nontrivial common factor and $\mathcal{F}_2$ is the algebraic set consisting of those pairs for which the corresponding pseudopolynomial $D(z, z^{-1})$ has a double root. Suppose then that $(\alpha^{(n)}, \gamma^{(n)})$ is a sequence in $\mathcal{D} \cap \Omega_e$ with limit

$$\lim_{n\to\infty} (\alpha^{(n)}, \gamma^{(n)}) = (\alpha, \gamma).$$

Of course, $(\alpha, \gamma) \in \mathcal{D}$ so our claim will follow if we show that $(\alpha, \gamma) \in \mathcal{D} - \mathcal{D} \cap \Omega_e$ implies $(\alpha, \gamma) \in \mathcal{F}_1 \cup \mathcal{F}_2$. From (iii) of Theorem 7.1, we know that if $(\alpha_0, \gamma_0) = (\alpha, \gamma)$, then

$$\lim_{t\to\infty} (\alpha_t, \gamma_t) = (\alpha_\infty, 0),$$

where $\alpha_\infty(z) = \tilde{\alpha}_\infty(z)\theta(z)$, where $\tilde{\alpha}_\infty(z)$ has all of its roots in the closed unit disc and $\theta = (a, b)$. If $\deg \theta \geq 1$, then $(\alpha, \gamma) \in \mathcal{F}_1$, so we suppose $\theta(z) \equiv 1$. In this case, to say $\alpha_\infty(z)$ has roots on the unit circle is to say $(\alpha, \gamma) \in \mathcal{F}_2$, so we may assume $\alpha_\infty(z)$ is a Schur polynomial, an assumption that we show is contrary to fact. If $\alpha_\infty(z)$ is a Schur polynomial, then $(\alpha_\infty, 0)$ belongs to the region $\mathcal{P}_n$ of all $(\alpha, \gamma) \in \mathbb{R}^{2n}$ satisfying the positive real conditions (2.13)–(2.15), and so $(\alpha_T, \gamma_T) \in \mathcal{P}_n$ for some finite $T > 0$. Since $\mathcal{P}_n$ is open and since the map on $\mathbb{R}^{2n}$

$$\Phi_T : (\alpha_0, \gamma_0) \mapsto (\alpha_T, \gamma_T),$$

defined by iterating the dynamical system (4.4) T times, is rational with no pole at (α, γ), there exists an $\varepsilon > 0$ such that

$$\Phi_T(B_\varepsilon(\alpha, \gamma)) \subset \mathcal{P}_n.$$

However, then no $(\alpha', \gamma') \in B_\varepsilon(\alpha, \gamma)$ can escape in finite time, contrary to the definition of (α, γ).

Finally, it is easy to modify the above argument to include the case $\alpha_n = 0$. Indeed, if for some $k < n$, $\alpha_n = \cdots = \alpha_{k+1} = 0$ and $\alpha_k \neq 0$, then $D(z, z^{-1})$ has degree k and the dynamical system (4.4) is reduced to a system of order $2k$ in $n - k$ steps. Therefore, $(\alpha(t), \gamma(t)) \to (\alpha_\infty, 0)$ if and only if $(\hat{\alpha}(t), \hat{\gamma}(t)) \to (\hat{\alpha}_\infty, 0)$, where the "hatted" quantities correspond to the reduced system. Then $\alpha_\infty(z) = z^{n-k}\hat{\alpha}_\infty(z)$ so that all statements concerning the reduced system also hold for the unreduced one. □

We remark that the initial conditions (α, γ) for which the pseudopolynomial $D(z, z^{-1})$ fails to be sign definite, and there consequently is no convergence, form an unbounded open set in $\mathbb{R}^{2n}$. As we illustrate in the next section, such points can be periodic or dense on some unbounded submanifold, depending on certain number theoretic considerations and leading to a remarkable sensitivity of the fast filtering algorithm to initial conditions in this region. We will return to this topic in a subsequent paper.

8. Examples and simulations. The purpose of this section is to illustrate our results for low-order problems, particularly the cases $n = 1$ and $n = 2$. Since these cases have been treated in [5] and [6], respectively, we quote only those results that best illustrate our main theorem.

In the first-order case the dynamical system (4.4) takes the form

$$\alpha_{t+1} = \frac{\alpha_t}{1 - \gamma_t^2}, \tag{8.1a}$$

$$\gamma_{t+1} = -\frac{\gamma_t \alpha_t}{1 - \gamma_t^2} \tag{8.1b}$$

corresponding to the rational function

$$v(z) = \frac{1}{2}\frac{b(z)}{a(z)}, \tag{8.2}$$

where

$$\begin{aligned} a(z) &= z + \alpha - \gamma, \\ b(z) &= z + \alpha + \gamma. \end{aligned} \tag{8.3}$$

This case was studied in detail in [5], where it was shown that points (α, γ) in the interior of the diamond I, with corners $(\pm 1, 0)$, $(0, \pm 1)$, depicted in Fig. 1 correspond to positive real $v(z)$, whereas the points (α, γ) in the shaded regions are precisely those for which $D(z, z^{-1})$ is sign definite on the unit circle, $v(1/z)$ being positive real in regions III and negative real in regions II. The dotted lines are the lines $\gamma = \pm 1$ of finite escape.

The invariant manifold X_D, defined by (5.9), becomes

$$1 + \alpha_t^2 - \gamma_t^2 = \frac{2}{\kappa}\alpha_t, \tag{8.4}$$

valid for all $\kappa \neq 0$ (including $\kappa = \infty$, corresponding to $d_0(\alpha, \gamma) = 0$); for $\kappa = 0$, the dynamical system (8.1) evolves along the axis $\alpha = 0$, converging in one step to the origin.

Fig. 2 depicts the invariant manifolds defined by (8.4) for certain values of κ. For $\kappa^2 < 1$ these manifolds are hyperbolas completely contained in the shaded, sign-definite region, and for $\kappa^2 = 1$ they degenerate into a pair of intersecting lines, in the boundary of the shaded region, intersecting in $(1, 0)$ or $(-1, 0)$. In fact, each point in the shaded region lies on such an invariant manifold and converges to the intersection of this hyperbola with the segment $\{(\alpha, 0) \mid -1 \leq \alpha \leq 1\}$.

Since $a(z)$ and $b(z)$, displayed in (8.3), can have a common pair of reciprocal roots only in $(1, 0)$ and $(-1, 0)$, Theorem 5.27 states that these are the only singular points, σ in Lemma 5.11 being zero everywhere else, a fact that is illustrated by the above analysis.

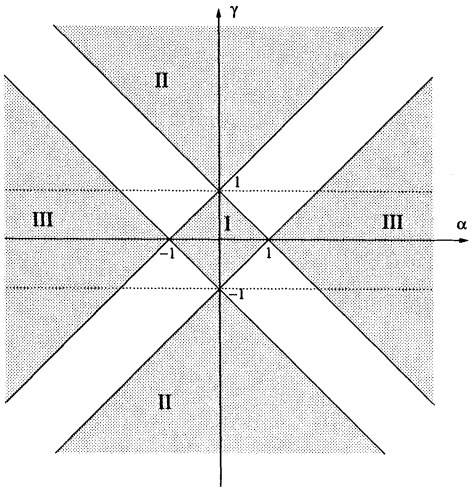


FIG. 1.

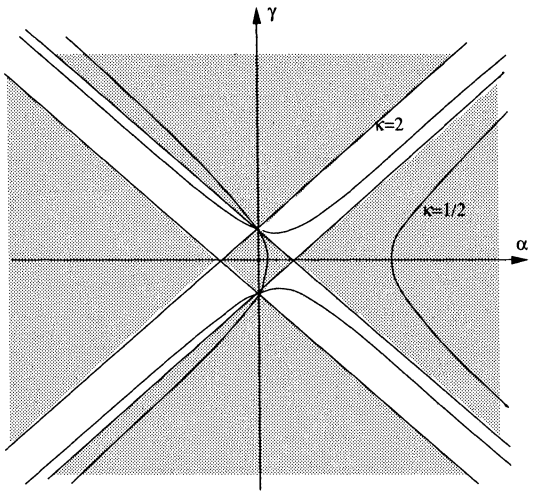


FIG. 2.

It is easy to check (see [5] for details) that hyperbolas for which $\kappa^2 < 1$ correspond to those symplectic matrices S that have two real eigenvalues, one inside the unit circle and the other outside, whereas the case $\kappa^2 = 1$ yields a symplectic matrix S that has an eigenvalue of multiplicity 2 either at 1 or at -1. Convergence in these cases is therefore in accordance with Theorem 7.1, since S has no eigenvalue of odd multiplicity on the unit circle. On the other hand, if $\kappa^2 > 1$, there is a complex pair of such eigenvalues, and the hyperbolas lie in the white region of Fig. 2, where the corresponding pseudopolynomials $D(z, z^{-1})$ are sign indefinite on the unit circle, implying nonconvergence by Theorem 7.1.

Those hyperbolas for which $\kappa^2 < 1$ intersect the α-axis in two points, one of which is a point α_∞ so that the polynomial $z + \alpha_\infty$ is Schur; i.e., so that $|\alpha_\infty| < 1$.

In [5] it has been shown that not only is the hyperbola (8.4) locally a stable manifold for $\kappa^2 < 1$, but rather it consists of a global stable manifold, excluding the unstable equilibrium and the measure zero set of points which escape in finite time. Also, $E = \{(\alpha_\infty, 0) : \alpha_\infty \in \mathbb{R}\}$ is a global center manifold through $(\alpha_\infty, 0)$. We note that the unstable equilibrium $(\alpha_\infty, 0)$, with $|\alpha_\infty| > 1$, has, by (5.2), a one-dimensional center manifold. In fact, these manifolds exist globally with the hyperbola being a global unstable manifold for the unstable equilibrium, on which trajectories either escape or evolve to the equilibrium, with the exception of the unstable equilibrium itself. E is again a global center manifold. The global convergence is completely understood in this case and described in [5].

If $\kappa^2 > 1$, then the hyperbolas do not intersect the α-axis, see Fig. 2, and indeed the dynamics are far more complex. In fact, in [5] it is shown that there are two alternatives that, taken together, prove that (8.1) is sensitive to initial conditions, in the technical sense (as in [11]). Explicitly, one knows that either (A) or (B) holds:

(A) $\arctan\sqrt{\kappa^2 - 1} \in \mathbb{Q}\pi$ and hence

$$\frac{1}{2}\arctan\sqrt{\kappa^2 - 1} = \frac{q}{p}\pi \quad \text{if } \kappa < -1$$

or

$$\frac{1}{2}\{\pi - \arctan\sqrt{\kappa^2 - 1}\} = \frac{q}{p}\pi \quad \text{if } \kappa > 1,$$

where p and q are coprime natural numbers. If p is odd, $2(p-1)$ points on the hyperbola escape in finite time and if p is even there are $(p-2)$ such points. All other points are periodic with period p and every period p, $p \geq 3$, is possible.

(B) $\arctan\sqrt{\kappa^2 - 1} \notin \mathbb{Q}\pi$ and a countably infinite set of points on the hyperbola escape in finite time. All other points generate a dense orbit.

Finally, consider the points $(\pm 1, 0)$, correspondingly to $\kappa^2 = 1$. According to Theorem 5.2, the center manifold is two-dimensional and consequently is global. In fact, hyperbolas of all types, containing periodic orbits and dense orbits or consisting of stable and unstable manifolds, intersect every neighborhood of either equilibrium $(\pm 1, 0)$ yielding a rather complicated mix of dynamics. However, points lying on the degenerate hyperbola for $\kappa = \pm 1$ do converge to the equilibrium $(\pm 1, 0)$, except for a countable set of points which escape in finite time.

In n dimensions, the n-folds (5.9) are defined for every value of $\kappa_1, \ldots, \kappa_n$. Moreover, setting $\kappa_n = \kappa_{n-1} = \cdots = \kappa_2 = 0$, we obtain an invariant subset of $\mathbb{R}^{2n}$ on which the n-dimensional algorithm restricts to the first-order algorithm on the hyperbola defined by $\kappa = \kappa_1$. Therefore, in addition to the equilibrium structure described in §5 and the convergence analysis in §6 yielding a parameterization of the global stable manifolds of these equilibria, we also know (see [5]) the following result.

PROPOSITION 8.1. *For any $p \geq 3$, there exist infinitely many periodic points of period p for the fast filtering algorithm. Arbitrarily close to any one of these initial conditions is an initial condition that generates an unbounded orbit. In particular, in the sign indefinite region (in which trajectories cannot converge to equilibria), the fast filtering algorithms can exhibit sensitivity to initial conditions.*

We refer the reader to [5] for further details of the various kinds of asymptotic behavior in the case where $n = 1$.

In the case where $n = 2$, the fast filtering algorithm (4.4) takes the form

$$\alpha_1(t+1) = \frac{1}{1-\gamma_{t+1}^2}\alpha_1(t) + \frac{\gamma_{t+1}}{1-\gamma_{t+1}^2}\frac{\gamma_t}{1-\gamma_t^2}\alpha_2(t), \tag{8.5a}$$

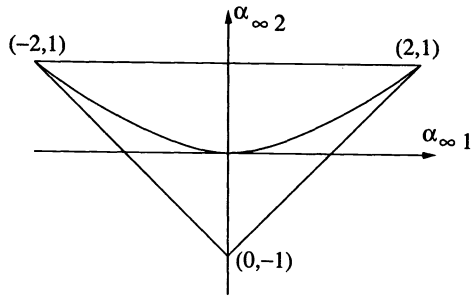


FIG. 3.

$$\alpha_2(t+1) = \frac{\alpha_2(t)}{1-\gamma_t^2}, \tag{8.5b}$$

$$\gamma_{t+1} = -\alpha_1(t)\gamma_t - \alpha_2(t)\gamma_{t-1}, \tag{8.5c}$$

and the invariant manifold X_D becomes

$$\begin{aligned} 2(r_1\alpha_1 + \alpha_1\alpha_2 + \gamma_0\gamma_1\alpha_2) &= \kappa_1(\alpha_2^2 + r_1\alpha_1^2 + r_2), \\ 2\alpha_2 &= \kappa_2(\alpha_2^2 + r_1\alpha_1^2 + r_2), \end{aligned} \tag{8.6}$$

where $r_1 = 1 - \gamma_0^2$ and $r_2 = (1-\gamma_0^2)(1-\gamma_1^2)$ are defined as in (3.7), as long as d_0, given by (4.11), is nonzero. The other cases are covered by dividing one or both of the equations in (8.6) by κ_1 and κ_2, respectively, and allowing these constants to take infinite values.

In the second-order case the invariant manifold X_D may have a singular point not only if $a(z)$ and $b(z)$ have a root $z = 1$ or $z = -1$ in common, which may occur outside of the equilibrium set, but also if

$$a(z) = b(z) = (z+\lambda)(z+1/\lambda),$$

which can only occur in an equilibrium point, because $\gamma_0 = \gamma_1 = 0$ in this case. By Lemma 5.1, equilibria are precisely the points of the form $(\alpha_\infty, 0)$. Inserting $(\alpha_\infty, 0)$ in (8.6) yields the constants κ_1, κ_2 defining the invariant manifold containing this point. In fact,

$$\begin{aligned} \kappa_1 &= \frac{2\alpha_{\infty 1}(1+\alpha_{\infty 2})}{\alpha_{\infty 2}^2 + \alpha_{\infty 1}^2 + 1}, \\ \kappa_2 &= \frac{2\alpha_{\infty 2}}{\alpha_{\infty 2}^2 + \alpha_{\infty 1}^2 + 1}. \end{aligned} \tag{8.7}$$

Conversely, it follows from Theorem 5.9 and Theorem 7.1 that to each point (κ_1, κ_2) such that $D(z, z^{-1})$ is sign definite, there corresponds a unique α_∞, such that $\alpha_\infty(z)$ is stable, i.e., all its zeros lie inside the unit circle. These α_∞ are precisely the points in the closed triangular stability region depicted in Fig. 3

Since the roots of

$$D(z, z^{-1}) = r_\infty \alpha_\infty(z)\alpha_\infty(1/z) \tag{8.8}$$

are the eigenvalues of the symplectic matrix S (Proposition 6.6), each point in the closed triangle depicted in Fig. 3 corresponds to a particular eigenvalue configuration

for which there is convergence. Excluding the segment $\alpha_{\infty 2} = 0$ that corresponds to the case where $n = 1$, the points in the interior of the triangle correspond to the situations when there are no eigenvalues on the unit circle. Below the parabola $\alpha_{\infty 2} = \alpha_{\infty 1}^2/4$, there are four real eigenvalues, while above there are two complex pairs. On the boundary of the triangle there are eigenvalues on the unit circle, but they are always of even multiplicity, as a simple application of (8.8) shows. The rest of the plane, outside of the triangle, corresponds to unstable solutions of the polynomial factorization problem (8.8) and hence to unstable equilibria $(\alpha_\infty, 0)$. To each point below the parabola in the interior of the triangle there corresponds one strictly unstable and two saddle equilibria outside the triangle. For an interior point above the parabola there is only one equilibrium outside the triangle and it is strictly unstable.

For all points in the interior of the triangle the invariant manifold X_D defined by (8.6) is a smooth surface. As shown in §6, not only is the invariant manifold through such a point locally a stable manifold, but it actually constitutes a global stable manifold, excluding the unstable equilibria, their stable manifolds, and the measure zero set of points that escape in finite time. Using the same argument as in the first-order case, $E = \{(\alpha_\infty, 0) : \alpha_\infty \in \mathbb{R}^2\}$ is a global center manifold through $(\alpha_\infty, 0)$ of dimension 2, for any α_∞ that does not lie on the lines through $(-2, 1)$, $(2, 1)$, and $(0, -1)$, i.e., on the boundary of the triangle.

The points on the boundary of the triangle are all singular. In fact, the invariant manifolds corresponding to the points on the line segment between $(-2, 1, 0, 0)$ and $(2, 1, 0, 0)$, as well as that of the point $(0, -1, 0, 0)$, have dimensions less than two. The center manifolds containing these points all have dimension four, while the center manifolds containing the points on the open boundary segments extending from the corner $(0, -1, 0, 0)$ have dimension three.

An initial condition $(\alpha, \gamma) \in \mathbb{R}^4$ for the fast filtering algorithm that does not belong to the plane $(\alpha_\infty, 0)$ of equilibria may or may not converge to an equilibrium. Figure 4 shows the plane $\alpha \mapsto (\alpha, \gamma)$ where γ is fixed so that, in this example, $\gamma_0 = 1/2$ and $\gamma_1 = \frac{1}{3}$. Each point in the bounded shaded region in Fig. 4 corresponds to a positive real function $v(z)$, and hence to a bona fide stochastic system, and converges, by classical results, to a stable equilibrium $(\alpha_\infty, 0)$ in the triangle of Fig. 3. This is precisely the solution set of the rational covariance extension problem for which the covariance data $\{c_1, c_2\}$ is prescribed so that the Schur parameters are $\gamma_0 = \frac{1}{2}$ and $\gamma_1 = \frac{1}{3}$. Initial conditions in the four unbounded shaded regions also correspond to orbits that converge to stable or unstable equilibria $(\alpha_\infty, 0)$ except for a zero measure set which escape in finite time.

As an example we may now choose the point $(0, 2, 1/2, 1/3)$, which lies in the topmost shaded unbounded region. We see from the simulation depicted in Fig. 5 that, using this point as an initial condition, the fast filtering algorithm (8.5) converges after having violated the positive real condition $|\gamma_t| < 1$ twice, showing that the corresponding $v(z)$ is not positive real. However, after six steps the iterate is inside the bounded positive real region and will remain there.

What happens if the initial condition (α, γ) lies in the white region of Fig. 4? These points correspond to sign indefinite $D(z, z^{-1})$ and according to Proposition 8.1 we have at least three kinds of behavior.

(i) (α, γ) is a periodic point;
(ii) the orbit of (α, γ) is dense on some manifold;
(iii) there is finite-time escape.

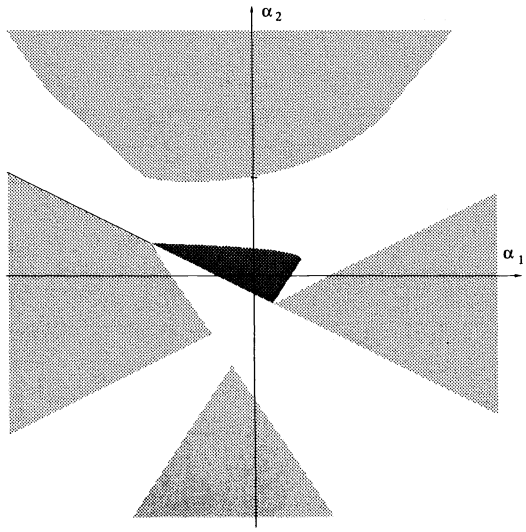


FIG. 4.

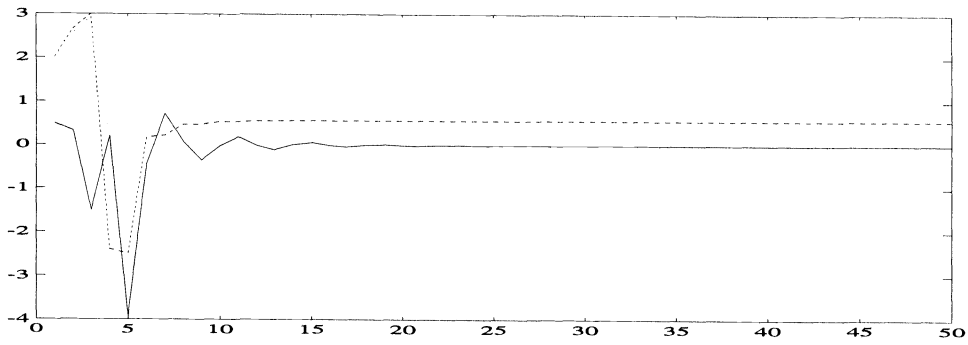


FIG. 5. *Plot of* α_1 (*dotted line*), α_2 (*dashed line*), γ (*solid line*).

These nonconvergent initial conditions and the invariant manifolds on which they lie correspond to the situations when S has eigenvalues of odd multiplicity on the unit circle. In the literature there has been a tendency to exclude the case with eigenvalues on the unit circle as being a rather complicated nongeneric case, but, as our analysis shows, this situation actually corresponds to an open unbounded set of initial conditions. Cases (i) and (iii), however, occur only for a measure zero subset of the white region. We refer the reader to [6] for simulations illustrating these types of dynamical behavior. Here we show only one simulation that illustrates that in the white region the fast filtering algorithm is extremely sensitive to the initial conditions. Consider the periodic point of period 144 corresponding to $\beta_1 = \sec \pi/8$

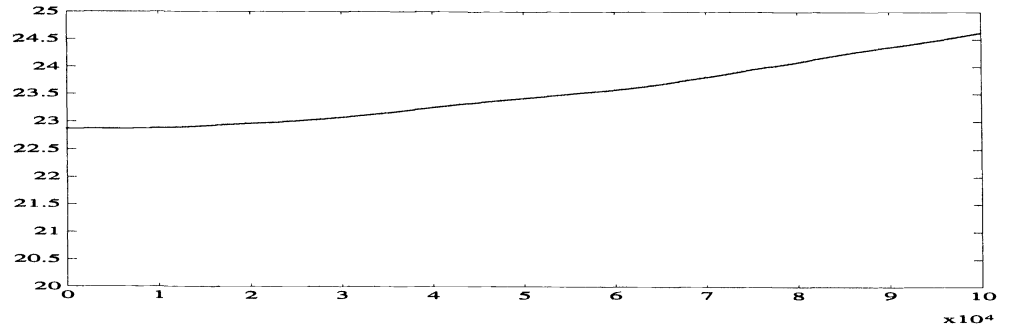


FIG. 6.

and $\beta_2 = \sec \pi/9$, where

$$\frac{2}{\beta_1} = \frac{-\kappa_1 + \sqrt{\kappa_1^2 - 8\kappa_2 + 8\kappa_2^2}}{2\kappa_2},$$
$$\frac{2}{\beta_2} = \frac{-\kappa_1 - \sqrt{\kappa_1^2 - 8\kappa_2 + 8\kappa_2^2}}{2\kappa_2}.$$

If we round off α_0 to the 15th digit, we obtain an orbit that is dense on the invariant manifold and part of whose γ-trajectory is depicted in Fig. 6. This dynamical behavior is apparently quite different to that of the periodic point which the new initial condition approximates.

If S has all its eigenvalues on the unit circle but at least two of them are real, then the dynamics degenerates to the first-order case. The interesting fact here is that $a(z)$ and $b(z)$ having a common pair of reciprocal roots which are not 1 corresponds to an equilibrium lying on the line $(\alpha_{\infty,1}, 1, 0, 0)$, where $|\alpha_{\infty,1}| > 2$, i.e., the part of the line that is not in the boundary of the triangle of Fig. 3. Moreover, this equilibrium is a saddle point.

A complete description of the positive real, sign definite and sign indefinite regions is available in the case $n = 2$, as reported in [19], where many simulation results are also given. Earlier graphical simulations of the positive real region, in the case $n = 2$, are contained in Georgiou's thesis [15]. Curiously, all graphical representations of the positive real region $\mathcal{A}_+(n)$ for γ fixed of which we are aware seem to be convex. Convexity of $\mathcal{A}_+(n)$ would in fact imply a Kharitonov-like property, namely, star-shapedness about the maximum-entropy filter, conjectured and established for the case $n = 1$ by Kimura. In this direction it is known that for reasons concerning the geometry of the spaces of real and of complex Schur polynomials, the convexity $\mathcal{A}_+(2)$ seems to be decidedly nontrivial. In general, although examples show [4] that $\mathcal{A}_+(n)$ can fail to be star-shaped for $n \geq 3$, $\mathcal{A}_+(n)$ is in fact always a Euclidian space [4].

Appendix. In this section, we provide the proofs deferred from §§4 and 6.

Proof of Theorem 4.3. Under the map $\mathcal{F}^{-1}$ of Corollary 3.4 the initial conditions (a, g) are transformed to (α, γ), where $\gamma_0, \gamma_1, \ldots, \gamma_{n-1}$ are the first n Schur parameters of $v(z)$ and $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the parameters in the Kimura–Georgiou parameterization (3.9). Under the same map, $(a(t), g(t))$ goes into $(\alpha(t), \gamma(t))$ which, according to Corollary 3.4, satisfies

$$\text{(A-1a)} \qquad a(t) = \varphi_n(\gamma(t)) + \Phi_n(\gamma(t))\alpha(t),$$

(A-1b) $$b(t) = \psi_n(\gamma(t)) + \Psi_n(\gamma(t))\alpha(t),$$

(A-1c) $$g(t) = \tfrac{1}{2}[b(t) - a(t)].$$

Now, by Lemma 4.2, $\{\gamma_t, \gamma_{t+1}, \gamma_{t+2}, \ldots\}$ is the Schur parameter sequence of $v_t(z)$, and consequently (4.7) must hold. Therefore, if we can prove (4.8), then we have shown that (4.4b) holds. To this end note that, in view of (4.7), the last of equations (A-1a) reads

$$a_n(t+1) = -\gamma_{t+n-1} - \gamma_{t+n-2}\alpha_1(t+1) - \cdots - \gamma_{t+1}\alpha_{n-1}(t+1) + \alpha_n(t+1).$$

Hence if we can prove that

(A-2) $$a_n(t+1) = (1+\gamma_t)\alpha_n(t+1),$$

then (4.8) follows. However, from (A-1) we see that

(A-3) $$\alpha_n(t) = a_n(t) + g_n(t)$$

for all $t = 0, 1, 2, \ldots$, and then (A-2) follows from the bottom equations in each of (2.16a) and (2.16b). This establishes (4.4b).

To prove (4.4a) first note that since the dynamical system is time-invariant it is enough to show that $\alpha(1) = A(\gamma)\alpha$, i.e., that

(A-4) $$\alpha = A(\gamma)^{-1}\alpha(1),$$

where

(A-5) $$A(\gamma)^{-1} = \begin{bmatrix} 1-\gamma_{n-1}^2 & -\gamma_{n-1}\gamma_{n-2} & -\gamma_{n-1}\gamma_{n-3} & \cdots & -\gamma_{n-1}\gamma_0 \\ 0 & 1-\gamma_{n-2}^2 & -\gamma_{n-2}\gamma_{n-3} & \cdots & -\gamma_{n-2}\gamma_0 \\ 0 & 0 & 1-\gamma_{n-3}^2 & \cdots & -\gamma_{n-3}\gamma_0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1-\gamma_0^2 \end{bmatrix}.$$

In view of (4.8), proving (A-4) amounts to proving that

(A-6) $$\alpha_j = \gamma_{n-j}\gamma_n + \gamma_{n-j}\gamma_{n-1}\alpha_1(1) + \cdots + \gamma_{n-j}\gamma_{n-(j-1)}\alpha_{j-1}(1) + \alpha_j(1),$$

for $j = 1, 2, \ldots, n$. Now, after the change of coordinates (A-1), the kth equation of (2.16a) reads

$$\begin{aligned} &\varphi_{nk}^{(1)} + \varphi_{n-1,k-1}^{(1)}\alpha_1(1) + \cdots + \varphi_{n-k+1,1}^{(1)}\alpha_{k-1}(1) + \alpha_k(1) \\ &\quad = \frac{1}{1-\gamma_0}\{\pi_{nk} - \rho_{n,k+1} + (\pi_{n-1,k-1} - \rho_{n-1,k})\alpha_1 + \cdots \\ &\qquad + (\pi_{n-(k-1),1} - \rho_{n-(k-1),2})\alpha_{k-1}\} + \alpha_k, \end{aligned} \tag{A-7}$$

where $\varphi_{tk}^{(1)} := \varphi_{tk}(\gamma(1))$, and ρ_{tk} and π_{tk} are the coefficients of polynomials (3.16) and (3.18), respectively. Note that $\{\varphi_{tk}^{(1)}\}$ are the coefficients of the Szegö polynomials $\{\varphi_t^{(1)}(z)\}$ corresponding to the shifted Schur parameter sequence $\{\gamma_1, \gamma_2, \gamma_3, \ldots\}$ that are related to $\{\varphi_t(z)\}$ through the algebraic identity [16] (see also [12], [9])

$$\begin{bmatrix} \psi_{t+1}(z) \\ \varphi_{t+1}(z) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} (1+\gamma_0)(z+1) & (1-\gamma_0)(z-1) \\ (1+\gamma_0)(z-1) & (1-\gamma_0)(z+1) \end{bmatrix}\begin{bmatrix} \psi_t^{(1)}(z) \\ \varphi_t^{(1)}(z) \end{bmatrix},$$

which can be inverted to yield

$$\begin{bmatrix} \psi_t^{(1)}(z) \\ \varphi_t^{(1)}(z) \end{bmatrix} = \frac{1}{2}\frac{1}{z(1-\gamma_0^2)} \begin{bmatrix} (1-\gamma_0)(1+z) & (1-\gamma_0)(1-z) \\ (1+\gamma_0)(1-z) & (1+\gamma_0)(1+z) \end{bmatrix} \begin{bmatrix} \psi_{t+1}(z) \\ \varphi_{t+1}(z) \end{bmatrix}.$$

Then

$$\varphi_t^{(1)}(z) = \frac{1}{2(1-\gamma_0)}[\pi_{t+1}(z) - z\rho_{t+1}(z)]. \tag{A-8}$$

Using recursion (3.5) it is easy to see that

$$\varphi_{t+1}(z) = z\varphi_t(z) + \gamma_t\gamma_{t-1}z\varphi_{t-1}(z) + \gamma_t\gamma_{t-2}z\varphi_{t-2}(z) + \cdots + \gamma_t\gamma_0 z - \gamma_t. \tag{A-9}$$

Similarly, changing the signs of the Schur parameters in (A-9), we also have

$$\psi_{t+1}(z) = z\psi_t(z) + \gamma_t\gamma_{t-1}z\psi_{t-1}(z) + \gamma_t\gamma_{t-2}z\psi_{t-2}(z) + \cdots + \gamma_t\gamma_0 z + \gamma_t. \tag{A-10}$$

Combining (A-9) with (A-10) yields the recursions

$$\pi_{t+1}(z) = z\{\pi_t(z) + \gamma_t\gamma_{t-1}\pi_{t-1}(z) + \cdots + \gamma_t\gamma_1\pi_1(z) + \gamma_t\gamma_0\}, \tag{A-11}$$

and

$$\rho_{t+1}(z) = z\rho_t(z) + \gamma_t\gamma_{t-1}z\rho_{t-1}(z) + \cdots + \gamma_t\gamma_1 z\rho_1(z) + \gamma_t. \tag{A-12}$$

Now, inserting (A-11) and (A-12) into (A-8), we obtain

$$\begin{aligned}\varphi_t^{(1)}(z) = \tfrac{1}{1-\gamma_0}\{&\pi_t(z) - z\rho_t(z) + \gamma_t\gamma_{t-1}[\pi_{t-1}(z) - z\rho_{t-1}(z)] \\ &+ \gamma_t\gamma_{t-2}[\pi_{t-2}(z) - z\rho_{t-2}(z)] + \cdots + \gamma_t\gamma_1[\pi_1(z) - z\rho_1(z)]\} - \gamma_t\end{aligned}$$

which, after identifying coefficients and observing that $\rho_{j1} = \gamma_0$ for $j = 1, 2, 3, \ldots,$ yields

$$\begin{aligned}\varphi_{tk}^{(1)} = \frac{1}{1-\gamma_0}&[\pi_{tk} - \rho_{t,k+1} + \gamma_t\gamma_{t-1}(\pi_{t-1,k-1} - \rho_{t-1,k}) + \cdots \\ &+\gamma_t\gamma_{t-k+1}(\pi_{t-k+1,1} - \rho_{t-k+1,2})] + \gamma_t\gamma_{t-k}, \qquad k = 1, \ldots, n.\end{aligned} \tag{A-13}$$

We now prove (A-6) by induction. For $j = 1$, (A-7) reads

$$\varphi_{n1}^{(1)} + \alpha_1(1) = \frac{1}{1-\gamma_0}[\pi_{n1} - \rho_{n2}] + \alpha_1. \tag{A-14}$$

On the other hand, (A-13) yields

$$\varphi_{n1}^{(1)} = \frac{1}{1-\gamma_0}[\pi_{n1} - \rho_{n2}] + \gamma_n\gamma_{n-1}$$

and therefore

$$\alpha_1 = \alpha_1(1) + \gamma_n\gamma_{n-1},$$

which shows that (A-6) is true for $j = 1$. Next, suppose that (A-6) holds for $j = 1, 2, \ldots, k-1$. We need to prove that (A-6) holds for $j = k$. To this end, use (A-6)

for $j = 1, 2, \dots, k-1$ to eliminate $\alpha_1, \alpha_2, \dots, \alpha_{k-1}$ from (A-7) and use (A-13) to eliminate $\varphi_{nk}^{(1)}, \varphi_{n-1,k-1}^{(1)}, \dots, \varphi_{n-k+1,1}^{(1)}$. This yields, after some simple calculations, (A-6) for $j = k$, as required.

Finally, formula (4.9) is obtained from (2.17) by merely inserting $a(t)$ and $g(t)$ as exhibited in (A-1). □

Proof of Lemma 4.4. It follows from (3.5) that $\{\pi_t\}$ as defined by (3.18) satisfies the recursion

$$\begin{aligned} \pi_{t+1}(z) &= z\pi_t(z) + \gamma_t \rho_t^*(z), \qquad \pi_0(z) = 1, \\ \rho_{t+1}^*(z) &= \rho_t^*(z) + \gamma_t z \pi_t(z), \qquad \rho_0^*(z) = 0, \end{aligned}$$

where $\rho_t^*(z) := z^n \rho_t(1/z)$, from which (4.12) is easily derived. Next, let $D(z, z^{-1})$ be defined by (2.13), and let $d(z)$ be the corresponding polynomial in (4.10). In view of (3.12),

$$D(z, z^{-1}) = \sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_i \alpha_j \sigma_{ij}(z, z^{-1}), \tag{A-15}$$

where $\alpha_0 = 1$ and

$$\sigma_{ij}(z, z^{-1}) := \tfrac{1}{2}[\varphi_i(z)\psi_j(1/z) + \psi_i(z)\varphi_j(1/z)]. \tag{A-16}$$

Now, it is well-known and easy to check that

$$\sigma_{ii}(z, z^{-1}) = r_i, \tag{A-17a}$$

and hence by using (3.5) we see that

$$\sigma_{i,i-1}(z, z^{-1}) = r_{i-1} z. \tag{A-17b}$$

Then, for $j = 1, \dots, i$, we obtain, by induction and repeated use of (3.5)

$$\sigma_{i,i-j}(z, z^{-1}) = r_{i-j}(z^j + p_1 z^{j-1} + \cdots + p_{j-1} z), \tag{A-17c}$$

where $p_1, p_2, \dots, p_{j-1}$ are functions of $\gamma_{i-j}, \gamma_{i-j+1}, \dots, \gamma_{i-1}$ only. Consequently, it follows from (A-15) that

$$d_0 = \alpha_n^2 + r_1 \alpha_{n-1}^2 + \cdots + r_n, \tag{A-18}$$

and that

$$\begin{aligned} d(z) - \tfrac{1}{2} d_0 &= \textstyle\sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \alpha_i \alpha_j \sigma_{n-i,n-j}(z, z^{-1}) \\ &= \textstyle\sum_{i=0}^{n-2} \sum_{j=i+1}^{n-1} \alpha_i \alpha_j \sigma_{n-i,n-j}(z, z^{-1}) + \alpha_n \sum_{j=1}^{n} \alpha_{n-j} \pi_j(z) \end{aligned} \tag{A-19}$$

because $\sigma_{j0} = \pi_j$. Now, writing $d^{(n)}$ instead of $d(z)$ to stress the fact that n is the dimension of α or γ (but *not* necessarily the degree of $d(z)$), we observe that the first term of (A-19) equals $d^{(n-1)}(z) - \frac{1}{2} d_0$ except that σ_{kl} has been replaced by $\sigma_{k+1,l+1}$, a replacement that, according to (A-17), amounts to exchanging $\{\gamma_0, \gamma_1, \dots, \gamma_{n-1}, \gamma_0, \gamma_1, \dots, \gamma_{n-1}\}$ by $\{\gamma_1, \gamma_2, \dots, \gamma_n, \gamma_1, \gamma_2, \dots, \gamma_n\}$. Consequently, since $d^{(1)}(z) = \frac{1}{2} d_0 + \alpha_1 z$, the coefficients of (4.10) are generated by the recursion in $d^{(k)}$ defined in the lemma.

Finally, suppose $d_i = 0$ for $i = 1, 2, \ldots, n$. Then, since $1 - \gamma_i^2 \neq 0$ for $i = 0, 1, 2, \ldots, n-1$, it follows from the recursion in $d^{(k)}$ that $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. But then, by (4.11), $d_0 = r_n := \prod_{i=0}^{n-1}(1 - \gamma_i^2) \neq 0$. Hence at least one of the coefficients must be nonzero as claimed. □

Proof of Lemma 6.8. Let $\begin{bmatrix} X \\ Y \end{bmatrix}$ be a matrix basis of $\mathcal{U}$ as in (6.19). Then

$$\mathcal{W} = \left\{ \begin{bmatrix} X \\ Y \end{bmatrix} z \mid z \in \ker Y \right\}. \tag{A-20}$$

The S-invariance of $\mathcal{U}$ implies that there is an $n \times n$ matrix R such that

$$S \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} R, \tag{A-21}$$

and therefore (i) holds if and only if $Rz \in \ker Y$ for all $z \in \ker Y$, i.e., $\ker Y$ is R-invariant. Set $z \in \ker Y$. Then, using (6.14), the second block of (A-21) yields

$$gg'(A')^{-1}Xz = YRz. \tag{A-22}$$

Since $\mathcal{U}$ is Lagrangian, $X'Y = Y'X$ so that $z'X'Y = 0$, and therefore

$$[z'X'g][g'(A')^{-1}Xz] = 0. \tag{A-23}$$

Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be the largest subspaces of $\ker Y$ for which $g'(A')^{-1}Xz = 0$ and $g'Xz = 0$, respectively. Then for (A-23) to hold for all $z \in \ker Y$, either $\mathcal{V}_1$ or $\mathcal{V}_2$ must be all of $\ker Y$. In fact, if there are two one-dimensional subspaces $l_1 \in \mathcal{V}_1$ and $l_2 \in \mathcal{V}_2$, then an arbitrary point in the plane spanned by l_1 and l_2 must belong to $\ker Y$ and hence to either $\mathcal{V}_1$ or $\mathcal{V}_2$ for (A-23) to hold; say $\mathcal{V}_1$. But then the whole plane must belong to $\mathcal{V}_1$, and hence also l_2.

Now, first suppose that $g'(A')^{-1}Xz = 0$ for all $z \in \ker Y$. Then, it follows from (A-22) that $\ker Y$ is R-invariant, and hence (i) holds.

Next, suppose that $g'Xz = 0$ for all $z \in \ker Y$. Since $\alpha_n \neq 0$, $\kappa_n \neq 0$, and hence S is nonsingular, (Proposition 6.6). Therefore, in view of (A-21), R is also nonsingular, so that

$$S^{-1} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} R^{-1} \tag{A-24}$$

and consequently

$$S^{-1}\mathcal{W} = \left\{ \begin{bmatrix} X \\ Y \end{bmatrix} R^{-1}z \mid z \in \ker Y \right\},$$

which belongs to $\mathcal{W}$ if and only if $R^{-1}z \in \ker Y$ for all $z \in \ker Y$. Now, by (6.12), $S^{-1} = \hat{J}^{-1}S'\hat{J}$ and hence (A-24) is equivalent to

$$S' \begin{bmatrix} -Y \\ X \end{bmatrix} = \begin{bmatrix} -Y \\ X \end{bmatrix} R^{-1}. \tag{A-25}$$

Taking $z \in \ker Y$ and remembering that $g'Xz = 0$, the top block of (A-25) yields $YR^{-1}z = 0$, which is what is required for condition (ii) to hold. Hence, we have proved that at least one of conditions (i) and (ii) holds.

Finally, we prove that these conditions are actually equivalent. Suppose $\dim \mathcal{W} = k$. If $k = 0$, the statement is trivial, so we assume that $k > 0$. Then

$$\mathcal{W} = \operatorname{Im} \begin{bmatrix} U \\ 0 \end{bmatrix} \tag{A-26}$$

for some full-rank $n \times k$ matrix U such that

$$S \begin{bmatrix} U \\ 0 \end{bmatrix} = \begin{bmatrix} U \\ 0 \end{bmatrix} T. \tag{A-27}$$

Since S is nonsingular, so is T, so (A-27) is equivalent to

$$S^{-1} \begin{bmatrix} U \\ 0 \end{bmatrix} = \begin{bmatrix} U \\ 0 \end{bmatrix} T^{-1}, \tag{A-28}$$

which holds if and only if $S^{-1}\mathcal{W} \subset \mathcal{W}$. This concludes the first part of the lemma. The proof of the second part concerning $\tilde{\mathcal{W}}$ is analogous. □

Acknowledgments. We thank Professors Gy. Michaletzky and Michael Benedicks for many helpful discussions, and the anonymous referees for their valuable advice.

REFERENCES

[1] N. I. AKHIEZER, *The Classical Moment Problem*, Hafner, New York, 1965.
[2] G. AMMAR AND C. MARTIN, *The geometry of matrix eigenvalue methods*, Acta Appl. Math., 5 (1986), pp. 239–279.
[3] R. S. BUCY, D. RAPPAPORT, AND L. M. SILVERMAN, *Correlated noise filtering and invariant directions for the Riccati equation*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 535–540.
[4] C. I. BYRNES AND A. LINDQUIST, *On the geometry of the Kimura–Georgiou parameterization of modelling filter*, Internat. J. Control, 50 (1989), pp. 2301–2321.
[5] C. I. BYRNES, A. LINDQUIST, AND T. MCGREGOR, *Predictability and unpredictability in Kalman filtering*, IEEE Trans. Automat. Control, 36 (1991), pp. 563–579.
[6] C. I. BYRNES, A. LINDQUIST, AND Y. ZHOU, *Stable, unstable and center manifolds for fast filtering algorithms*, in Modeling, Estimation and Control of Systems with Uncertainty, G. B. D. Masi, A. Gombani, and A. Kurzhanski, eds., Birkhauser, Boston, MA, 1991.
[7] S. W. CHAN, G. C. GOODWIN, AND K. S. SIN, *Convergence properties of the Riccati difference equation in optimal filtering of nonstablizable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 110–118.
[8] C. CHANG AND T. T. GEORGIOU, *On a Schur-algorithm based approach to spectral factorization: Connection with the Riccati equation*, Linear Algebra Appl., 171 (1992), pp. 233–248.
[9] P. DELSATRE, Y. GENIN, AND Y. KAMP, *Schur parameterization of positive definite block-Toeplitz systems*, SIAM J. Appl. Math., 36 (1979), pp. 34–46.
[10] C. J. DEMEURE AND C. T. MULLIS, *The Euclid algorithm and the fast computation of cross-covariance and autocovariance sequences*, IEEE Trans. Acoustics, Speech Signal Process., ASSP-37 (1989), pp. 545–552.
[11] R. L. DEVANEY, *An Introduction to Chaotic Dynamic System*, Addison–Wesley, Reading, MA, 1987.
[12] P. DEWILDE, A. VIEIRA, AND T. KAILATH, *On a generalized Szegö–Levinson realization algorithm for optimal linear predictors based on a network synthesis approach*, IEEE Trans. Circuits Systems, CAS-25 (1978), pp. 663–675.
[13] P. FAURRE, M. CLERGET, AND F. GERMAIN, *Opérateurs Rationnels Positifs*, Dunod, Paris, 1979.
[14] F. R. GANTMACHER, *The Theory of Matrix*, Chelsea, New York, 1959.
[15] T. T. GEORGIOU, *Partial realization of covariance sequences*, Ph.D. thesis, Center for Mathematical Systems Theory, Univ. Florida, Gainesville, 1983.
[16] ———, *Realization of power spectra from partial covariance sequences*, IEEE Trans. Acoustics, Speech, Signal Process., ASSP-35 (1987), pp. 438–449.

[17] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, Univ. California Press, Berkeley, Los Angeles, 1958.

[18] J. GUCKENHEIMER AND P. HOLMES, *Nonlinear Oscilations, Dynamical Systems and Bifurcation of Vector Fields*, Springer-Verlag, New York, 1983.

[19] M. HAGSTRÖM AND Y. ZHOU, *On the geometry of the sign-definite regions in the generalized Kimura–Georgiou parameterization*, to appear.

[20] A. HURWITZ, *Über die Bedingungen unter welchen eine Gleichung nur Wurzeln mit Negativen Reelen Teilen Besitzt*, Math. Ann., 46 (1895), pp. 273–284.

[21] R. E. KALMAN, *Realization of covariance sequences*, in Proc. Toeplitz Memorial Conference, Tel Aviv, Israel, 1981.

[22] H. KIMURA, *Positive partial realization of covariance sequences*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1987, pp. 499–513.

[23] V. KUČERA, *A review of the matrix Riccati equation*, Kybernetika, 9 (1973), pp. 42–61.

[24] N. LEVINSON, *The Wiener RMS (root mean squares) error in filter design and prediction, Appendix* 13, in Wiener, Extrapolation, Interpolation and Smoothing of Stationary Time Series, M.I.T. Press, Cambridge, MA, 1942.

[25] A. LINDQUIST, *A new algorithm for optimal filtering of discrete-time stationary processes*, SIAM J. Control, 12 (1974), pp. 736–746.

[26] ———, *On Fredholm integral equations, Toeplitz equations and Kalman–Bucy filtering*, Appl. Math. Optim., 1 (1975), pp. 355–373.

[27] ———, *Some reduced-order non-Riccati equations for linear least-squares estimation: the stationary, single-output case*, Internat. J. Control, 24 (1976), pp. 821–842.

[28] J. E. MARSDEN AND M. MCCRACKEN, *The Hopf Bifurcation and its Applications*, Springer-Verlag, New York, 1976.

[29] C. MARTIN, *Grassmannian manifolds, Riccati equations, and feedback invariants of linear systems*, in Geometrical Methods for the Theory of Linear Systems, C. I. Byrnes and C. Martin, eds., Reidel Publishing Company, 1980, pp. 195–211.

[30] B. PARLETT AND W. G. POOLE, JR, *A geometric theory for the QR, LU, and power iterations*, SIAM J. Numer. Anal., 10 (1973), pp. 389–412.

[31] M. PAVON, *Stochastic realization and invariant directions of the matrix Riccati equation*, SIAM J. Control Optim., 18 (1980), pp. 155–180.

[32] I. SCHUR, *On power series which are bounded in the interior of the unit circle I and II*, J. Angew. Math., 148 (1918), pp. 122–145.

[33] I. R. SHAFAREVITCH, *Basic Algebraic Geometry*, Springer-Verlag, Heidelberg, 1974.

[34] M. SHAYMAN, *Geometry of the algebraic Riccati equations, part* I *and* II, SIAM J. Control Optim., 21 (1983), pp. 375–394.

[35] ———, *Phase portrait of the matrix Riccati equations*, SIAM J. Control Optim., 24 (1986), pp. 1–65.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 790–807, May 1994

009

CONTROLLED INVARIANCE FOR SINGULAR DISTRIBUTIONS*

VISWANATH RAMAKRISHNA†

Abstract. For an analytic affine control system and an analytic locally controlled invariant distribution Δ with singularities, which is already invariant under the vector fields corresponding to the controls (in other words is input-insensitive), the problem of finding a feedback that will render Δ controlled invariant is studied. This can be done if one can solve a degenerate system of partial differential equations (PDEs). The latter can be achieved in the presence of scaling-like symmetries.

Key words. controlled invariance, distributions, Poincaré vector field, degenerate differential equations, overdetermined systems

AMS subject classifications. 93C10, 93C15, 93C25

1. Introduction.

Suppose we are given (i) an affine nonlinear control system

$$\dot{x} = f(x) + \sum_{i=1}^{m} g_i(x)u_i,$$

$x \in \mathbb{R}^n$, and f and g_i, smooth vector fields on $\mathbb{R}^n$; (ii) a smooth distribution Δ on $\mathbb{R}^n$. The controlled invariance problem then seeks feedback functions α_i and $\beta = (\beta_{ij})$ such that the resulting closed-loop system leaves Δ invariant; i.e., $[f + \sum_{i=1}^{m} g_i\alpha_i, \Delta] \subseteq \Delta$ and $[(g\beta)_i, \Delta] \subseteq \Delta$. This problem arises in a variety of applications, such as disturbance decoupling and noninteracting control [15], [23]. This problem is known to have a solution (under some assumptions) in a neighborhood of a point where Δ is nonsingular [15], [23]. The purpose of this paper is to provide some sufficient conditions under which the problem has a solution when Δ becomes singular. To clarify exactly the kind of distributions considered in this paper, we begin by reviewing some terminology from differential geometry.

A distribution Δ on U (U an open set of $\mathbb{R}^n$) is an assignment to each point x of U, a subspace $\Delta(x)$ of T_xU. A C^∞ vector field Y on U is said to belong to Δ if $Y(p) \in \Delta(p)$ for all $p \in U$. Δ is said to be smooth if, at each $p \in U, \Delta(p)$ is the linear hull of the vectors $\{X(p) | X \in D\}$, where D is the set of C^∞ vector fields on U that belong to Δ.

Given an open subset V of U, we denote by $\Delta|_V$ the restriction of the assignment Δ to V. Δ is said to be locally finitely generated if, about each $p \in U$, there is an open subset $V \subseteq U$ that contains p with the property that the set D of C^∞ vector fields on V that belong to $\Delta|_V$ is finitely generated over the ring of smooth (or analytic) functions on the open set V; i.e., there exist vector fields $X_1, \ldots, X_d \in D$ so that, if $Y \in D$, then we have

$$Y = \sum_{i=1}^{d} f_i X_i,$$

where the f_i are C^∞ (or analytic) functions on V.

The expression "Δ is generated by d smooth vector fields an open set V" means the same thing as the above equation. Note that, if Δ is generated on V by $X_1, \ldots, X_d$,

* Received by the editors April 3, 1991; accepted for publication (in revised form) October 28, 1992.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130. Current Address, Frick Laboratories, Princeton University, Princeton, New Jersey 08544.

then $X_1(p), \ldots, X_d(p)$ span the vector space $\Delta(p)$ whenever $p \in V$. We say Δ is involutive if, whenever X and Y belong to Δ, so does $[X, Y]$. All distributions in this article are involutive and smooth. In the analytic case, all the above definitions remain valid with "analytic" replacing "C^∞." There is one important difference, however—an analytic distribution is always locally finitely generated.

In a neighborhood U of a reference point $p \in \mathbb{R}^n$, we make the following assumptions.

(A) Δ is involutive and finitely generated over $C^\infty(U)$ (respectievly, $C^\omega(U)$) by the vector fields $X_1, \ldots, X_d$.

(B) Δ is locally controlled invariant (also known as weakly (f, g) invariant), in the sense that

$$[f, X_i] = V_i + \sum_{j=1}^{m} c_j^i g_j, \qquad 1 = 1, \ldots, d$$

and

$$[g_j, X_i] = W_{ij} + \sum_{k=1}^{m} b_{jk}^i g_k, \qquad j = 1, \ldots, m, \quad i = 1, \ldots, d,$$

where V_i and W_{ij} are vector fields that belong to Δ; (c_j^i) and (b_{jk}^i) belong to $C^\infty(U)$ (respectively, $C^\omega(U)$).

It is known, then, that the required feedback functions may be found by solving the following partial differential equations (PDEs):

$$L_{X_i}(\alpha_k) - \sum_{j=1}^{m} b_{jk}^i \alpha_j = c_k^i, \qquad i = 1, \ldots, d, \quad k = 1, \ldots, m, \tag{1}$$

$$L_{X_i}(\beta_{jk}) - \sum_{r=1}^{m} b_{rk}^i \beta_{jr} = 0, \qquad i = 1, \ldots, d, \quad j, k = 1, \ldots, m. \tag{2}$$

(C) Δ is input-insensitive, i.e., $[g_j, \Delta] \subseteq \Delta$. We relax this condition in some cases. However, unless explicitly stated to the contrary, (C) is operative throughout this paper. If (C) holds, then only (1), which now takes the form

$$L_{X_i}(\alpha_k) = c_k^i, \qquad i = 1, \ldots, d, \quad k = 1, \ldots, b, \tag{3}$$

must be solved.

(D) The control vector fields $g_1, \ldots, g_m$, are linearly independent on a dense subset of U.

(E) $G \cap \Delta = \{0\}$ on a dense subset of U.

Then it is known (see [15]) that (B), (D), and (E) provide the integrability conditions for (2) and (3) on a dense subset of U and hence everywhere. The proof of this fact is a simple consequence of the Jacobi identity and the above hypotheses. If now, in addition, Δ is nonsingular, then (2) and (3) do indeed have smooth (respectively, analytic) solutions. In [15] the proof that (B), (D), and (E) yield the integrability conditions is given under the assumption that these conditions hold everywhere, not just on a dense subset. However, the same proof carries over, under our assumptions, to yield the conclusion on a dense subset (and hence everywhere). It must be emphasised

that we require "nice" behavior only on a dense subset and not everywhere, not out of a desire to be perversely general, but because this situation is forced upon us. Loosely speaking, the joint requirements that Δ be singular and that it be input-insensitive will, more often than not, force $G+\Delta/\Delta$ also to pick up singularities. Unless otherwise stated, therefore, we assume that all of the conditions (A)–(E) hold throughout this paper. We emphasize that this means that the integrability conditions for (2) and (3) are valid on all of U. Thus we tacitly assume that the integrability conditions, which are obviously necessary for the local solvability of these systems of partial differential equations, always hold.

Remark 1.1. (a) If Δ is nonsingular and so is G, then (B) and (E) are equivalent to $[f,\Delta] \subseteq \Delta + G$ and $[g_i,\Delta] \subseteq \Delta + G$. In fact, the smooth (respectively, analytic) implicit function theorem proves the smoothness of the functions (c_j^i) and (b_{jk}^i) and, once we have specified a basis for the distributions Δ and G, their uniqueness as well. We must assume (D) in the singular case. Note, however, that once this is assumed, the uniqueness follows from the denseness assumption in (D) and (E).

(b) Note that (A) and (D) hold automatically in the analytic case.

(c) Note that (D) and (E) imply that $G + \Delta/\Delta$ is nonsingular on a dense subset. The converse is not true however. We make the more restrictive assumptions because we cannot ensure, without additional assumptions, the possibility of an analytic extension of a separating feedback to all of U.

In §3 we obtain a "canonical" form for a basis for a locally finitely generated distribution with singularities. After the submission of this paper, it was brought to our attention by Dr. K. Grasse that our Theorem 3.1 could be deduced from [26, Thm. 1]. We wish to thank him for this. In [26] the author shows that involutive and locally finitely generated vector field systems give rise to a foliation with singularities. In particular, this means that at each point there is a "privileged chart." Using such a chart, we can produce a basis of fields of the type in Theorem 3.1. However, the proof presented here is more elementary and is constructive to the extent that the flowbox theorem is. We must also mention that singular distributions have been studed previously in other contexts [19], [12], [28], [26]. See [13] for a different study of a canonical form for involutive distributions in connection with control theory.

This canonical form clearly displays that solving (3) entails solving a degenerate system of PDEs (see §2 for a definition of a degenerate PDE). In §4 we show that solving this degenerate system suffices, in that we may construct a solution to all the equations in (3) from a solution to the degenerate system in (3).

Very little is known about even a single degenerate PDE, let alone a system. If we assume that all data are analytic, then, for a single analytic equation, there are conditions on the first-order eigenvalues of the vector field that guarantee a solution. These conditions are described in [3] and [4]. Since we use their results in more than one way, they have been summarized in §2.

Using these results, we are able to give one situation where a solution to (1.3) can be obtained (Theorem 5.3). To obtain Theorem 5.3, we must make a suitable basis change in Δ, for which we again must solve degenerate PDEs. It turns out that the scaling vector field plays an important role in this result.

For a different approach to the problem of controlled invariance of singular distributions, see [6] and [7]. For an example of a rigid body problem involving a distribution with singularities, see [22] and [24]. Consult the text [23] in connection with the problem of controlled invariance for general nonlinear systems, which has the merit of providing a way of studying controlled invariance for nonlinear discrete time sys-

tems [11]. The same text also formulates the integrability conditions for the PDEs under study in terms of the existence of flat connections. Finally, we refer to [8]–[10], [14]–[17], [20], [21], [23], [29] for more on the problem of controlled invariance.

2. Formal and convergent solutions to PDEs. In this section, we summarize results of Bengel and Gerard on the convergence of formal solutions to a degenerate PDE. We consider the following two kinds of equations:

$$L_X v = s(x) \tag{4}$$

and

$$L_X v = s(x, v). \tag{5}$$

In (4) and (5), both v and s can be r-vectors. If $v = (v_1, \ldots, v_r)$, then $L_X v$ is to be interpreted as the r-vector $(L_X v_1, \ldots, L_X v_r)$. s is taken to be an analytic r vector. The PDEs (4), (5) are said to be degenerate (at the origin) if the differential operator on the left-hand side vanishes at the origin. This happens if and only if the vector field X vanishes at the origin. In [3], [4], vector fields of the form

$$X = \sum_{j=1}^{n} (\gamma_j x_j + p_j(x_1, \ldots, x_n)) \frac{\partial}{\partial x_j}, \tag{6}$$

with $\gamma_j \in \mathbb{R}$ and the p_j analytic functions vanishing to the second-order at the origin, are studied.

For a multi-index $m = (m_1, \ldots, m_n)$ where the m_i are nonnegative integers, denote by $P_0(m)$ the quantity

$$P_0(m) = \left(\sum_{j=1}^{n} \gamma_j m_j \right) I_r$$

(where I_r is the rth-order identity matrix). Furthermore, let $\mid m \mid = \sum_{i=1}^{n} m_i$. We now have the following theorems.

THEOREM 2.1 (Bengel and Gérard). *If the operator X in* (6) *satisfies the estimate*

$$\mid P_0(m) \mid \geq C \mid m \mid, \qquad m \neq (0, \ldots, 0)$$

for some fixed positive constant C, and the obvious necessary condition $s(0) = 0$ holds, then, in a neighborhood of the origin, there exists an analytic solution to (4).

THEOREM 2.2 (Bengel and Gérard). *If the estimate of Theorem* 2.1 *holds and we have, in addition, that*

$$s(0,0) = 0, \qquad \mathrm{Jac}_v s(0,0) = 0,$$

where $\mathrm{Jac}_v s$ is the Jacobian matrix of s with respect to the v variables, then there exists, in a neighborhood of the origin in R^n, an analytic solution to (5).

Now consider the following equation:

$$L_X v = s(x, y, v) \tag{7}$$

with X as in (6), s analytic in all three variables x, y, and v, and where $y \in \mathbb{R}^p$ is a parameter.

Notation. If $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $m = (m_1, \ldots, m_n)$ is a multi-index, then x^m denotes $x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n}$.

THEOREM 2.3 (Bengel and Gérard). *Let s in (7) satisfy $s(0,0,0) = 0$ and $(\partial s/\partial v)(0,0,0) = 0$. Assume (7) has a formal solution*

$$w = \sum_k w_k(y) x^k$$

(where k is a multi-index) with $w_k(0) = 0$. Then (7) has a solution

$$v = \sum_k v_k(y) x^k$$

that converges. Furthermore, given any ν, the v_k can be chosen so that w and v agree up to order ν.

The condition $P_0(m) \neq 0$ for all $m \neq 0$ allows us to set up formal solutions to the degenerate PDE (4) and to (5) when the latter is linear in v. For the more general case, we must linearize (5) and use a "Newton's method-" type argument. See [3] and [4] for more details. The estimate on $P_0(m)$ then allows us to establish convergence of the formal power series for the putative solution. Consult [2] for other situations that give rise to similar notions, especially in the study of *ordinary* differential equations (ODEs) with regular and irregular singularities.

Remark 2.1. (a) In both Theorems 2.1 and 2.2, the solution is unique up to the choice of an additive constant.

(b) We have not stated the results of [2] and [3] in full generality even for (4), (5), and (7). In this paper, we just want to give a flavor of the kind of estimates (usually called Poincaré or Siegel conditions) needed to obtain convergence of formal solutions. For a detailed description of Poincaré or Siegel conditions, consult [2].

(c) Theorem 2.2 would hold even if $(\partial f/\partial u)(0,0) \neq 0$. Denote by $\Lambda(m)$ the matrix $P_0(m) - (\partial f/\partial u)(0,0)$. If $\Lambda(m)$ satisfies the estimate $|\Lambda(m)| > C|m|$ for some positive constant C and all $m \neq 0$, then we can find an analytic solution to (5). Of course, now it may not be possible to prescribe the initial value of the solution.

We close this section with a definition that is required later.

DEFINITION 2.1. *A Poincaré vector field is an analytic vector field of the form in (6) satisfying the estimate in Theorem* 2.1.

3. Canonical basis for singular distributions. The goal of this section is to prove the following theorem, in which we work in a neighborhood V of the reference point, where Δ is locally finitely generated.

THEOREM 3.1. *Let Δ be an involutive, smooth distribution that is generated by d vector fields in a neighborhood V of $p \in \mathbb{R}^n$ (d can be greater than n). Suppose that $\dim\Delta(p) = d_0$, where $d_0 \leq d$. Then there exists a coordinate system $(U, (x_1, \ldots, x_n))$ centered at p such that Δ is spanned over $C^\infty(U)$ by*

$$X_l = \frac{\partial}{\partial x_l}, \qquad l = 1, \ldots, d_0, \tag{8}$$

$$X_l = \sum_{j=d_0+1}^{n} a_j^l(x_{d_0+1}, \ldots, x_n) \frac{\partial}{\partial x_j}, \qquad l = d_0 + 1, \ldots, d, \tag{9}$$

where the a_j^l are smooth functions that vanish at the origin in $\mathbb{R}^{n-d_0}$.

Proof. We induct on the triples $(\mathrm{n}, \mathrm{d}, d_0)$ ordered lexicographically. The result is vacuously true when the triplet takes the value (0,0,0). We assume the result is true for triplets $(\mathrm{m}, \mathrm{l}, l_0)$, which satisfy one of the following: (a) $m < n$, (b) $m = n, l < d$, (c) $m = n, l = d, l_0 < d_0$.

The inductive step. Before we resume the proof, we make two observations.

Observation A. We may assume that $0 < d_0 < d$. If $d_0 = 0$, the original coordinate system will suffice. If $d_0 = d$, the classical Frobenius theorem gives the result.

Observation B. Let Δ be an involutive distribution in R^n. Suppose that at some stage, we have found a basis for Δ, in some neighborhood, of the following form:

$$X_l = \frac{\partial}{\partial x_l}, \qquad l = 1, \ldots, q$$

and

$$X_l = \sum_{j=1}^{n} a_j^l(x) \frac{\partial}{\partial x_j}, \qquad l = q+1, \ldots, r.$$

Then Δ is spanned in that neighborhood by

$$X_l = \frac{\partial}{\partial x_l}, \qquad l = 1, \ldots, q$$

and

$$\tilde{X}_l = \sum_{j=q+1}^{n} a_j^l(x_1, \ldots, x_n) \frac{\partial}{\partial x_j}, \qquad l = q+1, \ldots, r.$$

In essence, we "chop off" from the expression displaying any element of Δ as a smooth linear combination of the X_l's, the $\partial/\partial x_j, j = 1, \ldots, q$ terms from the last $q+1$ to r terms and add them to the first q terms.

We now resume the proof of the theorem with the inductive step.

We can assume that Δ now has a basis of the form $X_1, \ldots, X_d$, where $X_1, \ldots, X_{d_0}$ are linearly independent and $X_l(0) = 0$ for $l = d_0 + 1, \ldots, d$. This is achieved by using the linear dependence relations at the origin among the basis elements of Δ, corresponding to the drop in rank of Δ there, to effect a linear basis change for the distribution Δ. In particular, since $X_1(0) \neq 0$, we can by the flowbox theorem (see [1], [5]) find a coordinate system in which $X_1 = \partial/\partial x_1$. In this coordinate system, Δ has a basis consisting of X_1 and

$$X_j = \sum_{i=1}^{n} a_i^j(x) \frac{\partial}{\partial x_i}, \qquad j = 2, \ldots, d. \tag{10}$$

Consequently, by Observation B, Δ is spanned in this neighborhood by

$$X_1 = \frac{\partial}{\partial x_1}, \quad \tilde{X}_j = \sum_{i=2}^{n} a_i^j(x) \frac{\partial}{\partial x_i}, \qquad j = 2, \ldots, d. \tag{11}$$

In particular, $\tilde{X}_2, \ldots, \tilde{X}_{d_0}$ are linearly independent; their span $\tilde{\Delta}$ is involutive (since Δ was, and none of the $\tilde{X}_j, j = 2, \ldots, d$ have a $\partial/\partial x_1$ component); and $\dim\tilde{\Delta}(0) = d_0 - 1$.

Note that Δ is spanned by any basis for $\tilde{\Delta}$ and X_1. We now seek a basis change for $\tilde{\Delta}$ with the property that all basis elements commute with X_1. To that end, let

$$[X_1, \tilde{X}_i] = \sum_{j=2}^{d} \alpha_j^i(x)\tilde{X}_j \tag{12}$$

be the equation expressing the involutivity of Δ. Note that on the right-hand side of (12) there is no component along X_1. We make a basis change of the form

$$X_i \doteq \left(\sum_{k=2}^{d} b_k^i \tilde{X}_k\right), \qquad i = 2, \ldots, d. \tag{13}$$

If $[X_1, X_i]$ is to be zero, we must have

$$\sum_{k=2}^{d} \left(\frac{\partial b_k^i}{\partial x_1}\tilde{X}_k + b_k^i \sum_{j=2}^{d} \alpha_j^k \tilde{X}_j\right) = 0. \tag{14}$$

This leads to $(d-1)$ separate systems of linear ODEs (with parameters $x_2, \ldots, x_n$), below:

$$\frac{db_j^i}{dx_1} + \sum_{k=2}^{n} b_k^i \alpha_j^k = 0, \qquad i, j = 2, \ldots, d. \tag{15}$$

These can be solved, and, by prescribing initial conditions $b_k^i(0)$ so that the matrix of initial conditions is invertible, we can ensure that the basis change thus found is invertible in a small neighborhood. Hence, $\tilde{\Delta}$ has a basis of the form

$$X_i \doteq \sum_{j=2}^{n} d_j^i(x_2, \ldots, x_n)\frac{\partial}{\partial x_j}, \qquad i = 2, \ldots, d. \tag{16}$$

The distribution $\hat{\Delta} \doteq \operatorname{span}_{C^\infty_{R^{n-1}}}\{X_2, \ldots, X_d\}$ is involutive, and it is spanned by $d-1$ vector fields; furthermore, its dimension at the origin is $d_0 - 1$. So, by the inductive hypothesis, we can find a coordinate system centered at the origin in R^{n-1}, $(V, (x_2, \ldots, x_n))$, so that $\tilde{\Delta}$ is spanned by the vector fields

$$X_l = \frac{\partial}{\partial x_l}, \qquad l = 2, \ldots, d_0 \tag{17}$$

and

$$X_l = \sum_{j=d_0+1}^{n} a_j^l(x_{d_0+1}, \ldots, x_n)\frac{\partial}{\partial x_j}, \qquad l = d_0 + 1, \ldots, d \tag{18}$$

with $a_j^l(0) = 0$. Now observe that Δ is spanned on $U = \mathbb{R} \times W$, where W is the neighborhood on which the basis in the penultimate two equations is valid, by X_1 and the basis in these two equations. Thus, finally we have found a coordinate system $(U, (x_1, \ldots, x_n))$ centered at the origin in R^n in which Δ is spanned over $C^\infty(R^n)$, in a neighborhood contained in U by the vector fields

$$X_l = \frac{\partial}{\partial x_l}, \qquad l = 1, \ldots, d_0, \tag{19}$$

$$X_l = \sum_{j=d_0+1}^{n} a_j^l(x_{d_0+1}, \ldots, x_n)\frac{\partial}{\partial x_j}, \qquad l = d_0+1, \ldots, d. \tag{20}$$

In the last equation, we have, of course, $a_j^l(0) = 0$. This finishes the proof of the theorem. □

Remark 3.1. If the distribution in question is analytic, then the construction of the theorem would produce a basis of analytic vector fields.

Remark 3.2. Note that the proof of the theorem is constructive to the extent that the proof technique of the flowbox theorem is constructive.

4. Solving a degenerate system suffices. According to Theorem 3.1, a locally finitely generated involutive C^∞ distribution has a basis consisting of coordinate vector fields and vector fields tangent to a lower-dimensional space and vanishing at the origin. Consequently, the PDEs (3) for rendering this distribution-controlled invariant contain in them a degenerate system of PDEs. We now show that, if we can find a solution to this degenerate system, then we can find a solution to the entire system of equations (3). More precisely, we have the following result.

PROPOSITION 4.1. *Let Δ be a distribution satisfying the hypothesis of Theorem* 3.1. *Let $X_i = \partial/\partial x_i$, $i = 1, \ldots, d_0$, and $X_{d_0+1}, \ldots, X_d$ be a basis for Δ in a neighborhood U of the reference point, as in Theorem* 3.1. *Let*

$$\frac{\partial}{\partial x_i}\alpha = c_i(x), \qquad i = 1, \ldots, d_0, \tag{21}$$

$$L_{X_j}\alpha = c_j(x), \qquad j = d_0+1, \ldots, d \tag{22}$$

be (3) *for Δ with respect to this basis. Assume that Δ satisfies hypotheses* (A)–(E) *of* §1, *so that the integrability conditions for the system of equations* (21), (22) *hold. Then, if there exists a solution $\alpha(x)$ to* (22), *we can find a function $\tilde{\alpha}(x)$ that solves the full system* (21) *and* (22).

Proof. Let $\alpha(x)$ be a solution to (22). Since Δ is weakly (f, g) invariant, we know from §1 that the integrability conditions for (21) and (22) hold on all of U. Since the X_j's for $j = 1, \ldots, d_0$ and the X_i's for $i = d_0+1, \ldots, d$ commute, these conditions contain equations of the form

$$\frac{\partial c_i}{\partial x_j}(x) = L_{X_i}c_j(x), \qquad i = d_0+1, \ldots, d, \quad j = 1, \ldots, d_0. \tag{23}$$

Now consider the following expression:

$$\begin{aligned} L_{X_i}&\left(\frac{\partial \alpha(x)}{\partial x_j} - c_j(x)\right) \\ &= \frac{\partial}{\partial x_j}(L_{X_i}\alpha) - L_{X_i}c_j \\ &= \frac{\partial c_i}{\partial x_j} - L_{X_i}c_j \\ &= 0, \end{aligned}$$

where the second equality results from the fact that the coordinate vector fields in the first d_0 coordinates commute with the X_i's, and the last equality is nothing other than the above integrability conditions. Now there is a second set of integrability

conditions arising by virtue of the commutativity of the coordinate vector fields among themselves. These read as

$$\frac{\partial c_k}{\partial x_l} = \frac{\partial c_l}{\partial x_k}, \qquad k,l = 1,\dots,d_0. \tag{24}$$

Let us now consider the expressions

$$\frac{\partial \alpha(x)}{\partial x_j} - c_j(x) = \gamma_j(x). \tag{25}$$

The γ_j are functions invariant under the X_l for $l > d_0$. The system

$$\frac{\partial \gamma}{\partial x_k} = \gamma_k, \qquad k = 1,\dots,d_0 \tag{26}$$

can be solved, provided the integrability conditions hold for it, namely,

$$\frac{\partial \gamma_q}{\partial x_k} = \frac{\partial \gamma_p}{\partial x_q} \tag{27}$$

(for $p,q = 1,\dots,d_0$). However, the last set of equations reads as

$$\frac{\partial}{\partial x_p}\left(\frac{\partial \alpha}{\partial x_q} - c_q\right) = \frac{\partial}{\partial x_q}\left(\frac{\partial \alpha}{\partial x_p} - c_p\right).$$

This holds since (24) holds. Thus (26) has a solution $\gamma(x)$. Observe that $L_{X_l}\gamma = 0$ whenever $l > d_0$. Consequently, the function $\tilde{\alpha} := \alpha - \gamma$ solves both (21) and (22).

5. Degenerate systems and scaling.

Blanket Assumption. For the remainder of the paper, unless otherwise specified, Δ, f, and g are all analytic. Note that analytic distributions are always locally finitely generated.

According to Proposition 4.1, it suffices to restrict our attention to involutive distributions with $d_0 = 0$. If $d_0 > 0$, we must solve a parameterized system of degenerate equations. If $d_0 = 0$, the parameters disappear. In this section, we consider the latter case.

THEOREM 5.1. *Let Δ be a distribution satisfying the hypotheses of Theorem* 3.1, *with $d_0 = 0$, and assume furthermore that Δ satisfies the hypotheses* (A)–(E) *of* §1. *Suppose that one of the basis vector fields for Δ, say X_1, is a Poincaré vector field. If the remaining basis elements commute with X_1, and $c_j^i(0) = 0$, then there exists an analytic feedback rendering Δ controlled invariant in a neighborhood of the origin.*

Proof. Since Δ satisfies (B), (D), and (E), the integrability conditions for (3) are satisfied. Since $[X_1, X_j] = 0$, these read as

$$L_{X_j} c_1 = L_{X_1} c_j. \tag{28}$$

Now, by Theorem 2.1, the equation

$$L_{X_1}\alpha = c_1 \tag{29}$$

has an analytic solution $\bar{\alpha}(x)$. Therefore $L_{X_1}(L_{X_i}\bar{\alpha} - c_i) = L_{X_i}L_{X_1}\bar{\alpha} - L_{X_1}c_i = L_{X_i}c_1 - L_{X_1}c_i = 0$ (by (28)). So $L_{X_i}\bar{\alpha} - c_i$ is an analytic function that satisfies the equation

$$L_{X_1}\gamma = 0. \tag{30}$$

However, the only analytic solution to (30) is γ =constant, as can be seen by setting up a formal power series solution to the PDE $L_{X_1}\gamma = 0$. Since $c_i(0) = 0$ and $X_i(0) = 0$ (for all $i = 2, \ldots, d$), it follows that $L_{X_i}\bar{\alpha} - c_i = 0$ identically for all $i = 2, \ldots, d$. Thus $\bar{\alpha}$ satisfies the remaining equations in (3), also.

We can see from the proof above that the condition $[X_1, X_i] = 0$ is required only to ensure that the solution to the equation $L_{X_1}\alpha = c_1$ satisfies the remaining equations, also. This condition can be weakened. However, doing so may violate the conditions necessary for solving the equation $L_{X_1}\alpha = c_1$.

DEFINITION 5.1. *For an analytic vector field* $X = \sum_{i=1}^{n} a_i(x)(\partial/\partial x_i)$ *on* $\mathbb{R}^n$, *vanishing at the origin, the linear part of* X *is the vector field* $Y := \sum_{i=1}^{n} b_i(x)(\partial/\partial x_i)$, *where* $b_i(x)$ *is the first-order term in the Taylor expansion of* a_i *about the origin.*

Remark 5.1. One common Poincaré vector field is described below.

A vector field X of the type (6) with $\gamma_j = 1$ for all j is said to be an LS vector field, because the flow of the linear part of such a vector field generates the scaling action on $\mathbb{R}^n$. More precisely, the scaling action is the action given in coordinates by

$$\phi[\lambda, (x_1, \ldots, x_n)] = (\lambda x_1, \ldots, \lambda x_n),$$

where λ is a nonzero real number. Here $\phi(,)$ denotes the action of the scaling group. More importantly, for our purposes, it is of great significance that to be an LS vector field is a property independent of the coordinate system. Indeed, if $X = \sum_{i=1}^{n} a_i(x)(\partial/\partial x_i)$, then its linear part is determined by its Jacobian matrix at 0, J. If $J = ((\partial a_i/\partial xj)(0)) = (\gamma_{ij})$, then $Y = \sum_{i=1}^{n}\sum_{j=1}^{n}(\gamma_{ij})x_j(\partial/\partial x_i)$. Now, for an LS vector field, $J = I_n$, the identity matrix, in every coordinate system. This is because the Jacobian matrix is transformed to a conjugate matrix under a coordinate change. However, I_n is the only matrix in its conjugacy class.

DEFINITION 5.2. *Let* Δ *be an analytic distribution on an open set* $U \subseteq \mathbb{R}^n$ *containing the origin. Let* Δ *satisfy the hypotheses of Theorem* 3.1 *with* $d_0 = 0$. *If* $\Delta = \mathrm{sp}\{X_i\}$, $i = 1, \ldots, d$ *for vector fields* X_i *analytic on* U, *then its first-order distribution is the distribution* $\tilde{\Delta}$, *defined on* U *as* $\tilde{\Delta} := \mathrm{sp}_{\mathbb{R}}\{Y_i\}$, *where the* Y_i *are as in Definition* 5.1. *Note that this definition does not depend on the choice of either a coordinate system or a basis for* Δ. Δ *is said to be nonsingular of the first order if* $\tilde{\Delta}$ *has maximal rank* d.

Remark 5.2. It is not necessary that, if Δ is nonsingular away from the origin, then the Y_i's are linearly independent away from the origin. As we see later, it is adequate for our purposes if they are linearly independent at any one point in U away from the origin.

DEFINITION 5.3. *With the notation of Definition* 5.2, Δ *is said to be abelian of first order if* $[Y_1, Y_j] = 0$ *for all* $j = 1, \ldots, d$, *where* Y_1 *is the linear part of the Poincaré vector field* X_1. *Once again, this is a property independent of the choice of either a coordinate system or any choice of basis for* Δ *containing* X_1.

Note that, if Δ contains in it an LS vector field, then Δ is automatically abelian of the first order. This is because, if X_1 is an LS vector field, then $Y_1 = \sum_{i=1}^{n} x_i(\partial/\partial x_i)$ and any linear vector field commutes with Y_1. Indeed, every matrix commutes with the identity matrix.

THEOREM 5.2. *Let* Δ *be a distribution satisfying the hypothesis of Theorem* 3.1 *with* $d_0 = 0$. *Suppose that* Δ *contains a vector field, say* $\tilde{X}_1$, *whose Jacobian at zero* J *is diagonalizable and has eigenvalues* γ_j *satisfying the estimate described in Theorem* 2.1. *Furthermore, assume that* Δ *is nonsingular of the first order and that* Δ *is abelian of the first order. Then we can find a basis for* Δ, $\{X_1, X_2, \ldots, X_d\}$ *with* X_1 *a Poincaré vector field, and* $[X_1, X_k] = 0$, $k = 1, \ldots, d$.

Proof. Since J is diagonalizable, we can, by linear change of coordinates, transform $\tilde{X}_1$ to X_1 with X_1 a Poincaré vector field. If the original basis of Δ was $\{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d\}$, then, under this coordinate change, $\tilde{X}_2, \ldots, \tilde{X}_d$ transform to some vector fields $\hat{X}_2, \ldots, \hat{X}_d$.

We now make a basis change of the form

$$T = (\beta_{ij}), \tag{31}$$

where the β_{ij}'s are analytic functions that satisfy $\beta_{ii}(0) = C_i$ for nonzero constants $C_i (i \geq 2)$ and $\beta_{ij}(0) = 0$ for $j > 1$, and $j \neq i$, $i = 2, \ldots, d$. The $\beta_{i1}(0)$ can be arbitrary. Note that T is invertible at 0 and hence in a neighborhood. We want the new basis $(X_1, X_2, \ldots, X_d) := T(X_1, \hat{X}_2, \ldots, \hat{X}_d)$ to satisfy $[X_1, X_i] = 0$, $i \geq 2$. Since Δ is involutive, we have

$$[X_1, \hat{X}_i] = \sum_{j=1}^{d} \alpha_{ij} \hat{X}_j, \quad \text{with } \hat{X}_1 = X_1, \tag{32}$$

and the α_{ij}'s are some analytic functions. If $[X_1, X_i]$ is to be zero, we must have

$$\sum_{j=1}^{d} (L_{x_1} \beta_{ij}) \hat{X}_j + \sum_{j=1}^{d} \beta_{ij} [X_1, \hat{X}_j] = 0. \tag{33}$$

Equation (33) shows that the PDEs for the rows of T form separate systems (i.e., we can solve each separately).

We illustrate the solvability of (33) by doing so for the second row of T. The argument for the other rows is exactly the same. The equations for the second row read as

$$\begin{aligned} L_{X_1} \beta_{21} &= -\beta_{22}\alpha_{21} - \cdots - \beta_{2d}\alpha_{d1}, \\ L_{X_1} \beta_{22} &= -\beta_{22}\alpha_{22} - \cdots - \beta_{2d}\alpha_{d2}, \\ &\vdots \\ L_{X_1} \beta_{2d} &= -\beta_{22}\alpha_{2d} - \cdots - \beta_{2d}\alpha_{dd}. \end{aligned} \tag{34}$$

We use Theorem 2.2 to prove that the last $(d-1)$ equations in (34) have analytic solutions. Then we plug these solutions into the first equation and use Theorem 2.1 to show that it has a solution. To that end, write the last $(d-1)$ equations as $L_X v = s(x, v)$, for a $(d-1)$ vector v. Clearly, $s(0,0) = 0$.

Futhermore, $(\partial s/\partial v)(0,0) = (-\alpha_{ij}(0))$. We must show that this is zero. To that end, let us proceed as follows: Since the $\hat{X}_i$ vanish at the origin and so does X_1, we have

$$\text{linear part of } [X_1, \hat{X}_i] = [Y_1, \hat{Y}_i] = 0 \quad \text{(for all x)}$$

(since Δ is abelian of the first order). Consequently, the linear part of $\sum_{j=1}^{d} \alpha_{ij} \hat{X}_j$ is also zero for all $x \in U$. Bearing in mind Remark 5.1, let us compute the Jacobian at zero of $\sum_{j=1}^{d} \alpha_{ij} \hat{X}_j$. This is equal to $\sum_{j=1}^{d} \alpha_{ij}(0) J_j$, where J_j is the Jacobian at zero of $\hat{X}_j$(as the $\hat{X}_j$ vanish at the origin). Since Δ is nonsingular of the the first order, this forces $\alpha_{ij}(0) = 0$ for all (i, j). Thus $(\partial s/\partial v)(0,0) = 0$. Consequently, Theorem

2.2 applies to provide analytic solutions whose values at zero may (by Remark 2.1(a)) be chosen as we please. If we plug these solutions into the first equation of (34), we get an equation of the type (5). Clearly, the right-hand side of this equation vanishes at $x^{\neg} = 0$. Consequently, Theorem 2.1 applies to yield a solution.

To give the reader a flavor of the kind of solutions that arise, we set up formal solutions to (34). For clarity of exposition, we limit ourselves to the case where $d = 3$ and the case where X_1 is an LS vector field. By Poincaré's linearization theorem [27], there exists a coordinate change preserving the origin so that X_1 transforms to the scaling vector field. Now the last two equations of (34) read as

$$L_{X_1}\beta_{22} = -\beta_{22}\alpha_{22} - \beta_{23}\alpha_{32}$$

and

$$L_{X_1}\beta_{23} = -\beta_{22}\alpha_{23} - \beta_{23}\alpha_{33}.$$

We seek solutions of the form

$$\beta_{22} = w = c_2 + \sum_{|m|\geq 1} w_m x^m,$$
$$\beta_{23} = v = \sum_{|m|\geq 1} v_m x^m,$$

where m stands for a multi-index. Clearly,

$$L_{X_1} w = \sum_{|m|\geq 1} |\, m \,|\, w_m x^m,$$
$$L_{X_1} v = \sum_{|m|\geq 1} |\, m \,|\, v_m x^m.$$

Let us also define

$$-\alpha_{22} = \alpha = \sum_{|m|\geq 1} \alpha_m x^m,$$
$$-\alpha_{32} = \beta = \sum_{|m|\geq 1} \beta_m x^m,$$
$$-\alpha_{23} = \gamma = \sum_{|m|\geq 1} \gamma_m x^m,$$
$$-\alpha_{33} = \delta = \sum_{|m|\geq 1} \delta_m x^m.$$

These functions are known analytic functions (since the distribution Δ is analytic). Consequently, the coefficients $\alpha_m, \beta_m, \gamma_m, \delta_m$ are predetermined. We know that the the nonsingularity and the abelianness in the first order of Δ entail that these functions vanish at the origin. Equating like powers of x, we have

$$w_m = c_2 \alpha_m, \tag{35}$$
$$v_m = c_2 \gamma_m, \tag{36}$$

for $\mid m \mid = 1$ and

$$w_m = \sum_l \sum_k \frac{\alpha_l w_k + \beta_l v_k}{\mid m \mid},$$
$$v_m = \sum_l \sum_k \frac{\gamma_l w_k + \delta_l v_k}{\mid m \mid} \tag{37}$$

for $\mid m \mid > 1$ and $l + k = m$.

We claim that (37) can be solved recursively for w_m, v_m. For $\mid m \mid = 1$, (35) shows that we can begin the induction. Let us assume the claim is true for $\mid m \mid = p$. Let $\mid m \mid = p + 1$ and let $m = (m_1, \ldots, m_n)$ be a multi-index with weight $p + 1$. Then

$$w_m = \sum_l \sum_k \frac{\alpha_l w_k + \beta_l v_k}{p+1}, \qquad l + k = m$$

and

$$v_m = \sum_l \sum_k \frac{\gamma_l w_k + \delta_l v_k}{p+1}, \qquad l + k = m.$$

Since $\alpha(0) = \beta(0) = \gamma(0) = \delta(0) = 0$, the u_k, v_k terms on the right-hand side of the last two equations all have weight $\leq p$ and hence have already been determined by the inductive hypothesis. Since the $\alpha_l, \beta_l, \gamma_l, \delta_l$ are known for all l, we are done. The procedure outlined above goes through verbatim when $d > 3$ (the only difference being an exponential increase in indices).

We now give sufficient conditions to ensure that $c_j^i(0) = 0$. Observe that, if Δ is a distribution with rank zero at the origin that contains even one vector field X_i whose Jacobian at the origin is a nonsingular matrix (if X_i has a hyperbolic equilibrium point at the origin, for instance) and is input-insensitive, then necessarily the control vector fields vanish at the origin. Indeed, we must have:

$$[g_j, X_i](0) \in \Delta(0) = 0.$$

Since each of the X_i vanish at the origin, we have that

$$[g_j, X_i](0) \ = \ -DX_i(0)g_j(0).$$

By hypothesis, the Jacobian of X_i at the origin, $DX_i(0)$, is nonsingular. This yields the conclusion and motivates the following definitions.

DEFINITION 5.4. *Assume the following:*

(i) $G(0) = 0$, *and denote by* $\tilde{g}$ *the linear parts of* g;

(ii) *The distribution* $\tilde{G} = \operatorname{span}_{\mathbf{R}}\{\tilde{g}\}$ *has maximal rank* m;

(iii) $\tilde{G} \cap \tilde{\Delta} = \{0\}$ *on a dense subset.*

We then say that the pair (G, Δ) *is nonsingular of the first order.*

Remark 5.3. If (G, Δ) is nonsingular of the first order $f(0) = 0$ and the linear part of f leaves $\tilde{\Delta}$ invariant, then $c_j^i(0) = 0$. Indeed, if in the equation $[f, X_i] = V_i + \Sigma c_j^i g_j$ we pass to linear parts, then we have that $\sum_{j=1}^m c_j^i(0)\tilde{g}_j = 0$ on a dense subset and so everywhere. Since $\tilde{G}$ has maximal rank m, there is at least some point where all the vectors obtained by evaluating the basis of $\tilde{G}$ are linearly independent (in fact, by analyticity most points are points where this rank condition is valid). So we have $c_j^i(0) = 0$.

Combining Theorems 5.1 and 5.2 and Remark 5.3, we have the following result.

THEOREM 5.3. *Assume the following*:

(a) Δ *satisfies assumptions* (A)–(E) *of* §1, *the hypotheses of Theorem* 3.1 *with* $d_0 = 0$, *and contains a Poincaré vector field*;

(b) Δ *is nonsingular of the first order*;

(c) Δ *is abelian of the first order*;

(d) *The pair* (G, Δ) *is nonsingular of the first order*;

(e) $f(0) = 0$, *and the linearization of* f *leaves* $\tilde{\Delta}$ *invariant.*

Then we can find an analytic feedback in a neighborhood of the origin that preserves the equilibrium of f *and renders* Δ *controlled invariant.*

Remark 5.4. (a) If the vector field in Δ is also an LS vector field, then (c) automatically holds.

(b) Results of a nature similar to Theorem 5.3 appear in [18], [25]. Their results are concerned with a finite-dimensional Lie algebra of analytic vector fields. However, the set of vector fields taking values in an involutive distribution, although a Lie algebra, is rarely finite-dimensional. This is, perhaps, why they could get by with coordinate changes alone.

Example. Consider in R^4 the involutive distribution $\operatorname{span}\{X_1, X_2\}$, where

$$X_1(x) = x_1 \frac{\partial}{\partial x_1} + x_2 \frac{\partial}{\partial x_2} + 2x_3 \frac{\partial}{\partial x_3} + x_4 \frac{\partial}{\partial x_4}$$

and

$$X_2(x) = x_2 \frac{\partial}{\partial x_1} - x_1 \frac{\partial}{\partial x_2} + x_3 \frac{\partial}{\partial x_3} + x_4 \frac{\partial}{\partial x_4}.$$

In fact, $[X_1, X_2] = 0$. Note, however, that Δ has singularities.

Let the control vector field g be given as

$$g(x) = x_3 \frac{\partial}{\partial x_3}.$$

Then Δ is input insensitive. In fact, $[g, X_i] = 0$, for $i = 1, 2$. In addition, $G \cap \Delta = 0$ on a dense subset $(x_1, x_2, x_4 \neq 0)$, but not on all of R^4. Also, $\tilde{\Delta}$ is nonsingular of the first order. Note also that the linear part of g spans a distribution of maximal rank 1 and that it intersects $\tilde{\Delta}$ in the zero section on a dense subset of R^4.

Finally, let the drift be given by

$$f(x) = (x_1 - x_2 + x_1^2 + x_2^2)\frac{\partial}{\partial x_1} + (x_1 + x_2)\frac{\partial}{\partial x_2} + (x_1^2 + x_2^2)x_3^2 \frac{\partial}{\partial x_3} + (x_1 x_4 + x_2 x_4 + x_4)\frac{\partial}{\partial x_4}.$$

We can also see that the linear part of f leaves $\tilde{\Delta}$ invariant. It is easy to calculate that

$$[f, X_1] = (x_1^2 + x_2^2)\frac{\partial}{\partial x_1} + 4(x_1^2 + x_2^2)x_3^2 \frac{\partial}{\partial x_3},$$

so that $c^1(x) = -2x_1 - x_2 + 4(x_1^2 + x_2^2)x_3$ and

$$[f, X_2] = (x_1^2 + x_2^2)x_3 \frac{\partial}{\partial x_2} + (x_1^2 + x_2^2)x_3 \frac{\partial}{\partial x_3},$$

and hence $c^2 = -2x_2 + x_1 + (x_1^2 + x_2^2)x_3$. It is easy to check that $X_2 c^1 = X_1 c^2$, so that the integrability conditions are satisfied, as they must be for theoretical reasons. To

find a solution to this system of PDEs, we first solve the equation corresponding to the Poincaré vector field X_1, this being easier of the two, since the solution is given by $\alpha_m = c_m^1/P_0(m)$, where, as usual, α_m and c_m^1 are the coefficients of the Taylor expansions of α and c^1. Thus the solution α is

$$\alpha(x) = (x_1^2 + x_2^2)x_3 - 2x_1 - x_2.$$

This α also solves $L_{X_2}\alpha = c^2$, as it should.

Remark 5.5. One might wonder whether one would ever encounter, in practice, a distribution that contains the scaling vector field. Let us consider the disturbance decoupling problem, for instance. If the output of the system (or the functions one has obtained at the instance of the first breakdown of the controlled invariant subdistribution algorithm [15], [23]) is an analytic function of the ratio of the states and if, furthermore, the disturbance vector field also vanishes at the origin, then we have precisely such a distribution. Such an output is discontinuous at the origin, but many practical measurement schemes, such as camera measurements, answer to such a description.

6. The "d_0 positive" case. Let us first suppose that $d_0 = d - 1$. Then, if the vector field X_d is a Poincaré vector field in $\mathbb{R}^{n-d_0}$, we can, by Theorem 2.3 and Proposition 4.1, find a solution to (3), so long as a formal solution to the equation

$$L_{X_d}\alpha = c_d(x) \tag{38}$$

exists. In general, (38) is an equation with parameters. Now it is easy to see that a formal solution to (38) exists if the Taylor expansion of c_d contains no first-order terms in the variables $x_1, x_2, \ldots, x_{d-1}$. While this condition can always be checked once c_d is known, there seems to be no clear-cut way to characterize this condition in terms of the vector fields f, g_i and the distribution Δ.

If $0 < d_0 < d - 1$, then we must, in general, solve a degenerate system with parameters. As of now, we do not know much about this situation, even if one of the degenerate vector fields is an LS vector field in $\mathbb{R}^{n-d_0}$. The argument presented in Theorem 5.1 would fail because now any analytic function depending only on $x_1, \ldots, x_{d_0}$ would be invariant under a Poincaré vector field. However, if it should turn out that the functions $c_{d_0+1}, \ldots, c_d$ do not depend on $x_1, \ldots, x_{d_0}$ (i.e., the parameters disappear) then we are back in the situation of Theorem 5.1, and we will be done. Put in a different way, we are demanding that the functions $c_{d_0+1}, \ldots, c_d$ be invariant under the group of translations in the variables $x_1, \ldots, x_{d_0}$. Sufficient conditions for these functions can be found in [24]. In our situation, it would suffice if, for instance, $[f, X_k]$ for $k = 1, \ldots, d_0$ were to lie in Δ already. We now discuss an example that illustrates this issue. This example is similar to the one in the previous section, except that the involutive distribution in question has nonzero dimension at the origin.

Example. Consider in R^5 the involutive distribution $\operatorname{span}\{X_1, X_2, X_3\}$, where

$$X_1(x) = \frac{\partial}{\partial x_1},$$

$$X_2(x) = x_1\frac{\partial}{\partial x_1} + x_2\frac{\partial}{\partial x_2} + x_3\frac{\partial}{\partial x_3} + 2x_4\frac{\partial}{\partial x_4} + x_5\frac{\partial}{\partial x_5},$$

and

$$X_3(x) = x_3\frac{\partial}{\partial x_2} - x_2\frac{\partial}{\partial x_3} + x_4\frac{\partial}{\partial x_4} + x_5\frac{\partial}{\partial x_5}.$$

To arrive at a basis of the form in Theorem 3.1, we just take X_1, X_3 as they are and modify X_2 to the field $x_2\partial/\partial x_2 + x_3\partial/\partial x_3 + 2x_4\partial/\partial x_5 + x_5\partial/\partial x_5$, which, for convenience, we label once more as X_2.

Note that $\dim\Delta(0) = 1$ and that its maximal rank is 3.

Let the control vector field g be given as

$$g(x) = \frac{\partial}{\partial x_1} + x_4\frac{\partial}{\partial x_4}.$$

Then Δ is input insensitive. In fact, $[g, X_i] = 0$, for $i = 1, 2$. In addition, $G \cap \Delta = 0$ on a dense subset but not on all of R^5. Note also that the linear part of g spans a distribution of maximal rank 1. Finally, let the drift be given by

$$f(x) = (x_2^2 + x_5^2 + x_1)\frac{\partial}{\partial x_1} + [x_1^2(x_2 - x_3) + x_2^2 + x_3^2]\frac{\partial}{\partial x_2} + [x_1^2(x_3 + x_2)]\frac{\partial}{\partial x_3}$$
$$+[x_1^2 + (x_2^2 + x_3^2)x_4]x_4\frac{\partial}{\partial x_4} + (x_2 + x_3)x_5\frac{\partial}{\partial x_5}.$$

It is easy to calculate that

$$[f, X_1] \in \Delta$$

so that $c^1(x) = 0$. We can also find $c^2(x) = -2x_2 - x_3 + 4(x_2^2 + x_2^3)x_4$ and $c^3(x) = -2x_3 + x_2 + (x_2^2 + x_2^3)x_4$. Note that neither c^2 nor c^3 depends on the variable x_1. This is a consequence of the translational symmetry in the variable x_1, which, in turn, follows because $[f, X_1] \in \Delta$. Once again, the integrability conditions for the PDEs are satisfied. Finally, a solution to the system is obtained by first solving the easier of the two equations, namely,

$$L_{X_2}\alpha(x) = c^2.$$

One solution is

$$\alpha(x) = (x_2^2 + x_3^2)x_4 - 2x_2 - x_3.$$

We can check easily that this solves the second equation $L_{X_3}\alpha(x) = c^3$, also. Since the solution does not involve x_1 at all, it also solves the equation $L_{X_1}\alpha = 0$.

7. The noninput insensitive case. If we are dealing with a single-input control system, then the theory for input-insensitivization parallels that of the PDEs (3). Indeed, in this case, (2) reads as

$$L_{X_i}(\beta) = b^i\beta, \tag{39}$$

where b^i is determined by $[g, X_i] = b^i g \ (\mathrm{mod}\Delta)$. Note that (39) can be written in a form analogous to (3), to wit,

$$L_{X_i}(\gamma) = b^i \tag{40}$$

with $\gamma = \ell n\beta$. The integrability conditions for (40) are $L_{X_i}b^j = L_{X_j}b^i$, which are (as they should be) the integrability conditions for (39). Consequently, Theorems 4.1, 5.1, and 5.3 with the obvious modifications carry over to this case.

If on the other hand there are many inputs, it seems difficult (at present, at least!) to generalize Theorems 4.1, 5.1, and so forth. The crux of the problem is that we do

not know what conditions, in addition to the integrability conditions, are needed for the local solvability of an overdetermined system of degenerate PDEs. At any rate, the proof technique of Theorems 4.1 and 5.1 would not carry over. If, on the other hand, we have $d = 1$ (and hence $d_0 = 0$) (thus the system is not overdetermined), then Theorem 2.2 would apply to produce an analytic feedback under suitable assumptions. If the pair (G, Δ) is nonsingular of the first order, then we can show, just as in Theorem 5.2, that (in the terminology of Theorem 2.2) $(\partial s/\partial v)(0,0) = 0$. Clearly, $s(0,0) = 0$. So Theorem 2.2 gives an analytic feedback, and by Remark (2.1(a)) we can arrange for this feedback to be invertible. On the other hand, if we assume that $G + \Delta/\Delta$ is of maximal rank at 0, then by calculating $[g, X_1](0)$ (where X_1 is the Poincaré vector field), we can see that $(\partial s/\partial v)(0,0) = -I_p$. In the terminology of Remark 2.5(c), $\Lambda(m) = (|m| + 1)I_p$. If we choose C to be 1, say, then Remark 2.5(c) would yield an analytic feedback that would render Δ input insensitive. Of course, we cannot make this feedback invertible—in fact, it will vanish at the origin. In related ongoing work, we are attempting to extend solutions to (1)–(3) past the singularities of Δ by finding group invariant solutions to these PDEs. This work suggests that, for Δ with $d_0 = 0$, it is very natural that $G + \Delta/\Delta$ also picks up singularities at 0.

8. Conclusion. In this paper, we have addressed the problem of rendering a locally controlled invariant analytic distribution with singularities controlled invariant. We showed that a solution to this problem involves solving a system of degenerate first-order PDEs. We have provided some instances of when this can be done. It would be interesting to see what bearing the results described in this paper have on the stabilization problem, since it is known [15] that the zero dynamics algorithm is closely related to the controlled invariance problem. Of course, we must first deal with input-insensitivization. In general, it seems that a delicate balance of techniques from symmetry groups, degenerate PDEs, and the blowing-up construction will be needed to address the problem of controlled invariance for singular distributions in general.

Acknowledgments. The author thanks Dr. K. Grasse for reading versions of the manuscript and offering innumerable suggestions to improve its quality; Dr. W. Dayawansa, Dr. K. Grasse, and Dr. H. Schättler for their comments, especially in streamlining the proof of Theorem 3.1; the anonymous referees for making several very important suggestions toward enhancing the readability of the paper; and Dr. C. Byrnes and Dr. D. Elliott for their remarks. I thank Ms. P. Johnson and Ms. C. Presswood for their clinical efficiency and infinite patience during their help in the typesetting process.

Note added in proof. Since the submission of this paper, we have been successful in obtaining further results on the subject matter of §§6 and 7. The details will be reported elsewhere.

REFERENCES

[1] R. Abraham and J. Marsden, *Foundations of Mechanics*, 2nd ed., Addison-Wesley, Reading, MA, 1985.

[2] D. V. Anosov and V. I. Arnold, eds., *Encyclopaedia of Mathematical Sciences*, Vol. 1, Springer-Verlag, New York, Berlin, 1988.

[3] G. Bengel and R. Gérard, *Formal and convergent solutions to singular partial differential equations*, Manuscripta Math., 38 (1982), pp. 343–373.

[4] G. Bengel, *Convergence of Formal Solutions of Singular Partial Differential Equations*, in Advances in Microlocal Analysis, H. Garnir, ed., D. Reidel, Boston, MA, 1986, pp. 1–14.

[5] W. BOOTHBY, *An Introduction To Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
[6] C. I. BYRNES, *Towards a global theory of (f,g) invariant distributions with singularities*, in Proc. of the 1983 Mathematical Theory of Networks and Systems, Beer Sheva, Israel, P. Fuhrmann, ed., pp. 149–165.
[7] ———, *Feedback decoupling of rotational disturbances for spherically constrained systems*, in Proc. of the 23rd CDC, December 1984, Las Vegas, NV, pp. 421–426.
[8] D. CHENG AND T. TARN, *New result on (f,g) invariance*, System Control Lett., 12 (1989), pp. 319–326.
[9] W. P. DAYAWANSA, D. CHENG, W. BOOTHBY, AND T. TARN, *Global (f,g) invariance of nonlinear systems*, SIAM J. Control Optim., 26 (1988), pp. 1119–1132.
[10] K. GRASSE, *Controlled invariance for fully nonlinear systems*, Internat. J. Control., 56 (1992), pp. 1121–1137.
[11] J. GRIZZLE, *Controlled invariance for discrete time nonlinear systems with applications to the disturbance decoupling problem*, IEEE Trans. Automat. Control, 30 (1985), pp. 868–874.
[12] R. HERMANN, *The differential geometry of foliations*, J. Math. Mech., 11 (1962), pp. 302–316.
[13] H. HERMES, *Involutive subdistributions and canonical forms for distributions and nonlinear systems*, in Theory and Applications of Nonlinear Systems, C. Byrnes and A. Lindquist, eds., North–Holland, Amsterdam, 1986, pp. 125–133.
[14] R. M. HIRSCHORN, *(A,B) invariant distributions and disturbance decoupling of nonlinear systems*, SIAM J. Control Optim., 19 (1981), pp. 1–19.
[15] A. ISIDORI, *Nonlinear Control Systems*, 2nd ed., Springer-Verlag, New York, Berlin, 1989.
[16] A. ISIDORI, A. KRENER, C. GORI-GEORGI, AND S. MONACO, *Nonlinear Decoupling via feedback—A differential geometric approach*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 331–345.
[17] A. KRENER, (f,g) *invariant distributions, connections and Pontryagin classes*, in Proc. of the 20th CDC, San Diego, CA, 1981, pp. 1322–1325.
[18] E. S. LIVINGSTON AND D. ELLIOTT, *Linearization of families of vector fields*, J. Differential Equations, 55 (1984), pp. 289–299.
[19] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.
[20] H. NIJMEIJER, *Controlled invariance for affine control systems*, Internat. J. Control, 24 (1981), pp. 825–833.
[21] H. NIJMEIJER AND A. VAN DER SCHAFT, *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 904–914.
[22] ———, *Controlled invariance for nonlinear systems: Two worked examples*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 361–364.
[23] ———, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, Berlin, 1990.
[24] V. RAMAKRISHNA AND H. SCHÄTTLER, *Controlled invariant distributions and group invariance*, J. Math. Systems Estimation Control, 1 (1991), pp. 209–240.
[25] J. L. SEDWICK AND D. ELLIOTT, *Linearization of Analytic Vector Fields in the Transitive Case*, J. Differential Equations, 25 (1977), pp. 377–390.
[26] P. STEFAN, *Accessible sets, orbits and foliations with singularities*, Proc. London Math. Soc., 29 (1974), pp. 699–713.
[27] S. STERNBERG, *Local contractions and a theorem of Poincaré*, Amer. J. Math., 79 (1957), pp. 809–823.
[28] H. SUSSMANN, *Orbits of families of vector fields and the integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.
[29] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, Berlin, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 808–830, May 1994

010

DIFFERENTIAL GAMES WITH INFORMATION LAGS*

XIAOJUN QIAN†

Abstract. Differential games of generalized pursuit and evasion are studied. The definitions of strategy and payoff follow those of Berkovitz. It is shown under appropriate hypotheses on the data of the problem that if the Isaacs condition holds, then there exists a saddle point.

Then differential games with information lags are studied, in which Berkovitz's definitions of strategy and payoff are generalized to games with lags. It is first shown through an example that if a game has a lag, then value of the game does not exist in general. For games of fixed duration with information lags, it is demonstrated that if the Isaacs condition holds, then as the lags tend to zero, the upper and lower values as functions of the lags will tend to the value in the game with no lags. The same results hold also for differential games of generalized pursuit and evasion and for games of survival if certain reasonable conditions on the data and the structure of the terminal set hold.

Key words. differential games, saddle points, information lags, upper and lower values

AMS subject classifications. 90D25, 90D26

Formulated intuitively, a differential game has its state $x(t) \in \mathbb{R}^n$ at time t determined by a system of differential equations

$$\begin{aligned} &\frac{dx}{dt} = f(t, x, y(t), z(t)) \qquad \text{for } t_0 < t \le T \\ &x(t_0) = x_0, \end{aligned} \tag{1}$$

where $y(t)$ is chosen from some preassigned set $Y \subset \mathbb{R}^r$ by Player I at each instant of time t and $z(t)$ is chosen from some preassigned set $Z \subset \mathbb{R}^s$ by Player II at each instant of time t. The payoff is

$$g(t_f, x_f) + \int_{t_0}^{t_f} f^0(s, \phi(s), y(s), z(s))ds, \tag{2}$$

where ϕ is the solution of (1), the terminal time t_f is the first time that the trajectory $(t, \phi(t))$ reaches some preassigned terminal set $\mathcal{T}$, $x_f = \phi(t_f)$, g is a function defined on $[t_0, T] \times \mathbb{R}^n$, and f^0 is a function defined on $[t_0, T] \times \mathbb{R}^n \times Y \times Z$. Player I wishes to choose y so as to maximize the payoff while Player II wishes to choose z so as to minimize it. Such a game is called a game of survival. If $g \equiv 0$, we call the game a game of generalized pursuit and evasion. If the terminal set $\mathcal{T}$ is given by $[T, \infty) \times \mathbb{R}^n$, we call the game a game of fixed duration.

In his series of papers [1]–[3], Berkovitz studied the three types of differential games, using a definition of strategy that is an adaptation of that of Friedman [5] and Karlin [7] and a definition of payoff that is an adaptation of that of Krasovskii and Subbotin [8]. Berkovitz showed that if the Isaacs condition holds and the data satisfy reasonable hypotheses, then the three types of differential games have values that are continuous functions of the initial time and state. He showed in [1] that under these

*Received by the editors July 22, 1991; accepted for publication (in revised form) November 13, 1992. This research was supported by a David Ross grant from Purdue University.

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

hypotheses, games of fixed duration have saddle points, but he did not obtain the existence of a saddle point in the other games.

In §1 of our paper we present a result that is of interest in its own right and that we use in our study of problems with lags, namely, that if the Isaacs condition holds, then games of generalized pursuit and evasion have saddle points.

Friedman showed in [6] that, among other assumptions, if the compact sets Y and Z are convex, if $f^0 \geq 0$, and if f and f^0 are linear in y and z, that is,

$$f(t,x,y,z) = f_0(t,x) + F_1(t,x)y + F_2(t,x)z,$$
$$f^0(t,x,y,z) = f_0^0(t,x) + F_1^0(t,x)y + F_2^0(t,x)z,$$

where $F_1(t,x) \in \mathbb{R}^{n\times r}$, $F_2(t,x) \in \mathbb{R}^{n\times s}$, $F_1^0(t,x) \in \mathbb{R}^{1\times r}$, and $F_2^0(t,x) \in \mathbb{R}^{1\times s}$, then under his definitions, a saddle point exists for games of generalized pursuit and evasion. Friedman also showed that if $f^0(t,x,y,z) = f^0(t,x)$ and if the set $f(t,x,Y,Z) = \{f(t,x,y,z) | y \in Y, z \in Z\}$ is convex for any $(t,x) \in [t_0,T] \times \mathbb{R}^n$, then a generalized saddle point exists. In our result, we do not require the convexity of the sets Y and Z, the linearity of f and f^0 in y and z, nor the conditions that f^0 is independent of y and z and that the set $f(t,x,Y,Z)$ is convex.

In §2 of this paper, we extend Berkovitz's definition of differential games to games with information lags. First, we show through an example that if a differential game of fixed duration has a lag, then the value of the game does not exist in general. In [1]–[3], Berkovitz assumed that at each instant time t, each player has complete knowledge of his own action and that of his opponent from the initial time t_0 of the game up to but not including the current time t. A more realistic model of most conflict situations is that both players need a certain period of time to process information about their opponent's actions, i.e., at time t, Player I knows Player II's history between t_0 and $t-\lambda$ while Player II knows Player I's history between t_0 and $t-\mu$. The real numbers λ, μ, $0 \leq \lambda, \mu \leq t-t_0$, are called the lags of Player I and Player II, respectively. It will be shown that if the Isaacs condition holds, then as the lags tend to zero, the upper and lower values as functions of the lags will tend to the value in the game with no lags. Although such results are to be expected on intuitive grounds, and should hold if the model is realistic, the mathematical arguments needed to establish these results are delicate. Finally, we will show that if certain reasonable conditions on the data and the structure of the terminal set hold, then the results obtained for the fixed-time game with lags also hold for games of pursuit and evasion and games of survival with lags.

Friedman [5, p. 268] defined differential games with only one player having a lag. If Player I has the lag, Friedman defined the value of game as the limit of lower δ-values. In games without lag, he calls this limit the lower value of the game. If Player II has the lag, he defined the value as the limit of upper δ-values. In games without lag he calls this limit the upper value of the game. It is immediate from his definitions that the lower or upper δ-values of the game are monotonic in δ and so converge. Hence value for games with lags always exists in Friedman's sense. It is our belief that Friedman's definition is unsatisfactory in that his values are really upper or lower values, as the case may be. It was shown in [5] that if the opposing control variables appear "separated" in both the differential system and the payoff, then the values of the games with lag tend to the value of the game without lag. Recall, however, that Friedman's values for games with lag are really upper or lower values, so that our example and his results are not contradictory. We think that our definitions of

strategy and value are more natural than Friedman's; moreover, our result does not require a separation assumption on the dynamics of the game.

1. The existence of saddle point in games of generalized pursuit and evasion. The game that we study in this section is governed by (1) and has the following payoff:

$$\int_{t_0}^{t_f} f^0(s, \phi(s), u(s), v(s))\, ds, \tag{3}$$

where ϕ is the solution of (1) and t_f is the first time that $(t, \phi(t))$ reaches some preassigned terminal set $\mathcal{T}$. We use the notation of strategy and payoff introduced in [1] and [2] and establish the existence of saddle points in games of generalized pursuit and evasion.

We denote the game with initial point (t_0, x_0) by $G(t_0, x_0)$. Let $\hat{x} = (x^0, x)$ where $x^0 \in \mathbb{R}$ and let $\hat{f} = (f^0, f)$. We make the following assumptions concerning $\hat{f}$.

ASSUMPTION I. (i) *The function $\hat{f}$ is continuous on $D = [t_0, T] \times \mathbb{R}^{n+1} \times Y \times Z$.*

(ii) *There exists a function k that is integrable on $[t_0, T]$ such that $\langle \hat{x}, \hat{f}(t, x, y, z)\rangle \leq k(t)(1 + |\hat{x}|^2)$ for all $(t, \hat{x}, y, z)$ in D.*

(iii) *For any $R > 0$, there exists a constant $K_R > 0$ such that for all t in $[t_0, T]$, y in Y, z in Z, and $|x| \leq R$, $|\bar{x}| \leq R$, $|\hat{f}(t, x, y, z) - \hat{f}(t, \bar{x}, y, z)| \leq K_R|x - \bar{x}|$.*

ASSUMPTION II. *Let F_1 be a closed domain in $(t_0, \infty) \times \mathbb{R}^n$ with $C^{(2)}$ boundary ∂F_1. Let F denote the intersection of F_1 with the slab $t_0 \leq t \leq T$, and let F be bounded. Let the terminal set $\mathcal{T}$ be given by $\mathcal{T} = F \cup ([T, \infty) \times \mathbb{R}^n)$. Let $\hat{\partial} F = (\partial F) \cap (\partial F_1)$. At each point $(t, x) \in \hat{\partial} F$, let*

$$\nu^0 + \langle \nu, f(t, x, y, z)\rangle < 0 \tag{4}$$

for all $y \in Y$ and all $z \in Z$, where (ν^0, ν) is the normal to ∂F_1 at (t, x) pointing to the exterior of F_1.

ASSUMPTION II′. *Let F_1 be as in Assumption* II *and let F_1 further satisfy the condition $F_1 \supset ([T, \infty) \times \mathbb{R}^n)$. Let $F = F_1$ and let $\mathcal{T} = F$, and let* (4) *hold at each $(t, x) \in \partial F$.*

For $\tau_0 \in [t_0, T]$ and $\hat{\xi}_0 \in \mathbb{R}^{n+1}$, we consider game $\hat{G}(\tau_0, \hat{\xi}_0)$ governed by

$$\frac{dx}{dt} = f(t, x, u(t), v(t)), \qquad x(\tau_0) = \xi_0$$

with payoff $\xi_0^0 + \int_{\tau_0}^{t_f} f^0(s, \phi(s), u(s), v(s))\, ds$ and terminal set $\hat{\mathcal{T}} = \mathbb{R} \times \mathcal{T}$.

If we let $W^{\pm}(t_0, x_0)$ be the upper and lower values of $G(t_0, x_0)$ and $\hat{W}^{\pm}(t_0, \hat{x}_0)$ be the upper and lower values of $\hat{G}(t_0, \hat{x}_0)$, then $\hat{W}^{\pm}(t_0, \hat{x}_0) = W^{\pm}(t_0, x_0) + x_0^0$.

Let $\hat{F} = \mathbb{R} \times F$ and let (ν^0, ν) be the exterior unit normal to ∂F, then $(\nu^0, 0, \nu)$ is the exterior unit normal to $\partial \hat{F}$. Let $\hat{\nu} = (0, \nu)$. Since $\hat{f} = (f^0, f)$, Lemma 1 follows from (4).

LEMMA 1. *For every $(t, \hat{x}) \in \partial \hat{F}$, $\nu^0 + \langle \hat{\nu}, \hat{f}(t, \hat{x}, y, z)\rangle < 0$ for all $y \in Y$ and all $z \in Z$.*

For a point $(t, \hat{x})$ let $\hat{\rho}(t, \hat{x})$ denote the signed distance of $(t, \hat{x})$ to $\partial \hat{F}$ with negative values assigned to points interior to $\hat{F}$. Let $(\partial \hat{F})_\epsilon = \{(t, \hat{x}) : |\hat{\rho}(t, \hat{x})| < \epsilon\}$. Then for

each $R > 0$, there exists an $\epsilon_0 > 0$ such that $\hat{\rho}$ is C^1 in $(\partial\hat{F})_\epsilon \cap \{(t,\hat{x}) : |\hat{x}| < R\}$, and if $(\nu^0, \hat{\nu})$ is the unit normal to $\partial\hat{F}$ at $(\tau,\hat{\xi})$ pointing to the exterior of $\hat{F}$, then

$$\lim_{(t,\hat{x})\to(\tau,\hat{\xi})} (\hat{\rho}_t(t,\hat{x}), \hat{\rho}_{\hat{x}}(t,\hat{x})) = (\nu^0,\hat{\nu}) \quad \text{for all } (\tau,\hat{\xi}) \text{ in } (\partial\hat{F})_{\epsilon_0} \cap \{(t,\hat{x}) : |\hat{x}| < R\}.$$

Let $\hat{F}_\mu = \{(t,\hat{x}) : (t,\hat{x}) \in \hat{F},\ |\hat{\rho}(t,\hat{x})| < \mu\}$. Since either Assumption II or II′ holds for F, there exists a $\mu_0 > 0$ such that for all $0 < \mu < \mu_0$, $(\hat{F} - \hat{F}_\mu) \cap \{(t,\hat{x}) : |\hat{x}| < R\} \neq \emptyset$.

We now define as in [2] a family of games of fixed duration. Let $\gamma(r) = 1 - r$ if $0 \le r \le 1$ and let $\gamma(r) = 0$ if $r > 1$. For each $0 < \mu < \mu_0$ let

$$f^0_\mu(t,\hat{x},y,z) = \begin{cases} f^0(t,x,y,z)\gamma(|\hat{\rho}(t,\hat{x})|/\mu) & \text{if } (t,\hat{x}) \in \hat{F}, \\ f^0(t,x,y,z) & \text{if } (t,\hat{x}) \notin \hat{F}. \end{cases}$$

The function f^0_μ is continuous on $D \times (0,\mu_0)$, and $f^0_\mu(t,\hat{x},y,z) = 0$ if $(t,\hat{x}) \in \hat{F} - \hat{F}_\mu$. For each integer k such that $\epsilon_k = \frac{1}{k} < \mu_0$ we consider the game $\hat{G}_k(t_0,\hat{x}_0) = \hat{G}_{\epsilon_k}(t_0,\hat{x}_0)$ of fixed duration with terminal time T, with dynamics given by (1) and with payoff

$$x^0_0 + \int_{t_0}^{T} f^0_{\epsilon_k}(s,\hat{\phi}(s),y(s),z(s))ds.$$

From now on, we let $(t_0,x_0) \in \mathbb{R}^{n+1}$ be fixed and let $\hat{x}_0$ be $(0,x_0)$.

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^{n+2}$ containing $(t_0,\hat{x}_0)$ as an interior point. Berkovitz showed in [3, p. 496, (4.2); p. 497, (4.5)] that if k is sufficiently large, there exists a $c_0 > 0$ such that for all $(\tau,\hat{\xi}) \in \mathcal{X}$

$$\hat{W}^+(\tau,\hat{\xi}) \le \hat{W}^+(\tau,\hat{\xi},k) + \frac{c_0}{k}, \qquad \hat{W}^-(\tau,\hat{\xi}) \ge \hat{W}^-(\tau,\hat{\xi},k) - \frac{c_0}{k}, \tag{5}$$

where $\hat{W}^\pm(\tau,\hat{\xi},k)$ are the upper and lower values of $\hat{G}_k(\tau,\hat{\xi})$.

Since ∂F is $C^{(2)}$, $\partial\hat{F}$ is also $C^{(2)}$. It follows that if $\hat{\mathcal{B}}$ is a compact subset of $\partial\hat{F}$, then there exists an $\epsilon_0 > 0$ such that $\hat{\rho}$ is C^1 on $\mathcal{N}_{\epsilon_0}(\hat{\mathcal{B}})$, where $\mathcal{N}_{\epsilon_0}(\hat{\mathcal{B}})$ is the closed ϵ_0-neighborhood of $\hat{\mathcal{B}}$ in $\mathbb{R}^{n+2}$. Also, at points $(t_1,\hat{x}_1)$ of $\partial\hat{F}$, $(\hat{\rho}_t(t,\hat{x}), \hat{\rho}_{\hat{x}}(t,\hat{x})) \to (\nu^0,\hat{\nu})$ as $(t,\hat{x}) \notin \hat{F}$ tends to $(t_1,\hat{x}_1)$. Moreover, on compact subsets of $\partial\hat{F}$, the convergence is uniform. Therefore the following lemma follows from Lemma 1.

LEMMA 2. *If $\hat{\mathcal{B}}$ is a compact subset of $\partial\hat{F}$, there exist a $c > 0$ and an $\epsilon > 0$ such that for all $(t,\hat{x}) \in \mathcal{N}_\epsilon(\hat{\mathcal{B}})$, $\hat{\rho}_t(t,\hat{x}) + \langle \hat{\rho}_{\hat{x}}(t,\hat{x}), \hat{f}(t,\hat{x},y,z)\rangle \le -c$.*

LEMMA 3. *Let F_1 satisfy either Assumption* II *or* II′. *Let $(t_0,\hat{x}_{0n})$ be in $\mathcal{X}$ such that $\lim_{n\to\infty} \hat{x}_{0n} = \hat{x}_0$. Let $\{\hat{\phi}_n\}$ be a sequence of solutions of the differential equation*

$$\frac{d\hat{x}}{dt} = \hat{f}(t,\hat{x},u_n(t),v_n(t)), \qquad \hat{x}(t_0) = \hat{x}_{0n}, \quad t \in [t_0,T]$$

and let $\hat{\phi}[\cdot]$ be a motion resulting from $\{\hat{\phi}_n\}$. If t_{f_n}, t_f are the terminal times of $\hat{\phi}_n(\cdot)$ and $\hat{\phi}[\cdot]$, respectively, then $\lim_{n\to\infty} t_{f_n} = t_f$.

Proof. As in [2, Lem. 4.1], we have $t_f \le \liminf_{n\to\infty} t_{f_n}$.

We show by contradiction that any convergent subsequence of $\{t_{f_n}\}$ has limit t_f. Suppose there were a subsequence of $\{t_{f_n}\}$, still denoted as $\{t_{f_n}\}$, converging to a

limit greater than t_f, i.e., $\lim_n t_{f_n} > t_f$. Since each $t_{f_n} \le T$, $t_f < T$. Since $(t_0, \hat{x}_0)$ is in $\mathcal{X}$, there exists an $R > 0$ such that for any$\{\hat{\phi}_n(\cdot)\}$ and any motion $\hat{\phi}[\cdot]$ resulting from $\{\hat{\phi}_n\}$, $\{(t, \hat{\phi}_n(t))\}$, $(t, \hat{\phi}[t])$ will lie in $[t_0, T] \times \hat{B}_R$ for all $t \in [t_0, T]$, where $\hat{B}_R$ denotes the closed ball of radius R in $\mathbb{R}^{n+1}$. Hence $(t_f, \hat{\phi}[t_f])$, $\{(t_{f_n}, \hat{\phi}_n(t_{f_n}))\}$ will lie in $([t_0, T] \times \hat{B}_R) \cap \partial\hat{T}$. Let $\hat{\mathcal{B}}_R = ([t_0, T] \times \hat{B}_R) \cap \partial\hat{F}$. Then $\hat{\mathcal{B}}_R$ is a compact subset of $\partial\hat{F}$. The inequality $t_f < T$ implies that $(t_f, \hat{\phi}[t_f]) \in \hat{\mathcal{B}}_R$.

Let $\epsilon_0 = \epsilon_0(\hat{\mathcal{B}}_R)$, $c = c(\hat{\mathcal{B}}_R)$ be the constants in Lemma 2 associated with $\hat{\mathcal{B}}_R$. Since $\hat{\phi}_n(\cdot) \to \hat{\phi}[\cdot]$ uniformly on $[t_0, T]$, for any $\epsilon > 0$ with $\epsilon \le \epsilon_0$, there exists an integer $N_1 > 0$ such that if $n \ge N_1$

$$|\hat{\phi}_n(t) - \hat{\phi}[t]| < \epsilon \quad \text{for all } t \in [t_0, T]. \tag{6}$$

Therefore $(t_f, \hat{\phi}_n(t_f)) \in \mathcal{N}_\epsilon(\hat{\mathcal{B}}_R)$ if $n \ge N_1$.

Since $\lim t_{f_n} > t_f$, there exists $N_2 > 0$ such that if $n \ge N_2$, $t_{f_n} > t_f$. Let $N = \max(N_1, N_2)$. Let $\theta_n(t) = \hat{\rho}(t, \hat{\phi}_n(t))$, then

$$\frac{d\theta_n}{dt} = \hat{\rho}_t(t, \hat{\phi}_n(t)) + \langle \hat{\rho}_x(t, \hat{\phi}_n(t)), \hat{f}(t, \hat{\phi}_n(t), u_n(t), v_n(t)) \rangle. \tag{7}$$

If $(t, \hat{\phi}_n(t))$ stays in $\mathcal{N}_\epsilon(\hat{\mathcal{B}}_R)$, the right-hand side of the above equation is $\le -c$.

Since $(t_f, \hat{\phi}_n(t_f)) \in \mathcal{N}_\epsilon(\hat{\mathcal{B}}_R)$, it follows from the continuity of $\hat{\phi}_n$ that there exists a maximal interval $[t_f, t_f + \alpha_n)$ with $t_f + \alpha_n \le t_{f_n}$ such that if $t \in [t_f, t_f + \alpha_n)$, then $(t, \hat{\phi}_n(t)) \in \mathcal{N}_\epsilon(\hat{\mathcal{B}}_R)$ and $(t, \hat{\phi}_n(t)) \notin \hat{T}$. Note that $\alpha_n > 0$. For all $t \in [t_f, t_f + \alpha_n)$, (7) gives

$$\theta_n(t) - \theta_n(t_f) \le -c(t - t_f). \tag{8}$$

We claim that $t_f + \alpha_n = t_{f_n}$. Otherwise, we would have $\theta_n(t_f) < \epsilon$ and $\theta_n(t_f + \alpha_n) = \epsilon$. Then (8) would give that $0 < -c\alpha_n$, which is impossible. Therefore (8) holds for all $t \in [t_f, t_{f_n}]$.

If $t_{f_n} < T$, $\theta_n(t_{f_n}) = 0$. Then $(t_{f_n} - t_f) \le c^{-1}\theta_n(t_f) < \frac{\epsilon}{c}$ for $n \ge N$. If $t_{f_n} = T$, $\theta_n(T) \ge 0$. Then $(T - t_f) \le c^{-1}(\theta_n(t_f) - \theta_n(T)) \le c^{-1}\theta_n(t_f) < \frac{\epsilon}{c}$ for $n \ge N$.

In any case, we have $|t_{f_n} - t_f| < \frac{\epsilon}{c}$ if $n \ge N$. It follows that $\lim t_{f_n} = t_f$, which is a contradiction. Hence the lemma is proved. □

The proof of Lemma 3 gives the following result.

COROLLARY 4. *Suppose that F_1 satisfies either Assumption* II *or* II′. *Let ϵ_0 and c be determined by $\hat{\mathcal{B}}_R$ and let t_f be the terminal time of the trajectory $\hat{\phi}(t)$. Let $0 \le \epsilon \le \epsilon_0$ and let $\bar{t} < t_f$. If $(\bar{t}, \hat{\phi}(\bar{t})) \in \mathcal{N}_\epsilon(\hat{\mathcal{B}}_R)$, then $t_f - \bar{t} \le \frac{\epsilon}{c}$.*

COROLLARY 5. *Under either Assumption* II *or* II′, *once a trajectory or a motion gets into the terminal set, it can never get out.*

Proof. If the statement were false, then for some trajectory $\hat{\phi}(t, t_0, x_0, u, v)$ there would exist t_1 and t_2 with $t_1 < t_2$ such that

(1) $(t_1, \hat{\phi}(t_1)) \in \hat{T}$ and $(t_2, \hat{\phi}(t_2)) \notin \hat{T}$,

(2) $(t, \hat{\phi}(t)) \in \mathcal{N}_\epsilon(\hat{\mathcal{B}}_R)$ for $t_1 \le t \le t_2$.

Let $\theta(t) = \hat{\rho}(t, \hat{\phi}(t))$. From (8), we get that

$$\theta(t_2) - \theta(t_1) \le -c(t_2 - t_1)$$

or

$$\theta(t_2) \le -c(t_2 - t_1) + \theta(t_1).$$

Since $(t_1, \hat{\phi}(t_1)) \in \hat{T}$, $\theta(t_1) \leq 0$. So, $\theta(t_2) < 0$, which contradicts the assumption that $(t_2, \hat{\phi}(t_2)) \notin \hat{T}$. □

We assume that the following Isaacs condition holds. For all vectors $\hat{s} = (1, s)$ in $\mathbb{R}^{n+1}$ and any $(t, \hat{x}) \in [t_0, T] \times \mathbb{R}^{n+1}$

$$(\hat{I}) \qquad \max_y \min_z \langle \hat{s}, \hat{f}(t, x, y, z) \rangle = \min_z \max_y \langle \hat{s}, \hat{f}(t, x, y, z) \rangle,$$

where y ranges over the set Y and z ranges over the set Z.

Let v be any real number. Let

$$C^{(k)}(v) = \{(\tau, \hat{\xi}) : t_0 \leq \tau \leq T, \hat{\xi} \in \mathbb{R}^{n+1}, \hat{W}^-(\tau, \hat{\xi}, k) \leq v\}$$

and for any $\alpha > 0$, let

$$C_\alpha^{(k)}(v) = \{(\tau, \hat{\xi}) : t_0 \leq t \leq T, \hat{\xi} \in \mathbb{R}^{n+1}, \ \operatorname{dist}\big((\tau, \hat{\xi}), C^{(k)}(v)\big) \leq \alpha\}.$$

Let $\nu_k = \hat{W}^-(t_0, \hat{x}_0, k)$. Let V_k^ϵ be extremal to $C^{(k)}(\nu_k)$ and let $\Delta_k^\epsilon = \Delta_k^\epsilon(V_k^\epsilon)$ be the corresponding feedback strategy (see [1, p. 189]). Let Γ be any strategy for Player I and let $\hat{\phi}_k[\ , t_0, \hat{x}_0, \Gamma, \Delta_k^\epsilon]$ be any motion resulting from $(\Gamma, \Delta_k^\epsilon)$. Let $\{\hat{\phi}_{k,n}(\ , t_0, \hat{x}_0, u_{k,n}, v_{k,n})\}$ be a sequence of trajectories converging uniformly to $\hat{\phi}_k[\cdot]$ on $[t_0, T]$. Let t_f be the terminal time of $\hat{\phi}[\cdot]$ and let t_{f_n} be the terminal time of $\hat{\phi}_{k,n}(\cdot)$ (since the terminal time of $\hat{\phi}_{k,n}(\cdot)$ does not depend on k). Lemma 3 asserts that $t_{f_n} \to t_f$ as $n \to \infty$.

Since $\mathcal{X}$ is a compact subset of $\mathbb{R}^{n+2}$ containing $(t_0, \hat{x}_0)$ as its interior point, it follows from Assumption I that there exists $R > 0$ such that all of the trajectories $(t, \hat{\phi}(t))$ and motions $(t, \hat{\phi}[t])$ initiated from $\mathcal{X}$ will lie in $[t_0, T] \times \hat{\mathcal{B}}_R$ where $\hat{\mathcal{B}}_R = \{\hat{x} \in \mathbb{R}^{n+1} : \ |\hat{x}| \leq R\}$. Let $\mathcal{D} = \{(\tau, \hat{\xi}) : \ \tau \in [t_0, T], \ |\xi| \leq R\}$. Let

$$(9) \qquad M = \max_{\substack{(t,\hat{x}) \in \mathcal{D} \\ y \in Y, z \in Z}} |\hat{f}(t, x, y, z)|.$$

LEMMA 6. *Let Assumption* I *and either Assumption* II *or* II′ *hold and let the Isaacs condition* $(\hat{I})$ *hold for* $\hat{f}$. *Then there exist a nonnegative nondecreasing function* α *and a* k_0 *such that* $\alpha(\delta) \to 0$ *as* $\delta \to 0$ *and for* $n \geq k_0$

$$(10) \qquad (t, \hat{\phi}_{k,n}(t, t_0, \hat{x}_0, u_{k,n}, v_{k,n})) \in C^{(k)}_{(((M+1)/c)+1)^2\alpha(1/n)}(\nu_k)$$

for all $t_0 \leq t \leq t_{f_n}$ *and all* k.

Proof. Let $\hat{\mathcal{B}}_R$ be defined as in the proof of Lemma 3. Then $\hat{\mathcal{B}}_R$ is compact in $\partial \hat{F}$. Let $\epsilon = \epsilon(\hat{\mathcal{B}}_R) > 0$, $c = c(\hat{\mathcal{B}}_R) > 0$ be the constants determined as in Lemma 2.

If $\hat{\phi}_{k,n}(\cdot)$ never leaves $C^{(k)}(\nu_k)$, (10) certainly holds. Assume that $t_1 \in (t_0, T]$ is the first time at which $\hat{\phi}_{k,n}(\cdot)$ leaves $C^{(k)}(\nu_k)$. Let

$$\Pi(\Delta_{k,n}^\epsilon) : \ t_0 = \tau_0 < \tau_1 < \cdots < \tau_{j_0} \leq t_1 < \tau_{j_0+1} < \cdots < \tau_{p_{\Delta_{k,n}^\epsilon}} = T$$

be the nth partition of Δ_k^ϵ. Note that $t_1 \in [\tau_{j_0}, \tau_{j_0+1})$.

Consider $t \in [t_0, t_{f_n}]$ and suppose that $t \in [\tau_j, \tau_{j+1}]$ with $j \geq j_0 + 1$. Let $\hat{x}^* = \hat{\phi}_{k,n}(\tau_j)$, let $t^* = \tau_j$, and let $\hat{w}^*$ be the point in $C^{(k)}(\nu_k) \cap H(\tau_j)$ in the definition of

$V_k^e(t^*, \hat{x}^*) = V_k^e(\tau_j, \hat{\phi}_{k,n}(\tau_j))$. Let $\hat{s}^* = \hat{x}^* - \hat{w}^*$ and let y^* be any point in Y such that (y^*, z^*) is a saddle point for the local game $(t^*, \hat{x}^*, \hat{s}^*) = (\tau_j, \hat{\phi}_{k,n}(\tau_j), \hat{s}^*)$ with payoff $\langle \hat{s}^*, \hat{f}_k(t^*, \hat{x}^*, y, z)\rangle$.

Case 1. Assume $(t^*, \hat{w}^*) \notin \hat{\mathcal{T}}$. By [1, Lem. 8.3], for each $t > \tau_j$, there exists a relaxed control ζ_t such that the corresponding relaxed trajectory $\hat{\psi}_k(\cdot) = \hat{\psi}_k(\cdot, \tau_j, \hat{w}^*, u_{k,n}, \zeta_t)$ has the property that

$$(t, \hat{\psi}_k(t, \tau_j, \hat{w}^*, u_{k,n}, \zeta_t) \in C^{(k)}(\nu_k). \tag{11}$$

Let $e_{k,n}(t)$ denote the distance from $(t, \hat{\phi}_{k,n}(t, t_0, \hat{x}_0, u_{k,n}, v_{k,n})$ to $C^{(k)}(\nu_k)$. Since $\hat{\phi}_{k,n}(s, t_0, \hat{x}_0, u_{k,n}, v_{k,n}) = \hat{\phi}_{k,n}(s, \tau_j, \hat{\phi}_{k,n}(\tau_j), u_{k,n}, v_{k,n})$ for $s > \tau_j$, we have that

$$e_{k,n}(t) \le |\hat{\phi}_{k,n}(t, \tau_j, \hat{\phi}_{k,n}(\tau_j), u_{k,n}, v_{k,n}) - \hat{\psi}_k(t, \tau_j, \hat{w}^*, u_{k,n}, \zeta_t)|.$$

Case 1(a). If $(t, \hat{\psi}_k(t, \tau_j, \hat{w}^*, u_{k,n}, \zeta_t)) \notin \hat{\mathcal{T}}$, then by Corollary 5, for all $\tau_j \le s \le t$, $(s, \hat{\psi}_k(s, \tau_j, \hat{w}^*, u_{k,n}, \zeta_t)) \notin \hat{\mathcal{T}}$. Since $\hat{\psi}_k$ is the solution of $dx/ds = f(s, x, u_{k,n}, \zeta_t)$ for $\tau_j \le s \le t$, following the argument given in [1, p. 191] we get that for n sufficiently large

$$e_{k,n}^2(t) \le e_{k,n}^2(\tau_{j_0+1}) e^{\beta K} + E(\delta_n)(e^{\beta K} - 1)/\beta \quad \text{for } \tau_{j_0+1} \le t, \tag{12}$$

where K is the Lipschitz constant of $\hat{f}$, β is $2K$, L is a constant such that $\|\Pi_n\| = \delta_n \le L/n$, and $E(\cdot)$ is a nondecreasing function defined on $[0, \infty)$ such that $E(\delta) \to 0$ as $\delta \to 0$. Since $e_{k,n}(t_1) = 0$, we get that $e_{k,n}(\tau_{j_0+1}) \le ML/n$. Let $\alpha(\cdot)$ be a nondecreasing function defined on $[0, \infty)$ such that

$$\Big(\frac{ML}{n}\Big)^2 e^{\beta K} + E\Big(\frac{L}{n}\Big)(e^{\beta K} - 1)/\beta \le \alpha^2\Big(\frac{1}{n}\Big)$$

and such that $\alpha(\delta) \to 0$ as $\delta \to 0$. Then (12) gives that $e_{k,n}(t) \le \alpha(1/n)$, and hence

$$(t, \hat{\phi}_{k,n}(t, t_0, \hat{x}_0, u_{k.n}, v_{k,n})) \in C^{(k)}_{\alpha(1/n)}(\nu_k) \subset C^{(k)}_{(((M+1)/c)+1)^2\alpha(1/n)}(\nu_k).$$

Case 1(b). If for some $t \in [\tau_j, \tau_{j+1})$ every relaxed control ζ_t with the property (11) satisfies

$$(t, \hat{\psi}_k(t, \tau_j, \hat{w}^*, u_{k,n}, \zeta_t)) \in \hat{\mathcal{T}},$$

then let t_{f_ψ} be the infimum of such t. Hence for any $t \in [t_0, t_{f_\psi})$, the argument in Case 1(a) implies that

$$(t, \hat{\phi}_{k,n}(t)) \in C^{(k)}_{\alpha(1/n)}(\nu_k).$$

The continuity of $\hat{\phi}_{k,n}$ and the closedness of $C^{(k)}_{\alpha(1/n)}(\nu_k)$ imply that the preceding inclusion holds for all $t \in [t_0, t_{f_\psi}]$.

If $t_{f_n} \le t_{f_\psi}$, we are done with Case 1(b), and hence Case 1. If not, we claim that $(t_{f_\psi}, \hat{\phi}_{k,n}(t_{f_\psi})) \in \mathcal{N}_{\alpha(1/n)}(\hat{\mathcal{B}}_R)$. In fact, for any small $\delta > 0$, there exist t_δ^+ and t_δ^- such that $|t_{f_\psi} - t_\delta^\pm| \le \delta$ and a relaxed control $\zeta_{t_\delta^-}$ such that

$$(t_\delta^-, \hat{\psi}_k(t_\delta^-, \tau_j, \hat{w}^*, u_{k,n}, \zeta_{t_\delta^-})) \in C^{(k)}(\nu_k),$$

$$(t_\delta^-, \hat{\psi}_k(t_\delta^-, \tau_j, \hat{w}^*, u_{k,n}, \zeta_{t_\delta^-})) \notin \hat{T},$$

and

$$(t_\delta^+, \hat{\psi}_k(t_\delta^+, \tau_j, \hat{w}^*, u_{k,n}, \zeta_{t_\delta^-})) \in \hat{T}.$$

Therefore, we get

$$\begin{aligned}
&|(t_{f_\psi}, \hat{\phi}_{k,n}(t_{f_\psi}, \tau_j, \hat{\phi}_{k,n}(\tau_j), u_{k,n}, v_{k,n})) - (t_\delta^+, \hat{\psi}_k(t_\delta^+, \tau_j, \hat{w}^*, u_{k,n}, \zeta_{t_\delta^-}))| \\
&\le |t_{f_\psi} - t_\delta^+| + |\hat{\phi}_{k,n}(t_{f_\psi}) - \hat{\phi}_{k,n}(t_\delta^-)| + |\hat{\phi}_{k,n}(t_\delta^-) - \hat{\psi}_k(t_\delta^-)| \\
&\qquad + |\hat{\psi}_k(t_\delta^-) - \hat{\psi}_k(t_\delta^+)| \\
&\le \delta + M\delta + \alpha\Big(\frac{1}{n}\Big) + 2M\delta \\
&= (1 + 3M)\delta + \alpha\Big(\frac{1}{n}\Big).
\end{aligned}$$

The arbitrariness of δ implies the claim.

By Corollary 4, we get that $t_{f_n} - t_{f_\psi} \le \frac{1}{c}\alpha(1/n)$. This gives that for $t \in [t_{f_\psi}, t_{f_n}]$

$$\begin{aligned}
e_{k,n}(t) &\le |(t, \hat{\phi}_{k,n}(t)) - (t_{f_\psi}, \hat{\phi}_{k,n}(t_{f_\psi}))| + e_{k,n}(t_{f_\psi}) \\
&\le |t - t_{f_\psi}| + |\hat{\phi}_{k,n}(t) - \hat{\phi}_{k,n}(t_{f_\psi})| + e_{k,n}(t_{f_\psi}) \\
&\le \frac{1}{c}\alpha\Big(\frac{1}{n}\Big) + M\frac{1}{c}\alpha\Big(\frac{1}{n}\Big) + \alpha\Big(\frac{1}{n}\Big) \\
&= \Big(\frac{M+1}{c} + 1\Big)\alpha\Big(\frac{1}{n}\Big).
\end{aligned}$$

Hence (10) holds.

Case 2. Assume that $(t^*, \hat{w}^*) \in \hat{T}$ for some $j \ge j_0 + 1$. Let t^* be the first partition point such that $(t^*, \hat{w}^*) \in \hat{T}$. Then Case 1 says that for $t_0 \le t \le t^*$, $e_{k,n}(t) \le (((M+1)/c) + 1)\alpha(1/n)$. Hence $(t^*, \hat{\phi}_{k,n}(t^*)) \in \mathcal{N}_{(((M+1)/c)+1)\alpha(1/n)}(\hat{\mathcal{B}}_R)$. Corollary 4 says if n is large enough, $t_{f_n} - t^* \le \frac{1}{c}(((M+1)/c) + 1)\alpha(1/n)$. Therefore, for $t \in [t^*, t_{f_n}]$

$$\begin{aligned}
e_{k,n}(t) &\le |(t, \hat{\phi}_{k,n}(t)) - (t^*, \hat{\phi}_{k,n}(t^*))| + e_{k,n}(t^*) \\
&\le |t - t^*| + |\hat{\phi}_{k,n}(t) - \hat{\phi}_{k,n}(t^*)| + e_{k,n}(t^*) \\
&\le \frac{1}{c}\Big(\frac{M+1}{c} + 1\Big)\alpha\Big(\frac{1}{n}\Big) + \frac{M}{c}\Big(\frac{M+1}{c} + 1\Big)\alpha\Big(\frac{1}{n}\Big) + \Big(\frac{M+1}{c} + 1\Big)\alpha\Big(\frac{1}{n}\Big) \\
&= \Big(\frac{M+1}{c} + 1\Big)^2\alpha\Big(\frac{1}{n}\Big).
\end{aligned}$$

Hence, (10) holds. □

THEOREM 7. *If Assumption* I *and either Assumption* II *or* II′ *hold and if Isaacs condition* $(\hat{I})$ *holds for* $\hat{f}$, *then there exists a strategy* Δ^* *for Player* II *such that for any strategy* Γ *for Player* I

$$P(t_0, x_0, \Gamma, \Delta^*) \le W(t_0, x_0).$$

Proof. First, note that under the assumptions of the theorem, we have $\hat{W}(\tau, \hat{\xi}, k) = \hat{W}^{\pm}(\tau, \hat{\xi}, k)$, $W(t_0, x_0) = W^{\pm}(t_0, x_0)$, and $\hat{W}(\tau, \hat{\xi}) = \hat{W}^{\pm}(\tau, \hat{\xi})$. Therefore, the following is true from (5):

$$\hat{W}(\tau, \hat{\xi}, k) - \frac{c_0}{k} \le \hat{W}(\tau, \hat{\xi}) \le \hat{W}(\tau, \hat{\xi}, k) + \frac{c_0}{k}.$$

Note also that $\hat{W}(t_0, \hat{x}_0) = W(t_0, x_0)$ since $\hat{x}_0 = (0, x_0)$.

Let Γ be a strategy for Player I. If $\Delta_k^e = \{\Delta_{k,n}^e\}$, let $\Delta^* = \{\Delta_{k,k}^e\}$. Let $\hat{\phi}[\ , t_0, \hat{x}_0, \Gamma, \Delta^*]$ be a motion and assume that $\{\hat{\phi}_k(t, t_0, \hat{x}_0, u_k, v_k)\}$ is a sequence of kth stage trajectories converging uniformly to $\hat{\phi}[\cdot]$.

Let t_f be the terminal time of $\hat{\phi}[\cdot]$ and let t_{f_k} be the terminal time of $\hat{\phi}_k(\cdot)$. Then by Lemma 3, $t_f = \lim_{k\to\infty} t_{f_k}$.

It follows from Lemma 6 that $(t_{f_k}, \hat{\phi}_k(t_{f_k})) \in C_{(((M+1)/c)+1)^2\alpha(1/k)}(\nu_k)$ for k sufficiently large. Since

$$\hat{W}(t_0, \hat{x}_0, k) \le \hat{W}(t_0, \hat{x}_0) + \frac{c_0}{k} = W(t_0, x_0) + \frac{c_0}{k} \quad \text{and} \quad \nu_k = \hat{W}(t_0, \hat{x}_0, k),$$

$$(t_{f_k}, \hat{\phi}_k(t_{f_k})) \in C_{(((M+1)/c)+1)^2\alpha(1/k)}\left(W(t_0, x_0) + \frac{c_0}{k}\right). \tag{13}$$

If we let

$$C = \bigcap_{k=1}^{\infty} C_{\left(((M+1)/c)+1\right)^2\alpha(1/k)}\left(W(t_0, x_0) + \frac{c_0}{k}\right),$$

then C is closed. Since $\{C_{(((M+1)/c)+1)^2\alpha(1/k)}\ (W(t_0, x_0) + c_0/k)\}$ is a sequence of decreasing sets, it follows from (13) that $(t_f, \hat{\phi}[t_f]) \in C$.

Let $C' = \{(\tau, \hat{\xi}) :\ t_0 \le \tau \le T, \hat{\xi} \in \mathbb{R}^{n+1}, \hat{W}(\tau, \hat{\xi}) \le \hat{W}(t_0, \hat{x}_0)\}$. We claim that $C = C'$.

If the claim is true, then we have $(t_f, \hat{\phi}[t_f]) \in C'$, i.e., $\phi^0[t_f] = \phi^0[t_f, t_0, x_0, \Gamma, \Delta^*] \le W(t_0, x_0)$. Hence, we have $P(t_0, x_0, \Gamma, \Delta^*) \le W(t_0, x_0)$ for any Γ.

Now, let us prove the claim. Assume that $(\tau, \hat{\xi}) \in C'$, so $\hat{W}(\tau, \hat{\xi}) \le \hat{W}(t_0, \hat{x}_0) = W(t_0, x_0)$. Then $\hat{W}(\tau, \hat{\xi}, k) \le \hat{W}(\tau, \hat{\xi}) + c_0/k \le W(t_0, x_0) + c_0/k$ for any k. Hence $(\tau, \hat{\xi}) \in C$, and we get $C' \subset C$. If $(\tau, \hat{\xi}) \in C$, $(\tau, \hat{\xi}) \in C_{(((M+1)/c)+1)^2\alpha(1/k)}(W(t_0, x_0) + c_0/k)$ for any k. Therefore, there exists $(\tau_k, \hat{\xi}_k) \in C(W(t_0, x_0) + c_0/k)$ such that $|(\tau_k, \hat{\xi}_k) - (\tau, \hat{\xi})| \le \left(((M+1)/c) + 1\right)^2\alpha(1/k)$. Hence, we have that

$$\hat{W}(\tau_k, \hat{\xi}_k) \le \hat{W}(\tau_k, \hat{\xi}_k, k) + \frac{c_0}{k} \le W(t_0, x_0) + \frac{2c_0}{k}. \tag{14}$$

Since $\hat{W}(t, \hat{x})$ is continuous [2] and since $\lim_{k\to\infty}(\tau_k, \hat{\xi}_k) = (\tau, \hat{\xi})$, it follows from (14) that $\hat{W}(\tau, \hat{\xi}) \le W(t_0, x_0)$, i.e., $(\tau, \hat{\xi}) \in C'$ and hence $C \subset C'$. This completes the proof of Theorem 7. □

Similarly, we can show using (5) that if Assumption I and either Assumption II or II′ hold and if Isaacs condition $(\hat{I})$ holds for $\hat{f}$, there exists a strategy Γ^* for Player I such that $W(t_0, x_0) \le P(t_0, x_0, \Gamma^*, \Delta)$ for any Δ. Then, clearly, (Γ^*, Δ^*) constitutes a saddle point in the game.

2. Differential games with information lags. In this section, we first study games of fixed duration with lags and then study games of generalized pursuit and evasion with lags and games of survival with lags.

2.1. Differential games of fixed duration with lags. Our game is governed by (1) with payoff

$$P(t_0, x_0, \Gamma, \Delta) = g(\phi[T]) = g(\phi[T, t_0, x_0, \Gamma, \Delta]) \tag{15}$$

where Γ is a strategy selected by Player I prior to the start of play, Δ is a strategy selected by Player II prior to the start of play, and $\phi[\cdot, t_0, x_0, \Gamma, \Delta]$ is a motion corresponding to (Γ, Δ). The concepts of strategy and motion for games with lags will be given later.

Let Player I and Player II have lags $\lambda \geq 0$ and $\mu \geq 0$, respectively. Let Y be a compact subset of $\mathbb{R}^r$ and let Z be a compact subset of $\mathbb{R}^s$. Let $t_0 < T$ and let $D = [t_0, T] \times \mathbb{R}^n \times Y \times Z$.

ASSUMPTION I′. *Assumption* I *in* §1 *holds with* $\hat{f}$ *replaced by* f *and the function* g *is continuous on* $\mathbb{R}^n$.

If Assumption I′ holds and if the Isaacs condition holds for f on $[t_0, T] \times \mathbb{R}^n$, i.e., for all vector s in $\mathbb{R}^n$

$$\max_{y \in Y} \min_{z \in Z} \langle s, f(t, x, y, z) \rangle = \min_{z \in Z} \max_{y \in Y} \langle s, f(t, x, y, z) \rangle$$

then from [1] we know that a game of fixed duration with initial condition (t_0, x_0) and payoff (15) has a value. We will denote the value by $W(t_0, x_0)$. Moreover, there exists a pair of strategies (Γ^*, Δ^*) such that $W(t_0, x_0) = g(\Phi[T, t_0, x_0, \Gamma^*, \Delta^*])$.

For $\sigma > 0$, let $Y_{[t,t+\sigma)}$ denote the set of all measurable functions u defined on $[t, t+\sigma)$ and satisfying $u(t) \in Y$ almost everywhere, and let $Z_{[t,t+\sigma)}$ denote the set of all measurable functions v defined on $[t, t+\sigma)$ and satisfying $v(t) \in Z$ almost everywhere. Such functions will be called control functions or controls.

A strategy Γ^λ for Player I is a choice of a sequence $\Pi = \{\Pi_n\}$ of partitions of $[t_0, T]$ and a choice of a sequence of maps $\Gamma^\lambda_\Pi = \{\Gamma^\lambda_{\Pi,n}\}$, where the $\Gamma^\lambda_{\Pi,n}$ will be defined below. Thus $\Gamma^\lambda = (\Gamma^\lambda_\Pi, \Pi)$. For simplicity, we write Γ^λ forΓ^λ_Π and Γ^λ_n for $\Gamma^\lambda_{\Pi,n}$. We restrict the choice of sequences of partitions to those such that the norm of Π_n, denoted by $\|\Pi_n\|$, satisfies $\|\Pi_n\| \leq L/n$, where L is a constant independent of n. Let the partition points of Π_n be $t_0 < t_1 < \cdots < t_p = T$, where $p \leq L_1 n$ and L_1 is a constant independent of n and $LL_1 > 1$. Each map Γ^λ_n is a collection of maps $\Gamma^\lambda_{n,1}, \ldots, \Gamma^\lambda_{n,p}$ as follows. The map $\Gamma^\lambda_{n,1}$ selects an element in $Y_{[t_0,t_1)}$. For $2 \leq j \leq p$, the map $\Gamma^\lambda_{n,j}$ is a map from $Y_{[t_0,t_{j-1})} \times Z_{[t_0,t_{j-1}-\lambda)}$ to $Y_{[t_{j-1},t_j)}$. If $t_{j-1} - \lambda < t_0$, we replace $Y_{[t_0,t_{j-1})} \times Z_{[t_0,t_{j-1}-\lambda)}$ by $Y_{[t_0,t_{j-1})}$.

A strategy Δ^μ for Player II is a choice of sequence of partitions $\bar{\Pi} = \{\bar{\Pi}_n\}$ of $[t_0, T]$ such that $\|\bar{\Pi}_n\| \leq L/n$ and a choice of sequence of maps $\{\Delta^\mu_n\}$. Each Δ^μ_n is a collection of maps $\Delta^\mu_{n,1}, \ldots, \Delta^\mu_{n,q}$ with $q \leq nL_1$, as follows. If $\bar{\Pi}_n$ has partition points $t_0 = s_0 < s_1 < \cdots < s_q = T$, then $\Delta^\mu_{n,1}$ selects a function v in $Z_{[s_0,s_1)}$. For $2 \leq j \leq q$, $\Delta^\mu_{n,j}$ is a map from $Y_{[s_0,s_{j-1}-\mu)} \times Z_{[s_0,s_{j-1})}$ to $Z_{[s_{j-1},s_j)}$. If for some j, $s_{j-1} - \mu < s_0$, replace $Y_{[s_0,s_{j-1}-\mu)} \times Z_{[s_0,s_{j-1})}$ by $Z_{[s_0,s_{j-1})}$ in the definition of $\Delta^\mu_{n,j}$.

A pair of nth stage strategies $(\Gamma^\lambda_n, \Delta^\mu_n)$ determines control functions (u_n, v_n) on $[t_0, T]$, where

$$\begin{aligned}
u_n(t) &= (\Gamma^\lambda_{n,1})(t), && t \in [t_0, t_1), \\
v_n(t) &= (\Delta^\mu_{n,1})(t), && t \in [s_0, s_1), \\
u_n(t) &= (\Gamma^\lambda_{n,j})(u_{[t_0,t_{j-1})}, v_{[t_0,t_{j-1}-\lambda)})(t), && t \in [t_{j-1}, t_j), \quad j = 2, \ldots, p, \\
v_n(t) &= (\Delta^\mu_{n,j})(u_{[s_0,s_{j-1}-\mu)}, v_{[s_0,s_{j-1})})(t), && t \in [s_{j-1}, s_j), \quad j = 2, \ldots, q.
\end{aligned}$$

The control functions (u_n, v_n) determined this way are called the nth stage outcomes of $(\Gamma^\lambda, \Delta^\mu)$.

In differential equations (1), if we replace y by $u_n(t)$, z by $v_n(t)$, and x_0 by x_{0n}, we obtain the system of differential equations

$$\frac{dx}{dt} = f(t, x, u_n(t), v_n(t)),$$
$$x(t_0) = x_{0n}.$$

The unique solution $\phi_n(\ , t_0, x_{0n}, u_n, v_n)$ defined on $[t_0, T]$ is called an nth stage trajectory.

Any uniform limit of a convergent subsequence of $\{\phi_n(\ , t_0, x_{0n}, u_n, v_n)\}$ where $x_{0n} \to x_0$ and (u_n, v_n) is the outcome of $(\Gamma_n^\lambda, \Delta_n^\mu)$, will be called a motion or motion of the game corresponding to strategies $(\Gamma^\lambda, \Delta^\mu)$. We denote a motion corresponding to $(\Gamma^\lambda, \Delta^\mu)$ by $\phi[\ , t_0, x_0, \Gamma^\lambda, \Delta^\mu]$. Under Assumption I′, there do exist motions. We denote the set of all motions corresponding to $(\Gamma^\lambda, \Delta^\mu)$ by $\Phi[\ , t_0, x_0, \Gamma^\lambda, \Delta^\mu]$.

In a game of fixed duration with initial point (t_0, x_0), $P(t_0, x_0, \Gamma^\lambda, \Delta^\mu)$, the payoff corresponding to a pair of strategies $(\Gamma^\lambda, \Delta^\mu)$, is then set valued and is defined as follows:

$$P(t_0, x_0, \Gamma^\lambda, \Delta^\mu) = g(\Phi[T, t_0, x_0, \Gamma^\lambda, \Delta^\mu]).$$

We denote this game as $G_{\lambda,\mu}(t_0, x_0)$, and hence games with no lags will be written as $G_{0,0}(t_0, x_0) = G(t_0, x_0)$. Let $W^-_{\lambda,\mu}(t_0, x_0) = \sup \inf_{\Gamma^\lambda \Delta^\mu} P(t_0, x_0, \Gamma^\lambda, \Delta^\mu)$ and let $W^+_{\lambda,\mu}(t_0, x_0) = \inf \sup_{\Gamma^\lambda \Delta^\mu} P(t_0, x_0, \Gamma^\lambda, \Delta^\mu)$. Then $W^-_{\lambda,\mu}(t_0, x_0) \le W^+_{\lambda,\mu}(t_0, x_0)$.

Now we want to construct an example in which value does not exist. Consider the following problem

$$\frac{dx}{dt} = u + v \qquad 0 < t \le 1,$$
$$x(0) = 0$$

with $|u|, |v| \le 1$. The payoff function is $g(x(1)) = x^2(1)$. We assume that $0 \le \lambda \le 1$ and $0 < \mu \le 1$.

First, we shall show that $W^-_{\lambda,\mu}(0, 0) = 0$. Fix a strategy Γ^λ for Player I. We want to find a strategy $\Delta^\mu(\Gamma^\lambda)$ for Player II such that $g(\Phi[1, \Gamma^\lambda, \Delta^\mu(\Gamma^\lambda)]) = 0$. Let $\Gamma^\lambda = (\Gamma_n^\lambda)$, where $\Pi_n(\Gamma^\lambda)$ is given by $0 = \tau_0 < \tau_1 < \cdots < \tau_{p_n} = 1$. Let $\Pi_n(\Delta^\mu) = \Pi_n(\Gamma^\lambda)$. Let (u_n, v_n) be the nth outcome of $(\Gamma^\lambda, \Delta^\mu)$. If $\Gamma^\lambda_{n,1} = u_n|_{[\tau_0, \tau_1)}$, define

$$\Delta^\mu_{n,1} = -u_n|_{[\tau_0, \tau_1)} = v_n|_{[\tau_0, \tau_1)}.$$

If $\Gamma^\lambda_{n,2}(u_n|_{[\tau_0, \tau_1)}, v_n|_{[\tau_0, \tau_1 - \lambda)}) = u_n|_{[\tau_1, \tau_2)}$, define

$$\Delta^\mu_{n,2}(u|_{[\tau_0, \tau_1 - \mu)}, v|_{[\tau_0, \tau_1)}) = -u_n|_{[\tau_1, \tau_2)} = v_n|_{[\tau_1, \tau_2)}.$$

In general, if $\Gamma^\lambda_{n,i}(u_n|_{[\tau_0, \tau_{i-1})}, v_n|_{[\tau_0, \tau_{i-1} - \lambda)}) = u_n|_{[\tau_{i-1}, \tau_i)}$, then we define

$$\Delta^\mu_{n,i}(u|_{[\tau_0, \tau_{i-1} - \mu)}, v|_{[\tau_0, \tau_{i-1})}) = -u_n|_{[\tau_{i-1}, \tau_i)} = v_n|_{[\tau_{i-1}, \tau_i)}.$$

Therefore, we have $u_n(t) = -v_n(t)$. This gives us $\phi_n(1) = 0$. Hence, for every Γ^λ, $\Phi[1, \Gamma^\lambda, \Delta^\mu(\Gamma^\lambda)] = 0$, and this leads to $W^-_{\lambda,\mu}(0, 0) \le 0$. Since $g(x(1)) \ge 0$, we have $W^-_{\lambda,\mu}(0, 0) = 0$.

Now, we want to show that the upper value of the game $G_{\lambda,\mu}(0,0)$ is not less than $\frac{1}{4}\mu^2$. Fix any strategy Δ^μ for Player II. Let the partition points of Δ^μ be given by

$$\Pi_n(\Delta^\mu): \quad 0=\tau_0<\tau_1<\cdots<\tau_{n_0-1}\le\mu<\tau_{n_0}<\tau_{n_0+1}<\cdots<\tau_{p_n}=1.$$

Let $\Gamma^\lambda(\Delta^\mu)$ be a strategy for Player I depending on Δ^μ defined as follows. First $\Pi_n(\Gamma^\lambda(\Delta^\mu))=\Pi_n(\Delta^\mu)$. Let $v_n|_{[\tau_0,\tau_{n_0})}$ be the control function of Player II determined by Δ_n^μ.

Let $E(v_n|_{[\tau_0,\tau_{n_0})}\ge 0)$ denote the Lebesgue measure of the set $\{t\in[\tau_0,\tau_{n_0}): v_n|_{[\tau_0,\tau_{n_0})}(t)\ge 0\}$.

Case 1. If $E(v_n|_{[\tau_0,\tau_{n_0})}\ge 0)\ge\frac{1}{2}\tau_{n_0}$, then let $\Gamma_n^\lambda(\Delta^\mu)$ be the constant strategy corresponding to $u(t)\equiv 1$. Therefore, $\phi_n(1)=\int_0^1(u_n+v_n)\,d\tau\ge\int_{E(v_n|_{[\tau_0,\tau_{n_0})}\ge 0)}d\tau\ge\frac{1}{2}\tau_{n_0}\ge\mu/2$.

Case 2. If $E(v_n|_{[\tau_0,\tau_{n_0})}<0)\ge\frac{1}{2}\tau_{n_0}$, then let $\Gamma_n^\lambda(\Delta^\mu)$ be the constant strategy corresponding to $u(t)\equiv -1$. Therefore, $\phi_n(1)=\int_0^1(u_n+v_n)\,d\tau\le-\int_{E(v_n|_{[\tau_0,\tau_{n_0})}<0)}d\tau\le-\frac{1}{2}\tau_{n_0}\le-\mu/2$.

So in any case, we have for any n, $|\phi_n(1)|\ge\frac{1}{2}\tau_{n_0}$. By the definition of motion, we get $|\Phi(1)|\ge\frac{1}{2}\mu$. So we have $g(\Phi[1,\Gamma^\lambda(\Delta^\mu),\Delta^\mu])\ge\frac{1}{4}\mu^2$, for any Δ^μ. This gives us $W^+_{\lambda,\mu}(0,0)\ge\frac{1}{4}\mu^2$. Since $W^-_{\lambda,\mu}(0,0)=0<\frac{1}{4}\mu^2\le W^+_{\lambda,\mu}(0,0)$ if $\mu>0$, value does not exist.

Now, let us study the properties of upper and lower values as functions of the lags.

LEMMA 8. *Suppose that f and g satisfy Assumption* I′ *and that the Isaacs condition holds for f on $[t_0,T]\times\mathbb{R}^n$. Then*

$$W^-_{\lambda,0}(t_0,x_0)\le W^+_{\lambda,0}(t_0,x_0)\le W(t_0,x_0)\le W^-_{0,\mu}(t_0,x_0)\le W^+_{0,\mu}(t_0,x_0), \qquad \text{if } \lambda,\mu\ge 0,$$
$$W^-_{\lambda,0}(t_0,x_0)\le W^-_{\lambda,\mu}(t_0,x_0)\le W^+_{\lambda,\mu}(t_0,x_0)\le W^+_{0,\mu}(t_0,x_0), \qquad \text{if } \lambda,\mu\ge 0.$$

Proof. We only show that $W^+_{\lambda,0}(t_0,x_0)\le W(t_0,x_0)$. The others either are obvious or can be shown in the same way.

Let Γ^λ be a strategy for Player I for the game $G_{\lambda,0}(t_0,x_0)$. We will define a strategy Γ^0 of Player I for the game $G(t_0,x_0)$ corresponding to Γ^λ as follows. If

$$\Pi_n: \ t_0=\tau_0<\tau_1<\cdots<\tau_{p_n}=T$$

is the nth partition of $\Gamma^\lambda=\{\Gamma_n^\lambda\}$, let Π_n be the nth partition associated with Γ_n^0. Define

$$\Gamma^0_{n,1}(t)=\Gamma^\lambda_{n,1}(t) \qquad \text{for } t\in[\tau_0,\tau_1),$$
$$\Gamma^0_{n,2}(Y|_{[\tau_0,\tau_1)},Z|_{[\tau_0,\tau_1)})=\Gamma^\lambda_{n,2}(Y|_{[\tau_0,\tau_1)},Z|_{[\tau_0,\tau_1-\lambda)}) \qquad \text{for } t\in[\tau_1,\tau_2).$$

Note that if $\tau_1-\lambda<\tau_0$, then $\Gamma^\lambda_{n,2}(Y|_{[\tau_0,\tau_1)},Z|_{[\tau_0,\tau_1-\lambda)})=\Gamma^\lambda_{n,2}(Y|_{[\tau_0,\tau_1)})$. Similarly, we can define $\Gamma^0_{n,i}$ for $i\le 2\le p_n$. Let S denote the mapping just defined from the set of all possible Γ^λ into the set of all possible Γ^0, i.e., $S(\Gamma^\lambda)=\Gamma^0$. The mapping is one to one, but need not be onto. Notice that for any $(\Gamma^\lambda,\Delta^0)$, the nth stage outcome of $(\Gamma^\lambda,\Delta^0)$ equals the nth stage outcome of $(S(\Gamma^\lambda),\Delta^0)$. Therefore $\Phi[\,,t_0,x_0,\Gamma^\lambda,\Delta^0]=$

$\Phi[\,, t_0, x_0, S(\Gamma^\lambda), \Delta^0]$. Hence $g(\Phi[T, t_0, x_0, \Gamma^\lambda, \Delta^0]) = g(\Phi[T, t_0, x_0, S(\Gamma^\lambda), \Delta^0])$ for any $(\Gamma^\lambda, \Delta^0)$, which implies that for any fixed Δ^0,

$$\{g(\Phi[T, t_0, x_0, \Gamma^\lambda, \Delta^0]) : \text{ all possible } \Gamma^\lambda\} \subset \{g(\Phi[T, t_0, x_0, \Gamma^0, \Delta^0]) : \text{ all possible } \Gamma^0\}.$$

It follows that

$$\sup_{\Gamma^\lambda} g(\Phi[T, t_0, x_0, \Gamma^\lambda, \Delta^0]) \le \sup_{\Gamma^0} g(\Phi[T, t_0, x_0, \Gamma^0, \Delta^0])$$

for any Δ^0, and hence $W^+_{\lambda,0}(t_0, x_0) \le W(t_0, x_0)$. □

Let $v_0 = W^-(t_0, x_0)$ and let $C(v_0) = \{(\tau, \xi) : t_0 \le \tau \le T, \xi \in \mathbb{R}^n, W^-(\tau, \xi) \le v_0\}$. Let $\mathcal{X}$ be a compact set in $\mathbb{R}^{n+1}$ with (t_0, x_0) in its interior. By Gronwall's lemma and Assumption I(ii), there exists a constant K_0 such that any solution ϕ of

$$\frac{dx}{dt} = f(t, x, u(t), v(t)), \qquad x(t_1) = x_1 \tag{16}$$

with $(t_1, x_1) \in \mathcal{X}$ satisfies $|\phi(t)| \le K_0$ for $t_1 \le t \le T$, independent of the choice of controls u and v. Therefore, there exists a compact set $E \subset \mathbb{R}^{n+1}$ such that the set of solutions of (16) obtained as (t_1, x_1) ranges over $\mathcal{X}$ and u and v range over all possible controls is contained in E.

Since the game $G(t_0, x_0)$ has a saddle point, let us denote the saddle point by (Γ^*, Δ^*). Then we have $(t, \Phi[t, t_0, x_0, \Gamma^*, \Delta^*]) \in E$ and $(t, \Phi[t, t_0, x_0, \Gamma^*, \Delta^*]) \in C(v_0)$ for each $t_0 \le t \le T$. Hence $C(v_0) \cap E \cap H(t) \ne \emptyset$ where $t_0 \le t \le T$ and $H(t)$ is the hyperplane in $\mathbb{R}^{n+1}$ at t. Define $\tilde{C}(v_0) = C(v_0) \cap E$ and $S(t) = \tilde{C}(v_0) \cap H(t)$. We have $S(t) \ne \emptyset$ for each $t_0 \le t \le T$.

Now let us define an extremal strategy to the set $\tilde{C}(v_0)$ with Player II having a lag μ. Let $\phi(t, t_0, x_0, u, v) = x_0 + \int_{t_0}^t f(s, \phi(s), u(s), v(s))ds$ for $t_0 \le t \le T$. For $t_0 \le t \le t_0 + \mu$, let $V_e^\mu(\phi(t), t)$ be any fixed $z \in Z$. For $\mu + t_0 \le t$, define $t^* = t - \mu$, $x^* = \phi(t^*)$ and $x = \phi(t)$. If $(t^*, x^*) \in \tilde{C}(v_0)$, let $V_e^\mu(\phi(t), t)$ be any fixed $z \in Z$. If $(t^*, x^*) \notin \tilde{C}(v_0)$, let w^* be a point in $S(t^*)$ that is at minimal distance from $\phi(t^*)$. If w^* is not unique, we may select any such w^*. Let $s^* = x^* - w^*$. Define $V_e^\mu(\phi(t), t) = z^*$ where (y^*, z^*) is any saddle point of the local game at (t^*, x^*, s^*) with payoff $\langle s^*, f(t^*, x^*, y, z)\rangle$. We say that V_e^μ defined as above is extremal to $\tilde{C}(v_0)$. Let $\Delta_e^\mu = \Delta_e^\mu(V_e^\mu)$ be the corresponding feedback strategy (see [1] for definition of feedback strategy). Δ_e^μ so defined is indeed a strategy since at time t Player II bases his decision on the information available to him at time $t^* = t - \mu$.

LEMMA 9. *Let Assumption* I′ *and the Isaacs condition hold for* f. *Let* V_e^μ *be extremal to* $\tilde{C}(v_0)$ *and let* $\Delta_e^\mu = \Delta_e^\mu(V_e^\mu)$ *be the corresponding feedback strategy of Player II for the game* $G_{0,\mu}(t_0, x_0)$. *Then there exists a nonnegative function* η *defined for* $\mu \ge 0$ *such that* $\eta(\mu) \to 0$ *as* $\mu \to 0$ *and every motion* $\phi[\,, t_0, x_0, \Gamma^0, \Delta_e^\mu]$ *lies in* $C_{\eta(\mu)}(v_0) = \{(t, x) : t \in [t_0, T], x \in \mathbb{R}^n, \operatorname{dist}((t, x), C(v_0)) \le \eta(\mu)\}$.

Proof. We will actually show that every motion $\phi[\,, t_0, x_0, \Gamma^0, \Delta_e^\mu]$ lies in $\{(t, x) : \operatorname{dist}((t, x), S(t)) \le \eta(\mu)\}$. Let

$$M = \max\{|f(t, x, y, z)| : (t, x, y, z) \in E \times Y \times Z\}. \tag{17}$$

Let $\phi[\cdot] = \phi[\,, t_0, x_0, \Gamma^0, \Delta_e^\mu]$ be any motion in $\Phi[\,, t_0, x_0, \Gamma^0, \Delta_e^\mu]$ and let $\{\phi_n(\cdot)\} = \{\phi_n(\,, t_0, x_0, u_n, v_n)\}$ be a sequence of nth stage trajectories converging uniformly to

$\phi[\ , t_0, x_0, \Gamma^0, \Delta_e^\mu]$. If $\phi[\ , t_0, x_0, \Gamma^0, \Delta_e^\mu]$ lies entirely in $\tilde{C}(v_0)$, then we have nothing to prove. Assume that $\phi[\ , t_0, x_0, \Gamma^0, \Delta_e^\mu]$ does not lie entirely in $\tilde{C}(v_0)$. Let $t_n = \inf\{t \in [t_0, T] : (t, \phi_n(t)) \notin \tilde{C}(v_0)\}$. Since $\tilde{C}(v_0)$ is closed, $(t_n, \phi_n(t_n)) \in \tilde{C}(v_0)$.

Let $e(t)$ be the distance from $(t, \phi[t])$ to $S(t) = \tilde{C}(v_0) \cap H(t)$. Let $e_n(t)$ be the distance from $(t, \phi_n(t))$ to $S(t)$. Then $\lim_{n\to\infty} e_n(t) = e(t)$ for all $t \in [t_0, T]$.

Let $I_1, \ldots, I_{p_n}$, where $I_j = [\tau_{j-1}, \tau_j)$, $j = 1, 2, \ldots, p_n$, be the intervals of the nth partition Π_n. Let $\delta_n = \|\Pi_n\|$. Let $k_1 = k_1(n)$ denote the integer such that $t_n + \mu \in I_{k_1} = [\tau_{k_1 - 1}, \tau_{k_1})$.

We first show that for $t_n \le t \le \tau_{k_1}$, $e_n(t) \le 2M(\delta_n + \mu)$. Let $t \in [t_n, \tau_{k_1}]$. Since $(t_n, \phi_n(t_n)) \in C(v_0)$, it follows from [1, Lem. 8.3] that there exists a relaxed control ζ such that the relaxed trajectory $\psi(\ , t_n, \phi_n(t_n), u_n, \zeta)$ has the property that $(t, \psi(t)) \in C(v_0)$. Hence $e_n(t) \le |\psi(t) - \phi_n(t)| = |\int_{t_n}^t f(s, \psi(s), u_n(s), \eta(s))ds - \int_{t_n}^t f(s, \phi_n(s), u_n(s), v_n(s))ds| \le 2M(\delta_n + \mu)$, for $t_n \le t \le \tau_{k_1}$, where M is as in (17). Define

$$\alpha_0 = 2M(\delta_n + \mu). \tag{18}$$

Then $e_n(t) \le \alpha_0$, for $t_n \le t \le \tau_{k_1}$.

Now suppose $t \in [\tau_k, \tau_{k+1})$ with $k \ge k_1$. Let $t^* = \tau_k - \mu$, let $x^* = \phi_n(t^*)$, and suppose $(t^*, x^*) \notin \tilde{C}(v_0)$. Let w^* be the point in $S(t^*)$ selected as being at minimum distance from x^* in $S(t^*)$ in the definition of V_e^μ. Let $s^* = x^* - w^*$ and let y^* be any point in Y such that (y^*, z^*) is a saddle point for the local game (t^*, x^*, s^*) with payoff $\langle s^*, f(t^*, x^*, y, z)\rangle$.

By [1, Lemma 8.3], there exists a relaxed control ζ such that the corresponding relaxed trajectory $\psi(\ ) = \psi(\ , t^*, w^*, y^*, \zeta)$ has the property that $(t, \psi(t)) \in C(v_0)$. Therefore, we have that $e_n(t) \le |\phi_n(t, t^*, x^*, u_n, z^*) - \psi(t, t^*, w^*, y^*, \zeta)|$. Let $\rho_n(t) = |\phi_n(t) - \psi(t)|$, then

$$\begin{aligned}\frac{d\rho_n^2(t)}{dt} &= 2\langle f(t, \phi_n(t), u_n, z^*) - f(t, \psi(t), y^*, \zeta),\ \phi_n(t) - \psi(t)\rangle \\ &= 2\langle f(t^*, x^*, u_n, z^*) - f(t^*, x^*, y^*, \zeta) + \Delta^\phi f - \Delta^\psi f, \\ &\qquad (x^* - w^*) + \phi_n(t) - x^* - \psi(t) + w^*\rangle,\end{aligned} \tag{19}$$

where

$$\Delta^\phi f = f(t, \phi_n(t), u_n, z^*) - f(t^*, x^*, u_n, z^*)$$

and

$$\Delta^\psi f = f(t, \psi(t), y^*, \zeta) - f(t^*, x^*, y^*, \zeta).$$

Since $t - t^* \le \delta_n + \mu$, we have that $|\phi_n(t) - x^*| \le M(\delta_n + \mu)$ and $|\psi(t) - w^*| \le M(\delta_n + \mu)$. Since f is continuous, there exists a function $\epsilon : [0, \infty) \to [0, \infty)$ such that $\lim_{s\to 0}\epsilon(s) = 0$ and $|f(t_1, x, y, z) - f(t_2, x, y, z)| \le \epsilon(|t_1 - t_2|)$ for (t_1, x, y, z) and (t_2, x, y, z) in the compact set $E \times Y \times Z$. So, $|\Delta^\phi f| \le |f(t, \phi_n(t), u_n, z^*) - f(\tau_k, \phi_n(t), u_n, z^*)| + |f(\tau_k, \phi_n(t), u_n, z^*) - f(t^*, \phi_n(t), u_n, z^*)| + |f(t^*, \phi_n(t), u_n, z^*) - f(t^*, x^*, u_n, z^*)| \le \epsilon(\delta_n) + \epsilon(\mu) + KM(\delta_n + \mu)$. Similarly, $|\Delta^\psi f| \le \epsilon(\delta_n) + \epsilon(\mu) + KM(\delta_n + \mu) + K|x^* - w^*|$, where K is the Lipschitz constant in Assumption I′. Since $e_n(t)$ is uniformly bounded for $t \in [t_0, T]$, by applying these estimates to (19), we get $(d\rho_n^2(t)/dt) \le 2Ke_n^2(\tau_k - \mu) + O(\delta_n) + O(\mu)$. Integrating from $\tau_k - \mu = t^*$

to t on both sides of the above inequality and noting that $e_n(t) \le \rho_n(t)$ and that $e_n(\tau_k - \mu) = \rho_n(\tau_k - \mu)$, we get that for $\tau_k \le t \le \tau_{k+1}$,

$$e_n^2(t) \le e_n^2(\tau_k - \mu)\big(1 + 2K(\delta_n + \mu)\big) + \big(O(\delta_n) + O(\mu)\big)(\delta_n + \mu). \tag{20}$$

Since $t_n < \tau_{k_1} - \mu \le \tau_{k_1}$, $e_n^2(\tau_{k_1} - \mu) \le \alpha_0^2$. So, for $t_n \le t \le \tau_{k_1+1}$,

$$e_n^2(t) \le \alpha_0^2\big(1 + 2K(\delta_n + \mu)\big) + \big(O(\delta_n) + O(\mu)\big)(\delta_n + \mu) \triangleq \alpha_1^2. \tag{21}$$

If $(t^*, x^*) \in \tilde{C}(v_0)$, the argument used to show (18) will give that $e_n^2(t) \le \alpha_0^2$. Hence (21) holds.

Let $k_2 > k_1 + 1$ be such that $\tau_{k_2+1} - \mu > \tau_{k_1+1}$ and if $k \le k_2$, then $\tau_k - \mu \le \tau_{k_1+1}$, i.e., τ_{k_2} is the greatest partition point in Π_n that is less than or equal to $\tau_{k_1+1} + \mu$. Then for any $k_1 + 1 < k \le k_2$, it follows from (21) that $e_n^2(\tau_k - \mu) \le \alpha_1^2$. Therefore, (20) gives that for $\tau_{k_1+1} \le t \le \tau_{k_2+1}$, $e_n^2(t) \le \alpha_1^2\big(1 + 2K(\delta_n + \mu)\big) + \big(O(\delta_n) + O(\mu)\big)(\delta_n + \mu)$. Together with (21), we get that for $t_n \le t \le \tau_{k_2+1}$, $e_n^2(t) \le \alpha_0^2\big(1 + 2K(\delta_n + \mu)\big)^2 + \big(O(\delta_n) + O(\mu)\big)(\delta_n + \mu)\sum_{i=0}^{1}\big(1 + 2K(\delta_n + \mu)\big)^i \triangleq \alpha_2^2$.

Suppose we have defined k_j. Let $k_{j+1} > k_j + 1$ be such that $\tau_{k_{j+1}+1} - \mu > \tau_{k_j+1}$ and if $k \le k_{j+1}$, then $\tau_k - \mu \le \tau_{k_j+1}$, i.e., $\tau_{k_{j+1}}$ is the greatest partition point in Π_n that is less than or equal to $\tau_{k_j+1} + \mu$.

Suppose for $t_n \le t \le \tau_{k_j+1}$, we have $e_n^2(t) \le= \alpha_0^2\big(1 + 2K(\delta_n + \mu)\big)^j + \big(O(\delta_n) + O(\mu)\big)(\delta_n + \mu)\sum_{i=0}^{j_1}\big(1 + 2K(\delta_n + \mu)\big)^i \triangleq \alpha_j^2$. Then, it is clear that for $t_n \le t \le \tau_{k_{j+1}+1}$, $e_n^2(t) \le \alpha_j^2\big(1 + 2K(\delta_n + \mu)\big) + \big(O(\delta_n) + O(\mu)\big) \le \alpha_0^2\big(1 + 2K(\delta_n + \mu)\big)^{j+1} + \big(O(\delta_n) + O(\mu)\big)(\delta_n + \mu)\sum_{i=0}^{j}\big(1 + 2K(\delta_n + \mu)\big)^i$.

Note that $\tau_{k_{j+1}+1} - \tau_{k_j+1} > \mu$. So, we can only iterate for at most $[(T - t_0)/\mu]$ steps. Therefore, for $t_n \le t \le T$,

$$e_n^2(t) \le \alpha_0^2\big(1+2K(\delta_n+\mu)\big)^{[(T-t_0)/\mu]} + \big(O(\delta_n)+O(\mu)\big)(\delta_n+\mu)\sum_{i=0}^{[(T-t_0)/\mu]-1}\big(1+2K(\delta_n+\mu)\big)^i.$$

Letting $n \to \infty$ and using (18), we get

$$\begin{aligned} e^2(t) &\le 4M^2\mu^2(1 + 2K\mu)^{[(T-t_0)/\mu]} + O(\mu)\mu\sum_{i=0}^{[(T-t_0)/\mu]-1}(1 + 2K\mu)^i \\ &\le 4M^2\mu^2 e^{2K(T-t_0)} + O(\mu)\frac{e^{2K(T-t_0)} - 1}{2K} \triangleq \eta^2(\mu). \end{aligned}$$

□

LEMMA 10. *Let Assumption* I′ *hold. If* $\Phi[\,, t_0, x_0, \Gamma^0, \Delta_e^\mu] \subset C_{\eta(\mu)}(v_0)$ *for all* Γ^0, *there exists a nonnegative function* $\tilde{\eta}_0$ *of* μ *having the following properties:*

(1) $\tilde{\eta}_0(\mu) \downarrow 0$ *as* $\mu \downarrow 0$,

(2) $W_{0,\mu}^+(t_0, x_0) \le W(t_0, x_0) + \tilde{\eta}_0(\mu)$.

Proof. It follows from the continuity of g, from Assumption I′, and from the definition of motion that for any given $\epsilon > 0$, there exists $\delta > 0$ such that for any two pairs of strategies (Γ, Δ) and (Γ', Δ') and any motions $\phi[\,, t_1, x_1, \Gamma, \Delta]$, $\phi[\,, t_1, x_1, \Gamma', \Delta']$ the inequality

$$|g(\phi[T, t_1, x_1, \Gamma, \Delta]) - g(\phi[T, t_1, x_1, \Gamma', \Delta'])| < \epsilon$$

holds for any $(t_1, x_1) \in E$ whenever $|T - t_1| < \delta = \delta(\epsilon)$.

Let $\phi[\ , t_0, x_0, \Gamma^0, \Delta_\epsilon^\mu]$ be a motion $\subset C_{\eta(\mu)}(v_0)$. Let t_1 satisfy $|T - t_1| < \delta$ and let $x_1 = \phi[t_1, t_0, x_0, \Gamma^0, \Delta_\epsilon^\mu]$. By Lemma 8, there exists a w_1 such that $(t_1, w_1) \in C(v_0)$ and $|x_1 - w_1| \leq \eta(\mu)$. It is easy to see that for any Γ and any Δ

$$|\phi[T, t_1, x_1, \Gamma, \Delta] - \phi[T, t_1, w_1, \Gamma, \Delta]| \leq |x_1 - w_1| + 2M\delta \leq \eta(\mu) + 2M\delta. \tag{22}$$

From the continuity of g and (22), there exists a function $\sigma \geq 0$ satisfying $\sigma(s) \to 0$ as $s \to 0$ such that $|g(\phi[T, t_1, x_1, \Gamma, \Delta]) - g(\phi[T, t_1, w_1, \Gamma, \Delta])| \leq \sigma(\eta(\mu)) + \sigma(2M\delta)$. Note that the segment of the motion $\phi[t_0, x_0, \Gamma^0, \Delta_\epsilon^\mu]$ on the interval $[t_1, T]$ is again a motion, say $\phi[\ , t_1, x_1, \Gamma', \Delta']$. Therefore $g(\phi[T, t_0, x_0, \Gamma^0, \Delta_\epsilon^\mu]) = g(\phi[T, t_1, x_1, \Gamma', \Delta'])$.

Since $(t_1, w_1) \in C(v_0)$, it follows from [1, Lemma 8.2] that there exists a strategy $\Delta(\Gamma')$ in the game without lag and a motion $\phi[\ , t_1, w_1, \Gamma', \Delta(\Gamma')]$ such that

$$g(\phi[T, t_1, w_1, \Gamma', \Delta(\Gamma')]) < v_0 + \epsilon.$$

Hence

$$\begin{aligned}
&g(\phi[T, t_0, x_0, \Gamma^0, \Delta_\epsilon^\mu]) \\
&= g(\phi[T, t_1, x_1, \Gamma', \Delta']) - g(\phi[T, t_1, x_1, \Gamma', \Delta(\Gamma')]) + g(\phi[T, t_1, x_1, \Gamma', \Delta(\Gamma')]) \\
&\quad - g(\phi[T, t_1, w_1, \Gamma', \Delta(\Gamma')]) + g(\phi[T, t_1, w_1, \Gamma', \Delta(\Gamma')]) \\
&\leq \epsilon + \sigma(\eta(\mu)) + \sigma(2M\delta) + v_0 + \epsilon \\
&= 2\epsilon + v_0 + \sigma(\eta(\mu)) + \sigma(2M\delta),
\end{aligned}$$

From above inequality and the arbitrariness of ϵ and ϕ, we get $g(\Phi[T, t_0, x_0, \Gamma^0, \Delta_\epsilon^\mu]) \leq v_0 + \tilde{\eta}(\mu)$ with $\tilde{\eta}(\mu) = \sigma(\eta(\mu))$. The lemma follows from this inequality. □

Following the same ideas as used in the proofs of Lemmas 9 and 10, we can obtain the following result.

LEMMA 11. *If Assumption* I′ *and the Isaacs condition hold for* f, *there exists a nonnegative function* $\tilde{\eta}_1$ *of* λ *having the following properties:*

(1) $\tilde{\eta}_1(\lambda) \downarrow 0$ *as* $\lambda \downarrow 0$,

(2) $W^-_{\lambda,0}(t_0, x_0) \geq W(t_0, x_0) - \tilde{\eta}_1(\lambda)$.

THEOREM 12. *If Assumption* I′ *and the Isaacs condition hold for* f, *then*

$$0 \leq W^+_{\lambda,\mu}(t_0, x_0) - W^-_{\lambda,\mu}(t_0, x_0) \leq \tilde{\eta}_0(\lambda) + \tilde{\eta}_1(\mu),$$

where $\tilde{\eta}_i$, $i = 0, 1$, *are as in Lemmas* 10 *and* 11, *and hence*

$$\lim_{\substack{\lambda \to 0 \\ \mu \to 0}} W^\pm_{\lambda,\mu}(t_0, x_0) = W(t_0, x_0).$$

The proof is an immediate consequence of Lemmas 8, 10, and 11.

2.2. Differential games of generalized pursuit and evasion. The game is governed by (1) with payoff (3) and terminal set $\mathcal{T}$.

Let (t_0, x_0) with $(t_0, x_0) \notin \mathcal{T}$ be an initial point of the game. A strategy Γ^λ for Player I with lag $\lambda \geq 0$ on the interval $[t_0, T]$ and a strategy Δ^μ for Player II with lag $\mu \geq 0$ on the interval $[t_0, T]$ are defined as in §2.1, where games of fixed duration with information lags are defined.

Corresponding to $(\Gamma^\lambda, \Delta^\mu)$, we obtain a sequence of controls (u_n, v_n) and a sequence of nth stage trajectories $\hat{\phi}_n(\ ) \equiv (\phi_n^0, \phi_n)$ satisfying

$$\phi_n^{0\prime}(t) = f^0(t, \phi_n(t), u_n(t), v_n(t)), \qquad \phi_n^0(t_0) = x_{0n}^0,$$
$$\phi_n'(t) = f(t, \phi_n(t), u_n(t), v_n(t)), \qquad \phi_n(t_0) = x_{0n},$$

with $(x_{0n}^0, x_{0n}) \to (0, x_0)$. A motion $\hat{\phi}[\ , t_0, x_0, \Gamma^\lambda, \Delta^\mu]$ is a uniform limit of nth stage trajectories $\hat{\phi}_n(\ , t_0, x_0, u_n, v_n)$.

By the terminal time t_{f_n} of the nth stage trajectory $\hat{\phi}_n(\ )$ we mean the first time such that $(t, \phi(t)) \in \mathcal{T}$. The terminal time t_f of a motion $\hat{\phi}[\ , t_0, x_0, \Gamma^\lambda, \Delta^\mu]$ is defined similarly.

The payoff $P(t_0, x_0, \Gamma^\lambda, \Delta^\mu)$ resulting from a pair of strategies $(\Gamma^\lambda, \Delta^\mu)$ is defined as

$$P(t_0, x_0, \Gamma^\lambda, \Delta^\mu) \equiv \cup\{g(t_f, \phi[t_f, t_0, x_0, \Gamma^\lambda, \Delta^\mu]) + \phi^0[t_f, t_0, x_0, \Gamma^\lambda, \Delta^\mu]\}$$

where the union is taken over all motions $\hat{\phi}[\ , t_0, x_0, \Gamma^\lambda, \Delta^\mu]$ resulting from $(\Gamma^\lambda, \Delta^\mu)$. We designate the game just defined as $G_{\lambda,\mu}(t_0, x_0)$. We define the upper value of the game by $W_{\lambda,\mu}^+(t_0, x_0)$ and the lower value by $W_{\lambda,\mu}^-(t_0, x_0)$. Thus

$$W_{\lambda,\mu}^+(t_0, x_0) = \inf_{\Delta^\mu} \sup_{\Gamma^\lambda} P(t_0, x_0, \Gamma^\lambda, \Delta^\mu),$$
$$W_{\lambda,\mu}^-(t_0, x_0) = \sup_{\Gamma^\lambda} \inf_{\Delta^\mu} P(t_0, x_0, \Gamma^\lambda, \Delta^\mu).$$

An argument similar to that used to prove Lemma 8 gives the following lemma.

LEMMA 13. *Let Assumption* I *and either Assumption* II *or* II′ *hold. If the Isaacs condition holds, then*

$$W_{\lambda,0}^-(t_0, x_0) \le W_{\lambda,0}^+(t_0, x_0) \le W(t_0, x_0) \le W_{0,\mu}^-(t_0, x_0) \le W_{0,\mu}^+(t_0, x_0);$$

$$W_{\lambda,0}^-(t_0, x_0) \le W_{\lambda,\mu}^-(t_0, x_0) \le W_{\lambda,\mu}^+(t_0, x_0) \le W_{0,\mu}^+(t_0, x_0),$$

where $W(t_0, x_0)$ is the value of game of generalized pursuit and evasion without lags.

Let $\epsilon_k = 1/k$. We consider a fixed-duration game $G_{\lambda,\mu}(t_0, x_0, \epsilon_k)$ as in [2], which we henceforth will write as $G_{\lambda,\mu}(t_0, x_0, k)$, and we will let $W_{\lambda,\mu}^\pm(t_0, x_0, k)$ denote the upper and lower values of the game $G_{\lambda,\mu}(t_0, x_0, k)$.

LEMMA 14. *Let Assumption* I *and either Assumption* II *or* II′ *hold. If $\mathcal{C}$ is a compact set in $(([t_0, \infty) \times \mathbb{R}^n) - \mathcal{T}) \cup (\partial\mathcal{T})$, then there exists a constant $c > 0$ such that for all k sufficiently large and all (t_0, x_0) in $\mathcal{C}$, $W_{\lambda,0}^-(t_0, x_0) \ge W_{\lambda,0}^-(t_0, x_0, k) - c/k$.*

The proof of this lemma is identical to the proof of [2, Lem. 4.3], which does not involve the structure of the strategies Γ for Player I. The proof requires the construction of a strategy Δ^* for Player II. In the present lemma, Player II does not have an information lag, so that the strategy Δ^* constructed for Player II in the present case will be identical to the construction of the strategy Δ^* in [2]. The following lemma is similar.

LEMMA 15. *Let Assumption* I *and either Assumption* II *or* II′ *hold. If $\mathcal{C}$ is a compact set in $(([t_0, \infty) \times \mathbb{R}^n) - \mathcal{T}) \cup (\partial\mathcal{T})$, then there exists a constant $c > 0$ such that for all k sufficiently large and all (t_0, x_0) in $\mathcal{C}$, $W_{0,\mu}^+(t_0, x_0) \le W_{0,\mu}^+(t_0, x_0, k) + c/k$.*

For $\tau_0 \in [t_0, T]$ and $\hat{\xi}_0 \in \mathbb{R}^{n+1}$, as in [2], we define the game $\hat{G}_{\lambda,\mu}(\tau_0, \hat{\xi}_0)$, which is governed by

$$\frac{dx}{dt} = f(t, x, u(t), v(t)), \qquad x(\tau_0) = \xi_0$$

with payoff $\xi_0^0 + \int_{\tau_0}^{t_f} f^0(s, \phi(s), u(s), v(s))ds$ and terminal set $\hat{T} = \mathbb{R} \times T$. Let $\hat{W}^{\pm}_{\lambda,\mu}(\tau_0, \hat{\xi}_0)$ denote the upper and lower values of $\hat{G}_{\lambda,\mu}(\tau_0, \hat{\xi}_0)$. Then $\hat{W}^{\pm}_{\lambda,\mu}(\tau_0, \hat{\xi}_0) = \xi_0^0 + W^{\pm}_{\lambda,\mu}(\tau_0, \xi_0)$. Let $\hat{F}_1 = \mathbb{R} \times F_1$ and let $\hat{F} = \mathbb{R} \times F$.

For $\epsilon_k = 1/k$, consider the fixed duration game $\hat{G}_{\lambda,\mu}(\tau_0, \hat{\xi}_0, k)$ as defined in §1 of this paper. Let $\hat{W}^{\pm}_{\lambda,\mu}(\tau_0, \hat{\xi}_0, k)$ denote the upper and lower values of $\hat{G}_{\lambda,\mu}(\tau_0, \hat{\xi}_0, k)$.

Let $\nu_{0,k} = \hat{W}^{+}_{0,0}(t_0, \hat{x}_0, k)$ with $\hat{x}_0 = (0, x_0)$. Let $C(\nu_{0,k}) = \{(t, \hat{x}) : t_0 \le t \le T, \hat{x} \in \mathbb{R}^{n+1}, \hat{W}^{+}_{0,0}(t, \hat{x}, k) \le \nu_{0,k}\}$ and let $C_\alpha(\nu_{0,k}) = \{(t, \hat{x}) : t_0 \le t \le T, \hat{x} \in \mathbb{R}^{n+1}, \operatorname{dist}((t, \hat{x}), C(\nu_{0,k})) \le \alpha\}$.

Let $\mathcal{X}$ be a compact set in $\mathbb{R}^{n+2}$. By Gronwall's lemma and Assumption I(ii), there exists a compact set $E \subset \mathbb{R}^{n+2}$ such that the set of solutions of

$$\frac{d\hat{x}}{dt} = \hat{f}(t, \hat{x}, u(t), v(t)), \qquad \hat{x}(t_1) = \hat{x}_1$$

obtained as $(t_1, \hat{x}_1)$ ranges over $\mathcal{X}$ and u and v range over all possible controls is contained in E. Let $\tilde{C}(\nu_{0,k}) = C(\nu_{0,k}) \cap E$. Let $M = \max|\hat{f}(t, \hat{x}, y, z)|$ for all $(t, \hat{x}, y, z) \in E \times Y \times Z$.

Let $\mathcal{B} = E \cap \hat{\partial}\hat{F}$ where $\hat{\partial}\hat{F} = (\partial\hat{F}) \cap (\partial\hat{F}_1)$. Since E is compact, $\mathcal{B}$ is a compact subset of $\hat{\partial}\hat{F}$. Let $\epsilon_0 = \epsilon_0(\mathcal{B})$ and $c_0 = c_0(\mathcal{B})$ be determined as in Lemma 2. Let $\mathcal{N}_\alpha(\mathcal{B}) = \{(t, \hat{x}) : t_0 \le t \le T, \hat{x} \in \mathbb{R}^{n+1}, \operatorname{dist}\big((t, \hat{x}), \mathcal{B}\big) \le \alpha\}$.

LEMMA 16. *Let Assumption* I *and either Assumption* II *or* II′ *hold and let the Isaacs condition* $(\hat{I})$ *hold for* $\hat{f}$ *(as defined in* §1*). Let* V_e^μ *be extremal to* $\tilde{C}^{(k)}(\nu_{0,k})$ *and let* $\Delta_e^\mu = \Delta_e^\mu(V_e^\mu)$ *be the corresponding feedback strategy of Player* II *for game* $\hat{G}_{0,\mu}(t_0, \hat{x}_0, k)$ *with initial point* $(t_0, \hat{x}_0)$*. Then there exist a* $\mu_0 > 0$ *and a nonnegative function* η *defined for* $0 \le \mu \le \mu_0$ *such that* $\eta(\mu) \to 0$ *as* $\mu \to 0$ *and having the following property. For any* Γ^0 *and any* $0 \le \mu \le \mu_0$*, if* t_f *is the terminal time for* $\hat{\phi}[\ , t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]$*, then there is a* t_{f_ψ}*,* $t_0 \le t_{f_\psi} \le t_f$ *such that* $t_f - t_{f_\psi} \le 1/c_0\eta(\mu)$ *and such that*

$$(t, \hat{\phi}[t, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]) \in C^{(k)}_{\eta(\mu)}(\nu_{0,k}) \cap H(t) \quad \text{for } t_0 \le t \le t_{f_\psi}.$$

Proof. Let Γ^0 be any strategy for Player I, let $\hat{\phi}[\ , t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]$ be any motion resulting from strategies (Γ^0, Δ_e^μ) and let $\{\hat{\phi}_{k,n}(\ , t_0, \hat{x}_0, u_n, v_0)\}$ be a sequence of nth stage trajectories converging uniformly to $\hat{\phi}[\ , t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]$. Let t_{f_n} be the terminal time of $\hat{\phi}_{k,n}(\ , t_0, \hat{x}_0, u_n, v_n)$. Note that the terminal time of $\hat{\phi}_{k,n}(\ , t_0, \hat{x}_0, u_n, v_n)$ does not depend on k. Let $\hat{e}_{k,n}(t)$ be the distance from

$$(t, \hat{\phi}_{k,n}(t, t_0, \hat{x}_0, u_n, v_n)) \quad \text{to } \tilde{C}^{(k)}(\nu_{0,k}) \cap H(t)$$

and let $\hat{e}_k(t)$ be the distance from

$$(t, \hat{\phi}[t, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]) \quad \text{to } \tilde{C}^{(k)}(\nu_{0,k}) \cap H(t).$$

Let $I_1, I_2, \ldots, I_{p_n}$ be the intervals of the nth partition Π_n. Let $\delta_n = \|\Pi_n\| \le L/n$. Let t_n be the first time at which $\hat{\phi}_{k,n}(t, t_0, \hat{x}_0, u_n, v_n)$ gets out of the set $C^{(k)}(\nu_{0,k})$. Let $j_1 = j_1(n)$ denote the integer such that $t_n + \mu \in I_{j_1} = [\tau_{j_1-1}, \tau_{j_1})$. Consider $t \in [t_n, t_{f_n}]$.

Case 1. Assume that $(t_n+\mu, \hat{\phi}_{k,n}(t_n+\mu, t_0, \hat{x}_0, u_n, v_n)) \in \hat{T}$. Then $t_{f_n} \le t_n+\mu$. Consider $t_n \le t \le T_{f_n}$. Since $(t_n, \hat{\phi}_{k,n}(t_n)) \in C^{(k)}(\nu_{0,k})$, according to [1, Lem. 8.3], there exists a relaxed control ζ_t such that the relaxed trajectory $\hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)$ has the property that $(t, \hat{\psi}_k(t, t_n, \hat{\psi}_{k,n}(t_n), u_n, \zeta_t)) \in C^{(k)}(\nu_{0,k})$.

If for every $t \in [t_n, t_{f_n}]$, we have $(t, \hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)) \notin \hat{T}$, then we get that for $t_n \le t \le t_{f_n}$

$$\begin{aligned}
\hat{e}_{k,n}(t) &\le |\hat{\phi}_{k,n}(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, v_n) - \hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)| \\
&\le |\hat{\phi}_{k,n}(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, v_n) - \hat{\phi}_{k,n}(t_n)| + |\hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t) - \hat{\phi}_{k,n}(t_n)| \\
&\le 2M(t_{f_n} - t_n) \\
&\le 2M(t_n + \mu - t_n) \\
&= 2M\mu.
\end{aligned}$$

If for some $t \in [t_n, t_{f_n}]$ every relaxed control ζ_t having the property that

$$(t, \hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)) \in C^{(k)}(\nu_{0,k})$$

satisfies $(t, \hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)) \in \hat{T}$, let $t_{f_{\psi_n}}$ be the infimum of such t. Then we have that for $t_n \le t \le t_{f_{\psi_n}}$, $\hat{e}_{k,n}(t) \le 2M\mu$. An argument as in the proof of Lemma 16 gives that $(t_{f_{\psi_n}}, \hat{\phi}_{k,n}(t_{f_{\psi_n}})) \in \mathcal{N}_{2M\mu}(\mathcal{B})$. Since $2M\mu \le \epsilon_0$ for sufficiently small μ, we have $t_{f_n} - t_{f_{\psi_n}} \le (1/c_0)2M\mu$.

Case 2. Assume that $(t_n+\mu, \hat{\phi}_{k,n}(t_n+\mu)) \notin \hat{T}$. Hence $(t_n, \hat{\phi}_{k,n}(t_n)) \notin \hat{T}$ by Corollary 5. Consider $t_n \le t \le \tau_{j_1}$. Since $(t_n, \hat{\phi}_{k,n}(t_n)) \in C^{(k)}(\nu_{0,k})$, it follows from [1, Lem. 8.3] that there is a relaxed control ζ_t such that the relaxed trajectory $\hat{\psi}_k(\ , t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)$ has the property that $(t, \hat{\psi}_k(t)) \in C^{(k)}(\nu_{0,k})$.

If for some $t \in [t_n, \tau_{j_1}]$, all the relaxed controls ζ_t satisfying the above property also satisfy $(t, \hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)) \in \hat{T}$, then let $t_{f_{\psi_n}}$ be the infimum of such t. Then an argument similar to that in Case 1 gives that $\hat{e}_{k,n}(t) \le 2M(\delta_n+\mu)$ for $t_n \le t \le t_{f_{\psi_n}}$ and $(t_{f_{\psi_n}}, \hat{\phi}_{k,n}(t_{f_{\psi_n}})) \in \mathcal{N}_{2M(\delta_n+\mu)}(\mathcal{B})$. Hence for sufficiently small μ and sufficiently large n, $t_{f_n} - t_{f_{\psi_n}} \le (1/c_0)2M(\delta_n+\mu)$.

If for every $t \in [t_n, \tau_{j_1}]$ there is a relaxed control ζ_t that not only has the property that

$$(t, \hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)) \in C^{(k)}(\nu_{0,k}),$$

but also has the property that $(t, \hat{\psi}_k(t, t_n, \hat{\phi}_{k,n}(t_n), u_n, \zeta_t)) \notin \hat{T}$, then we have $\hat{e}_{k,n}(t) \le 2M(\delta_n+\mu)$ for $t \in [t_n, \tau_{j_1}]$.

Consider $\tau_{j_1} \le t \le \tau_{j_1+1}$. Since $\tau_{j_1} - \mu \le \tau_{j_1}$, $\hat{e}_{k,n}(\tau_{j_1} - \mu) \le 2M(\delta_n+\mu)$. Let $t^* = \tau_{j_1} - \mu$, let $\hat{x}^* = \hat{\phi}_{k,n}(t^*)$ and let $\hat{w}^*$ be the point in $S(t^*) = C^{(k)}(\nu_{0,k}) \cap H(t^*)$ selected as being at minimum distance from $\hat{x}^*$ in the definition of V_e^{μ}. Then $(t^*, \hat{w}^*) \in C^{(k)}(\nu_{0,k})$. If $(t^*, \hat{w}^*) \in \hat{T}$, then since $\hat{e}_{k,n}(t^*) \le 2M(\delta_n+\mu)$, $(t^*, \hat{\phi}_{k,n}(t^*)) \in \mathcal{N}_{2M(\delta_n+\mu)}(\mathcal{B})$. Hence $0 \le t_{f_n} - t^* \le (1/c_0)2M(\delta_n+\mu)$ if $n \ge 2L/\epsilon_0$ and $\mu \le \epsilon_0/2M$, and $\hat{e}_{k,n}(t) \le 2M(\delta_n+\mu)$ for $t_0 \le t \le t^* = \tau_{j_1} - \mu$. If $(t^*, \hat{w}^*) \notin \hat{T}$, then the analysis in the proof of Lemma 9 gives that

$$\hat{e}_{k,n}^2(t) \le \hat{e}_{k,n}^2(\tau_{j_1} - \mu)(1 + 2K(\delta_n+\mu)) + (O(\delta_n) + O(\mu))(\delta_n+\mu).$$

If $(t, \hat{\psi}(t)) \in \hat{T}$ for some $t \in [\tau_{j_1}, \tau_{j_1+1}]$, then if we denote the terminal time of $\hat{\psi}(\ , t^*, \hat{w}^*, u_{k,n}, \zeta)$ as $t_{f_{\psi_n}}$, we have $t^* = \tau_{j_1} - \mu \le t_{f_{\psi_n}} \le \tau_{j_1+1}$. For $t_n \le t \le t_{f_{\psi_n}}$,

$$\begin{aligned}\hat{e}^2_{k,n}(t) &\le \hat{e}^2_{k,n}(\tau_{j_1} - \mu)(1 + 2K(\delta_n + \mu)) + (O(\delta_n) + O(\mu))(\delta_n + \mu) \\ &\le 4M^2(\delta_n + \mu)^2(1 + 2K(\delta_n + \mu)) + (O(\delta_n) + O(\mu))(\delta_n + \mu) \\ &= \alpha_1^2(\delta_n + \mu).\end{aligned}$$

Since $\hat{e}_{k,n}(t_{f_{\psi_n}}) \le \alpha_1(\delta_n + \mu)$ and $\alpha_1(\delta_n + \mu) \to 0$ as $\delta_n + \mu \to 0$, $(t_{f_{\psi_n}}, \hat{\phi}_{k,n}(t_{f_{\psi_n}})) \in \mathcal{N}_{\alpha_1(\delta_n+\mu)}(\mathcal{B})$ and hence $0 \le t_{f_n} - t_{f_{\psi_n}} \le (1/c_0)\alpha_1(\delta_n + \mu)$.

Let $j_2 > j_1 + 1$ be such that τ_{j_2} is the greatest partition point in Π_n that is less than or equal to $\tau_{j_1+1} + \mu$. Let $j_m > j_{m-1} + 1$ be such that τ_{j_m} is the greatest partition point in Π_n that is less than or equal to $\tau_{j_{m-1}} + \mu$, for $m = 1, 2, \ldots$. For $t_n \le t \le \tau_{j_m+1}$, an analysis similar to that used in the case when $\tau_{j_1} \le t \le \tau_{j_1+1}$ shows that there exists a $t_{f_{\psi_n}} \in [t_0, t_{f_n}]$ such that $0 \le t_{f_n} - t_{f_{\psi_n}} \le (1/c_0)\alpha(\delta_n + \mu)$ and such that for $t_0 \le t \le t_{f_{\psi_n}}$, $\hat{e}_{k,n}(t) \le \alpha(\delta_n + \mu)$ where $\alpha^2(\delta_n + \mu) = 4M^2(\delta_n + \mu)^2(1+2K(\delta_n+\mu))^{[(T-t_0)/\mu]} + (O(\delta_n)+O(\mu))(\delta_n+\mu)\sum_{i=0}^{[(T-t_0)/\mu]-1}(1+2K(\delta_n+\mu))^i$. Let $t_{f_\psi} = \inf \lim_{n\to\infty} t_{f_{\psi_n}}$. Since $\lim_{n\to\infty} t_{f_n} = t_f$, on letting $n \to \infty$, we get that $0 \le t_f - t_{f_\psi} \le (1/c_0)\alpha(\mu)$ and that for $t_0 \le t \le t_{f_\psi}$, $\hat{e}_{k,n}(t) \le \alpha(\mu)$ where

$$\begin{aligned}\alpha^2(\mu) &= 4M^2\mu^2(1 + 2K\mu)^{[(T-t_0)/\mu]} + O(\mu)\frac{(1 + 2K\mu)^{[(T-t_0)/\mu]} - 1}{2K} \\ &\le 4M^2\mu^2 e^{2K(T-t_0)} + O(\mu)\frac{e^{2K(T-t_0)} - 1}{2K} \\ &= \eta^2(\mu). \qquad \square\end{aligned}$$

LEMMA 17. *Let Assumption* I *and either Assumption* II *or* II′ *hold. Let u and v be any control functions in Y and Z, respectively. If t_f is the terminal time of a trajectory $\hat{\phi}(\ , \tau_0, \hat{\xi}_0, u, v, k)$ with $(\tau_0, \hat{\xi}_0) \in \mathcal{X}$, then for sufficiently large k,*

$$|\phi^0(T, \tau_0, \hat{\xi}_0, u, v, k) - \phi^0(t_f, \tau_0, \hat{\xi}_0, u, v, k)| \le \frac{M}{c_0 k}.$$

Proof. If $t_f = T$, then the inequality is clearly true. Assume $t_f < T$. Let $\hat{\theta}(t) = \hat{\rho}(t, \hat{\phi}(t))$, then $d\hat{\theta}/dt = \hat{\rho}_t(t, \hat{\phi}(t)) + \langle \hat{\rho}_{\hat{x}}(t, \hat{\phi}(t)), \hat{f}_k(t, \hat{\phi}(t), u(t), v(t))\rangle$. Let $k > 1/\epsilon_0$. If $(t, \hat{\phi}(t))$ stays in $\mathcal{N}_{1/k}(\mathcal{B})$, the right side of the preceding equation is $\le -c_0$. Let $t_f \le t \le T$ and let $(t, \hat{\phi}(t)) \in \mathcal{N}_{1/k}(\mathcal{B})$. We have that $\hat{\theta}(t) - \hat{\theta}(t_f) \le -c_0(t - t_f)$. Since $\hat{\theta}(t_f) = 0$, $(t - t_f) \le -\hat{\theta}(t)/c_0 \le 1/c_0 k$. Hence $|\phi^0(T, \tau_0, \hat{\xi}_0, u, v, k) - \phi^0(t_f, \tau_0, \hat{\xi}_0, u, v, k)| \le M/c_0 k$. $\square$

COROLLARY 18. *Let Assumption* I *and either Assumption* II *or* II′ *hold. If t_f is the terminal time of a motion $\hat{\phi}[\ , \tau_0, \hat{\xi}_0, \Gamma, \Delta, k]$ with $(\tau_0, \hat{\xi}_0) \in \mathcal{X}$, then for sufficiently large k, $|\phi^0[T, \tau_0, \hat{\xi}_0, \Gamma, \Delta, k] - \phi^0[t_f, \tau_0, \hat{\xi}_0, \Gamma, \Delta, k]| \le M/c_0 k$.*

LEMMA 19. *Let Assumption* I, *either Assumption* II *or* II′, *and the Isaacs condition (Î) hold for $\hat{f}$. Then*

$$W^+_{0,\mu}(t_0, x_0) \le W(t_0, x_0) + \left[\frac{2M}{c_0}\left(\frac{M+1}{c_0} + 1\right)\right]\eta(\mu).$$

Proof. From Lemma 16, there exists a $\bar{t} \in [t_0, t_f]$ such that

$$(\bar{t}, \hat{\phi}[\bar{t}, t_0, \hat{x}_0, \Gamma^0, \Delta^\mu_e, k]) \in C^{(k)}_{\eta(\mu)}(\nu_{0,k} \cap H(\bar{t})$$

and $0 \le t_f - \bar{t} \le (1/c_0)\eta(\mu)$ if μ is sufficiently small, where t_f is the terminal time of $\hat{\phi}[t, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]$.

Since $(\bar{t}, \hat{\phi}[\bar{t}, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]) \in C^{(k)}_{\eta(\mu)}(\nu_{0,k})$, there exists a $\hat{w} \in \mathbb{R}^{n+1}$ such that $(\bar{t}, \hat{w}) \in C^{(k)}(\nu_{0,k})$ and

$$|\hat{\phi}[\bar{t}, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] - \hat{w}| \le \eta(\mu). \tag{23}$$

Since $(\bar{t}, \hat{w}) \in C^{(k)}(\nu_{0,k})$, it follows from [1, Lem. 8.2] that for any $\epsilon > 0$ and any strategy Γ for Player I over $[\bar{t}, T]$ there exists a strategy $\Delta(\Gamma)$ for Player II over $[\bar{t}, T]$ such that

$$\phi^0[T, \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k] \le \nu_{0,k} + \epsilon. \tag{24}$$

Note that if $\hat{\phi}[t] = \hat{\phi}[t, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]$, then

$$\begin{aligned}
|(\bar{t}, \hat{w}) - (t_f, \hat{\phi}[t_f])| &\le |\bar{t} - t_f| + |\hat{w} - \hat{\phi}[t_f]| \\
&\le |\bar{t} - t_f| + |\hat{w} - \hat{\phi}[\bar{t}]| + |\hat{\phi}[\bar{t}] - \hat{\phi}[t_f]| \\
&\le \frac{1}{c_0}\eta(\mu) + \eta(\mu) + \frac{M}{c_0}\eta(\mu) \\
&= \Big(\frac{M+1}{c_0} + 1\Big)\eta(\mu).
\end{aligned}$$

Since $(t_f, \hat{\phi}[t_f, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k]) \in \mathcal{B}$, $(\bar{t}, \hat{w}) \in \mathcal{N}_{(((M+1)/c_0)+1)\eta(\mu)}(\mathcal{B})$.

If $(\bar{t}, \hat{w}) \notin \hat{T}$ and if $\bar{t}_f$ is the terminal time of $\hat{\phi}[\ , \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k]$, then Corollary 4 asserts that for sufficiently small μ

$$\bar{t}_f - \bar{t} \le \frac{1}{c_0}\left(\frac{M+1}{c_0} + 1\right)\eta(\mu). \tag{25}$$

By Corollary 18 we have that there exists a k_0, independent of Γ and Δ, such that if $k \ge k_0$ and if t_f is the terminal time of $\hat{\phi}[t, \tau_0, \hat{\xi}_0, \Gamma, \Delta, k]$, then

$$|\phi^0[T, \tau_0, \hat{\xi}_0, \Gamma, \Delta, k] - \phi^0[t_f, \tau_0, \hat{\xi}_0, \Gamma, \Delta.k]| \le \frac{M}{c_0 k}. \tag{26}$$

Therefore

$$\begin{aligned}
&|w^0 - \phi^0[T, \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k]| \\
\le &|w^0 - \phi^0[\bar{t}_f, \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k]| \\
&\quad + |\phi^0[\bar{t}_f, \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k] - \phi^0[T, \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k]| \\
\le &M(\bar{t}_f - \bar{t}) + \frac{M}{c_0 k} \qquad \text{(by (26))} \\
\le &\frac{M}{c_0}\left(\frac{M+1}{c_0} + 1\right)\eta(\mu) + \frac{M}{c_0 k} \qquad \text{(by (25))}.
\end{aligned} \tag{27}$$

If $(\bar{t}, \hat{w}) \in \hat{T}$, Corollary 18 implies that

$$|w^0 - \phi^0[T, \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k]| \le \frac{M}{c_0 k} \le \frac{M}{c_0}\left(\frac{M+1}{c_0} + 1\right)\eta(\mu) + \frac{M}{c_0 k}.$$

Hence, in any case, (27) holds.

Now, we have

$$\begin{aligned}
&\phi^0[T, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] \\
&= \phi^0[\bar{t}, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] + \phi^0[t_f, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] - \phi^0[\bar{t}, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] \\
&\quad + \phi^0[T, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] - \phi^0[\bar{t}, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] \\
&\le \phi^0[\bar{t}, t_0, \hat{x}_0, \Gamma^0, \Delta_e^\mu, k] + M(t_f - \bar{t}) + \frac{M}{c_0 k} \qquad \text{(by (26))} \\
&\le w^0 + \eta(\mu) + \frac{M}{c_0}\eta(\mu) + \frac{M}{c_0 k} \qquad \text{(by (23) and the definition of } \bar{t}) \\
&\le \phi^0[T, \bar{t}, \hat{w}, \Gamma, \Delta(\Gamma), k] + \frac{M}{c_0}\left(\frac{M+1}{c_0} + 1\right)\eta(\mu) + \frac{M}{c_0 k} \\
&\quad + \eta(\mu) + \frac{M}{c_0}\eta(\mu) + \frac{M}{c_0 k} \qquad \text{(by (27))} \\
&\le \nu_{0,k} + \epsilon + \left[\frac{M}{c_0}\left(\frac{M+1}{c_0} + 1\right)\right]\eta(\mu) + \frac{2M}{c_0 k} \qquad \text{(by (24))}.
\end{aligned}$$

By the arbitrariness of ϵ and the definition of $\hat{W}^+_{0,\mu}(t_0, \hat{x}_0, k)$, we get

$$\hat{W}^+_{0,\mu}(t_0, \hat{x}_0, k) \le \hat{W}^+_{0,0}(t_0, \hat{x}_0, k) + \left[\frac{2M}{c_0}\left(\frac{M+1}{c_0} + 1\right) + 1\right]\eta(\mu) + \frac{2M}{c_0 k}.$$

If k is sufficiently large, it follows from Lemma 15 that

$$W^+_{0,\mu}(t_0, x_0) \le W(t_0, x_0) + \frac{2c_0}{k} + \left[\frac{2M}{c_0}\left(\frac{M+1}{c_0} + 1\right) + 1\right]\eta(\mu) + \frac{2M}{c_0 k}.$$

The lemma is shown by letting $k \to \infty$. □

Under the same assumptions as in Lemma 19, we have $W^-_{\lambda,0}(t_0, x_0) \ge W(t_0, x_0) - ((M/c_0) + 1)\eta(\lambda)$. Hence, by Lemma 13, we have the following result.

THEOREM 20. *If Assumption* I *and either Assumption* II *or* II′ *hold and if the Isaacs condition (Î) holds, there exist a* $\delta > 0$ *and a nonnegative function* σ *such that* $\sigma(s) \to 0$ *as* $s \to 0$ *and*

$$0 \le W^+_{\lambda,\mu}(t_0, x_0) - W^-_{\lambda,\mu}(t_0, x_0) \le \sigma(\lambda) + \sigma(\mu)$$

for $0 \le \lambda, \mu \le \delta$, *and hence*

$$\lim_{\substack{\lambda \to 0 \\ \mu \to 0}} W^\pm_{\lambda,\mu}(t_0, x_0) = W(t_0, x_0).$$

2.3. Differential games of survival with lags. The game is governed by (1) with payoff (2). Applying the technique that has been used in [3], for differential games of survival with information lags, we have the following lemma.

LEMMA 21. *Assume that Assumption* I′ *and the Isaacs condition* ($\hat{I}$) *hold for* $\hat{f}$ *and that* F *satisfies either Assumption* II *or* II′. *If* g *is* $C^{(2)}$, *then there exist a* $\delta > 0$ *and a nonnegative function* σ *such that* $\sigma(s) \to 0$ *as* $s \to 0$ *and*

$$W^{-}_{\lambda,0}(t_0, x_0) + \sigma(\lambda) \geq W(t_0, x_0), \qquad \text{if } 0 \leq \lambda \leq \delta,$$
$$W^{+}_{0,\mu}(t_0, x_0) \leq W(t_0, x_0) + \sigma(\mu), \qquad \text{if } 0 \leq \mu \leq \delta.$$

Proof. We can assume without loss of generality that $g(t_0, x_0) = 0$. Then we can write

$$g(t_f, x_f) + \int_{t_0}^{t_f} f^0(s, \phi(s), u(s), v(s))\, ds$$
$$= \int_{t_0}^{t_f} [g_t(s, \phi(s)) + \langle Dg(s, \phi(s)), f(s, \phi(s), u(s), v(s))\rangle + f^0(s, \phi(s), u(s), v(s))]\, ds.$$

Thus the game G can be written as a game $\tilde{G}$ with $\tilde{g} \equiv 0$ and $\tilde{f}^0$ given by

$$\tilde{f}^0(t, x, y, z) = g_t(t, x) + \langle Dg(t, x), f(t, x, y, z)\rangle + f^0(t, x, y, z).$$

It is immediate to verify that if G satisfies the assumptions in Lemma 21, $\tilde{G}$ satisfies assumptions in Lemma 19. So, Lemma 21 holds. □

Remark. The condition on g in Lemma 21 can be weakened. If g is $C^{1,1}$, i.e., g is $C^{(1)}$ and g' is Lipschitz continuous, then the results of Lemma 21 hold.

Lemma 13 and Lemma 21 give us the following theorem.

THEOREM 22. *Under the same assumptions as in Lemma* 21, *the result of Theorem* 20 *holds.*

REFERENCES

[1] L. D. BERKOVITZ, *The existence of value and saddle point in games of fixed duration*, SIAM J. Control Optim., 23 (March 1985), pp. 172–196.

[2] ———, *Differential games of generalized pursuit and evasion*, SIAM J. Control Optim., 24 (1986 pages 361–373).

[3] ———, *Differential games of survival*, J. Math. Anal. and Appl., 29 (1988), pp. 493–504.

[4] ———, *Optimal feedback controls*, SIAM J. Control Optim., 27 (1989), pp. 991–1006.

[5] A. FRIEDMAN, *Differential Games*, John Wiley, New York, London, Sydney, Toronto, 1971.

[6] ———, *Differential Games*, CBMS Regional Conference Series in Applied Mathematics 18, American Mathematical Society, Providence, RI, 1974.

[7] R. ISAACS, *Differential Games*, John Wiley, New York, London, Sydney, 1965.

[8] N. N. KRASOVSKII AND A. I. SUBBOTIN, *Positional Differential Games*, Nauka, Moscow, 1974. (In Russian.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 831–854, May 1994

011

A DISSIPATIVE FEEDBACK CONTROL SYNTHESIS FOR SYSTEMS ARISING IN FLUID DYNAMICS*

KAZUFUMI ITO† AND SUNGKWON KANG‡

Abstract. A dissipative feedback control synthesis is constructed to regulate the systems arising in fluid dynamics. The feedback law is obtained by utilizing nonlinear dynamic programming techniques. The control law is designed for driving the system to a prescribed equilibrium state and enhancing the energy dissipation effects of the dynamical system. Two-dimensional Navier–Stokes equations and Burgers equation are used for numerical experiments to illustrate the effects of the feedback synthesis and the theoretical results.

Key words. feedback control, dynamic programming, Burgers' equation, Navier–Stokes equations

AMS subject classifications. 49J20, 76D05, 93B52

1. Introduction.

During the past years considerable attention has been given to the problem of active control of fluid flows. This interest is motivated by a number of potential applications such as control of flow separation, combustion, fluid-structure interaction, and super maneuverable aircraft. In this direction, Burns, Ito, and Kang ([BIK], [BK1], [BK2]) developed several computational algorithms for active control design for the Burgers equation, a simple model for convection-diffusion phenomena such as shock waves, traffic flows, supersonic flow around airfoils, etc. Using linearization techniques and linear quadratic regulator (LQR) theory, the feedback control laws were constructed to obtain a certain desired degree of stability for the closed-loop nonlinear system.

In this paper we construct and analyze a dissipative feedback control synthesis that regulates the systems arising in fluid dynamics. The control law is obtained by utilizing nonlinear dynamic programming techniques and designed for enhancing the energy dissipation effects of the dynamical systems and driving the systems to specific equilibrium states.

Let H, U, and V be separable Hilbert spaces. Assume that V is densely and continuously embedded into H and let V^* be the (strong) dual space of V. Consider a control problem governed by the following semilinear dynamics:

$$\frac{d}{dt}x(t) + \epsilon \mathcal{A}x(t) + \mathcal{F}(x(t)) = \mathcal{B}u(t) + f(t), \qquad x(0) = x_0, \tag{1.1}$$

where $\epsilon > 0$, $\mathcal{A}$ is a nonnegative selfadjoint operator defined on H with $\mathcal{D}(\mathcal{A}^{1/2}) = V$, $\mathcal{F}$ is a locally Lipschitz mapping from V into V^*, $\mathcal{B} \in \mathcal{L}(U, H)$ is a control input operator, and we assume $U = \mathbf{R}^m$. Thus, we have $\mathcal{B}u = \sum_{i=1}^m b_i u_i$ with $b_i \in H$. A class of problems described by (1.1) includes the Navier–Stokes equations, Burgers equation, and reaction-diffusion equations. Let x_e be an equilibrium state of the

* Received by the editors November 25, 1991; accepted for publication (in revised form) November 23, 1992. This research was supported by Air Force Office of Scientific Research grant AFOSR-90-0091 and Office of Naval Research grant N00014-91-J-1526.

† Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, California 90089-1113. Current address, Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

‡ Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, California 90089-1113. Current address, Department of Mathematics, Chosun University, Kwangju, 501-759, Korea.

system (1.1) with $u(t) \equiv 0$ and $f(t) \equiv 0$. In this paper, we consider a specific feedback control mechanism $u(t)$ of the form

$$
\begin{aligned}
u(t) &= (u_1(t), \dots, u_m(t)) \in \mathbf{R}^m, \\
u_i(t) &= -\gamma_i(t)\langle b_i, x(t) - x_e\rangle_H, \qquad i = 1, 2, \dots, m,
\end{aligned}
\tag{1.2}
$$

where b_i are control distribution functions in H and the feedback laws $\gamma_i(t) > 0$ will be determined by finding a suboptimal solution to the Hamilton–Jacobi–Bellman (HJB) equation. Note that the control law (1.2) is of the form $\mathcal{B}^* x(t)$ and this form of feedback is used commonly in control of flexible structure (e.g., in [B]) and is called the co-located rate sensors or the passive feedback form. This feedback law (1.2) is derived from the following considerations. First, the passive form (1.2) is an essential part of a linear optimal control law (see Lemma 2.2). Second, the closed-loop system (with $f \equiv 0$) is dissipative (see Theorem 2.6). Finally, it is easily implementable practically. In practice, the control distribution functions $b_i \in H$ are chosen as locally supported, and hence the feedback forms $\langle b_i, x(t) - x_e\rangle_H$ become local operations. As will be seen in §3, the control force generated through locally supported control distributions b_i's makes a significant change in global patterns of the flow.

The outline of the paper is as follows. In §2, detailed derivation of the feedback control law (1.2) is described. Also, properties of the feedback law, as well as well-posedness of the closed-loop system, are established. The feedback synthesis (1.2) is applied to the Burgers equation and the two-dimensional Navier–Stokes equations in §3. To see how this controller affects the global nature of the transient flow, such as achievement of the desired asymptotic behavior, energy dissipation effects, and time-dependent behavior of the solution, several numerical computations are performed.

Throughout this paper notation is very standard. We will use the notation $|\cdot|$ without any subindex for vector or operator norm. In all such cases the appropriate index for $|\cdot|$ will be understood from the context. A given Banach space X, X^* and $\langle\cdot,\cdot\rangle_{X^*,X}$ denote the strong dual space of X and the dual product, respectively. If X is a Hilbert space, $\langle\cdot,\cdot\rangle$ is the scalar inner product.

2. A dissipative feedback control law. In this section we consider a specific control law (1.2) and a "suboptimal" control law $\gamma_i(t)$ is derived. The control laws $\gamma_i(t)$ are determined based on the dynamic programming technique [FR]. Let H be a Hilbert space. Consider the following control problem

$$
\frac{d}{dt}x(t) + \epsilon\mathcal{A}x(t) + \mathcal{F}(x(t)) = \mathcal{B}u(t) + f(t), \qquad x(0) = x_0,
\tag{2.1}
$$

where $\epsilon > 0$, $\mathcal{A}$ is a nonnegative selfadjoint operator on H, and $\mathcal{B} \in \mathcal{L}(U, H)$ is the control input operator defined by

$$
\mathcal{B}u = \sum_{i=1}^{m} u_i b_i,
\tag{2.2}
$$

with $U = \mathbf{R}^m$ and $b_i \in H$. The function f is a source term and $\mathcal{F}$ is a locally Lipschitz nonlinear operator from $\mathcal{D}(\mathcal{A}^{1/2})$ into $\mathcal{D}(\mathcal{A}^{1/2})^*$.

We consider a regulation problem for the solution of (2.1) to an *equilibrium* $x_e \in \mathcal{D}(\mathcal{A}^{1/2})$ satisfying

$$
\epsilon\mathcal{A}x_e + \mathcal{F}(x_e) = 0 \quad \text{in } \mathcal{D}(\mathcal{A}^{1/2})^*.
$$

Assume the following:

(A1) $\langle \mathcal{F}(x) - \mathcal{F}(x_e),\, x - x_e \rangle = 0$ for all $x \in \mathcal{D}(\mathcal{A}^{1/2})$,

(A2) $\mathcal{F}(x)$ is Fréchet differentiable at x_e with derivative $\mathcal{F}'(x_e) \in \mathcal{L}(\mathcal{D}(\mathcal{A}^{1/2}), \mathcal{D}(\mathcal{A}^{1/2})^*)$.

Then the linearization of (2.1) about the equilibrium x_e becomes

$$\frac{d}{dt}y(t) + \epsilon \mathcal{A}y(t) + \mathcal{F}'(x_e)y(t) = \mathcal{B}v(t), \tag{2.3}$$

where $y(t)$ represents the variation of $x(t)$ from x_e.

Let $V = \mathcal{D}(\mathcal{A}^{1/2})$ with norm $\sqrt{|x|^2 + \langle \mathcal{A}x, x \rangle}$, and $\sigma_{x_e}(\cdot, \cdot)$ be a sesquilinear form on $V \times V$ defined by

$$\sigma_{x_e}(\phi, \psi) = \epsilon \langle \mathcal{A}\phi, \psi \rangle + \langle \mathcal{F}'(x_e)\phi, \psi \rangle.$$

Then it follows from (A1) and (A2) that σ_{x_e} is continuous and

$$\sigma_{x_e}(\phi, \phi) = \epsilon \langle \mathcal{A}\phi, \phi \rangle \quad \text{for all } \phi \in V.$$

In fact, $\mathcal{F}'(x_e)$ is skew-adjoint and

$$\langle \mathcal{F}'(x_e)\phi, \phi \rangle = 0 \quad \text{for all } \phi \in V.$$

Moreover, it follows from [P, p. 81] or [Ta] that the operator $\mathcal{A}_{x_e} = -\epsilon \mathcal{A} - \mathcal{F}'(x_e)$ generates an analytic semigroup on H.

2.1. Linear quadratic regulator problem. To give a mathematical motivation for considering the feedback law of form (1.2), we consider the following linear quadratic regulator (LQR) problem for the linearlized control systems (2.3): *Find the optimal control v that minimizes the quadratic cost functional*

$$J(v) = \int_0^\infty \left(|\mathcal{A}^{\frac{1}{2}} y(t)|^2 + |v(t)|^2 \right) dt, \tag{2.4}$$

subject to the control system (2.3).

Here, $|\mathcal{A}^{1/2}y|^2 = \langle y, \mathcal{A}y \rangle$ represents an energy stored in the system and the weight on control consumption is included in the definition of the operator $\mathcal{B}$. Assume that the pair $(\mathcal{A}_{x_e}, \mathcal{B})$ is *exponentially stabilizable*; i.e., there exists an operator $\mathcal{K} \in \mathcal{L}(V, U)$ such that $\mathcal{A}_{x_e} + \mathcal{B}\mathcal{K}$ generates an exponentially stable semigroup on H. It then follows from [PS], [BI] that the (LQR) problem has a unique optimal solution that is given by

$$v^{op}(t) = -\mathcal{B}^*\mathcal{P}y(t), \tag{2.5}$$

where the nonnegative selfadjoint operator $\mathcal{P} \in \mathcal{L}(H) \cap \mathcal{L}(V)$ satisfies the Riccati equation

$$\langle \mathcal{A}_{x_e}\phi, \mathcal{P}\psi \rangle + \langle \mathcal{P}\phi, \mathcal{A}_{x_e}\psi \rangle - \langle \mathcal{B}^*\mathcal{P}\phi,\, \mathcal{B}^*\mathcal{P}\psi \rangle + \langle \mathcal{A}\phi, \psi \rangle = 0 \tag{2.6}$$

for all $\phi, \psi \in V$. There are two difficulties associated with the feedback law (2.5); one is that the closed-loop operator $\mathcal{A}_{x_e} - \mathcal{B}\mathcal{B}^*\mathcal{P}$ may not generate an exponentially stable

semigroup, and the other is that $\mathcal{P}$ is not readily available and finding the solution $\mathcal{P}$ of (2.6) is nontrivial in general. Thus, we consider the following cost functional

$$J(v) = \int_0^\infty \left(|\mathcal{A}^{\frac{1}{2}} y(t)|^2 + \left| \frac{1}{2\epsilon} \mathcal{B}^* y(t) \right|^2 + |v(t)|^2 \right) dt. \tag{2.7}$$

Then we have the following result.

THEOREM 2.1. *Assume that $(\mathcal{A}_{x_e}, \mathcal{B})$ is exponentially stabilizable. Then there exists a unique optimal control that minimizes the cost* (2.7) *subject to the control system* (2.3) *and it is given by*

$$v^{op}(t) = -\frac{1}{2\epsilon} \mathcal{B}^* y(t). \tag{2.8}$$

Moreover, $\mathcal{A}_{x_e} - \frac{1}{2\epsilon}\mathcal{B}\mathcal{B}^*$ generates an exponentially stable (analytic) semigroup on H.

Proof. Note that $\bar{\mathcal{P}} = \frac{1}{2\epsilon} I \in \mathcal{L}(H) \cap \mathcal{L}(V)$ satisfies the Riccati equation

$$\begin{aligned} &\langle \mathcal{A}_{x_e}\phi, \bar{\mathcal{P}}\psi \rangle + \langle \bar{\mathcal{P}}\phi, \mathcal{A}_{x_e}\psi \rangle - \langle \mathcal{B}^*\bar{\mathcal{P}}\phi, *\mathcal{B}^*\bar{\mathcal{P}}\psi \rangle \\ &\qquad + \langle \mathcal{A}\phi, \psi \rangle + \tfrac{1}{4\epsilon^2} \langle \mathcal{B}^*\phi, \mathcal{B}^*\psi \rangle = 0 \end{aligned} \tag{2.9}$$

for all $\phi, \psi \in V$. Hence, the theorem follows from [BI, Thm. 3.5] and [PS, Thms. 3.4, 3.5 and Rem. 3.5], since $(\mathcal{A}_{x_e}, \mathcal{B}^*)$ is detectable. Therefore, the Riccati equation (2.9) has a unique nonnegative solution and the optimal control is given by (2.8). □

It follows from (2.2) that

$$(\mathcal{B}^*\phi)_i = \langle b_i, \phi \rangle \quad \text{for all } \phi \in H.$$

Thus the feedback law (2.8) is of form (2.1). Next, we show that the feedback law (2.8) is an essential part of linear feedback laws based on the Riccati equation in the following sense.

LEMMA 2.2. *Given a nonnegative selfadjoint operator $\mathcal{Q}$, assume that a nonnegative selfadjoint operator $\mathcal{P} \in \mathcal{L}(H) \cap \mathcal{L}(V)$ satisfies the Riccati equation*

$$\langle \mathcal{A}_{x_e}\phi, \mathcal{P}\psi \rangle + \langle \mathcal{P}\phi, \mathcal{A}_{x_e}\psi \rangle - \langle \mathcal{B}^*\mathcal{P}\phi, \mathcal{B}^*\mathcal{P}\psi \rangle + \langle \mathcal{A}\phi, \psi \rangle + \langle \mathcal{Q}\phi, \psi \rangle = 0 \tag{2.10}$$

for all $\phi, \psi \in V$. Then if V is compactly embedded in H, $\Sigma = \mathcal{P} - \frac{1}{2\epsilon} I$ is compact on H.

Proof. From (2.9) and (2.10) we have

$$\langle \mathcal{A}_{x_e}\phi, \Sigma\psi \rangle + \langle \Sigma\phi, \mathcal{A}_{x_e}\psi \rangle - \langle \mathcal{B}^*\mathcal{P}\phi, \mathcal{B}^*\mathcal{P}\psi \rangle + \langle \mathcal{Q}\phi, \psi \rangle = 0$$

for all $\phi, \psi \in V$. Thus we obtain

$$\Sigma = \int_0^t e^{s\mathcal{A}_{x_e}^*} (\mathcal{Q} - \mathcal{P}\mathcal{B}\mathcal{B}^*\mathcal{P}) e^{s\mathcal{A}_{x_e}} ds + e^{t\mathcal{A}_{x_e}^*} \Sigma e^{t\mathcal{A}_{x_e}},$$

where $e^{t\mathcal{A}_{x_e}}$, $t \geq 0$, denotes the analytic semigroup on H generated by $\mathcal{A}_{x_e}$. The lemma follows from the fact that $\Sigma x \in V$ for $x \in H$, since $|e^{t\mathcal{A}_{x_e}}|_{\mathcal{L}(H,V)} \leq M/t^{1/2}$, $t > 0$, for some constant $M > 0$ (see [Ta]). □

2.2. A nonlinear feedback law. Now, we seek a feedback control law of the form

$$u_i(t) = -\gamma_i(t)\langle b_i,\, x(t) - x_e\rangle, \tag{2.11}$$

where $\gamma_i(t) \geq 0$ is chosen so that the transition of the solution of the closed-loop system is accounted; i.e., the cost functional

$$\frac{1}{2}\int_0^\infty \left(\langle x(t) - x_e, \mathcal{A}(x(t) - x_e)\rangle + \langle \mathcal{Q}(x(t) - x_e), x(t) - x_e\rangle + |u(t)|^2\right) dt \tag{2.12}$$

is minimized subject to the control system

$$\begin{aligned} &\frac{d}{dt}x(t) + \epsilon\,\mathcal{A}x(t) + \mathcal{F}(x(t)) = \mathcal{B}u(t), \\ &u_i(t) = -\gamma_i(t)\langle b_i,\, x(t) - x_e\rangle, \qquad (1 \leq i \leq m). \end{aligned} \tag{2.13}$$

Before determining the feedback control law $\gamma(t) = (\gamma_1(t), \ldots, \gamma_m(t))$ we state the dissipative property of the closed-loop system (2.13).

LEMMA 2.3. *Given $\gamma(t) \in L^\infty(0,T;\mathbf{R}_+^m)$ and $x(0) = \phi \in H$, assume that* (2.13) *has a solution $x(t) \in L^2(0,T;V) \cap C(0,T;H) \cap H^1(0,T;V^*)$. Then we have*

$$\begin{aligned} &\frac{1}{2}|x(t) - x_e|_H^2 + \int_0^t \Big(\epsilon\langle x(s) - x_e, \mathcal{A}(x(s) - x_e)\rangle \\ &\qquad + \sum_{i=1}^m \gamma_i(s)|\langle b_i,\, x(t) - x_e\rangle|^2\Big)\, ds \\ &\qquad = \frac{1}{2}|x(0) - x_e|_H^2, \qquad t \in [0,T], \end{aligned} \tag{2.14}$$

where $\mathbf{R}_+^m = \{\,\eta = (\eta_1, \ldots, \eta_m) \in \mathbf{R}^m \,:\, \eta_i \geq 0,\ 1 \leq i \leq m\,\}$.

Proof. It follows from (2.13) that

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}|x(t) - x_e|_H^2 + \epsilon\,\langle \mathcal{A}(x(t) - x_e), x(t) - x_e\rangle + \langle \mathcal{F}(x(t)) - \mathcal{F}(x_e), x(t) - x_e\rangle \\ &\qquad + \sum_{i=1}^m \gamma_i(t)|\langle b_i,\, x(t) - x_e\rangle|^2 = 0. \end{aligned}$$

Integration of the above equation and assumption (A1) yield the estimate (2.14). □

Remark. In Lemma 2.8, we give a condition on $\mathcal{F}$ in which (2.13) has a unique global solution. Lemma 2.3 implies that the optimal feedback law provides additional energy dissipation.

We now construct the feedback laws $\gamma_i(t)$ in (2.11). Determination of the "optimal" control $\gamma_i(t)$ involves solving the HJB equation (see [FR]). It is not our interest here to establish a complete theory of feedback solutions based on the HJB equation. Rather we will use the HJB equation as a means of construction of feedback synthesis $\gamma_i(t)$. Thus, the following discussions are by no means rigorous. It is of our interest in future to verify the details. Also, we refer to a recent paper [Sr] and the references

therein for a mathematical treatment of the HJB equation in which the finite time horizon problem is treated.

Consider the HJB equation

(2.15)
$$\mathrm{Min}_{\gamma\geq 0}\left[\langle p, -\epsilon\mathcal{A}(x-x_e)-(\mathcal{F}(x)-\mathcal{F}(x_e))-\sum_{i=1}^{m}\gamma_i b_i\langle b_i, x-x_e\rangle\rangle\right.$$
$$\left.+\frac{1}{2}\left(\sum_{i=1}^{m}|\langle b_i, x-x_e\rangle|^2|\gamma_i|^2+\langle\mathcal{A}(x-x_e), x-x_e\rangle+\langle\mathcal{Q}(x-x_e), x-x_e\rangle\right)\right]=0,$$

where $p=\frac{\partial}{\partial x}W(x)$ and $W(x)$ is the value function defined below. The minimum of the problem (2.15) is attained when

(2.16)
$$\gamma_i=\begin{cases}\dfrac{(\langle b_i,p\rangle)(\langle b_i, x-x_e\rangle)}{|\langle b_i, x-x_e\rangle|^2}, & \text{if } (\langle b_i,p\rangle)(\langle b_i, x-x_e\rangle)\geq 0,\\ 0, & \text{otherwise.}\end{cases}$$

By Hamilton–Jacobi theory (e.g., see [FR], [L] for the finite-dimensional case and [CL], [BD] for the infinite-dimensional case), the value function

$$W(x)=\mathrm{Min}_{\gamma}J(\gamma)$$

satisfies the HJB equation (2.15), where $J(\gamma)$ is defined by (2.12). Conversely, assuming that (2.15), (2.16) has a solution $W(x)$ that is (at least) Lipschitz in V, then (2.16) with $p=\frac{\partial}{\partial x}W(x)$, $x\in V$, provides an optimal solution. To obtain the optimal feedback control γ^{op} we must solve the nonlinear partial differential equations (2.15), (2.16) for $W(x)$ in V. Instead of solving the equations, we look for p having the following form:

(2.17)
$$p(x)=c(x)(x-x_e),$$

where $c(x)$ is a scalar function in $x\in V$. Substitution of $p(x)$ into (2.15), (2.16) yields the equation

$$|\mathcal{B}^*(x-x_e)|^2c(x)^2+2\epsilon\langle x-x_e, \mathcal{A}(x-x_e)\rangle c(x)$$
$$-(\langle\mathcal{Q}(x-x_e), x-x_e\rangle+\langle x-x_e, \mathcal{A}(x-x_e)\rangle)=0.$$

The positive solution of the above equation is given by

(2.18)
$$c(x)=\frac{-\epsilon\langle x-x_e, \mathcal{A}(x-x_e)\rangle+\sqrt{g_1(x)+g_2(x)|\mathcal{B}^*(x-x_e)|^2}}{|\mathcal{B}^*(x-x_e)|^2},$$

where

$$g_1(x)=(\epsilon\langle x-x_e, \mathcal{A}(x-x_e)\rangle)^2,$$
$$g_2(x)=\langle x-x_e, \mathcal{A}(x-x_e)\rangle+\langle\mathcal{Q}(x-x_e), x-x_e\rangle.$$

Hence, letting $\gamma_i(t)\equiv c(x(t))$ for each i, we obtain the nonlinear feedback law

(2.19)
$$u(t)=-c(x)\mathcal{B}^*(x-x_e).$$

2.3. Properties of the proposed feedback law. Now we investigate properties of the feedback law (2.19).

THEOREM 2.4. *Suppose that* $\mathcal{Q} = \frac{1}{4\epsilon^2}\mathcal{B}\mathcal{B}^*$. *Then* $W(x) = \frac{1}{4\epsilon}|x - x_e|_H^2$ *satisfies the* HJB *equation* (2.15), (2.16).

Proof. Note that $\frac{\partial}{\partial x}W(x) = \frac{1}{2\epsilon}(x - x_e)$ and $\gamma_i(t) = \frac{1}{2\epsilon}$. Thus the left-hand side of (2.15) is equal to

$$\frac{1}{2\epsilon}\left\langle x - x_e, -\epsilon\mathcal{A}(x - x_e) - (\mathcal{F}(x) - \mathcal{F}(x_e)) - \sum_{i=1}^{m}\frac{1}{2\epsilon}b_i\langle b_i, x - x_e\rangle\right\rangle$$

$$+\frac{1}{2}\left(\sum_{i=1}^{m}\frac{1}{4\epsilon^2}|\langle b_i, x - x_e\rangle|^2 + \langle\mathcal{A}(x - x_e), x - x_e\rangle\right) = 0. \qquad \square$$

The next theorem is concerned with the bound of $c(x)$.

THEOREM 2.5. *Suppose that* $\mathcal{Q} = \alpha\mathcal{B}\mathcal{B}^*$ *with* $0 < \alpha \le 1/4\epsilon^2$. *Then* $0 \le c(x) \le \frac{1}{2\epsilon}$, *where* $c(x)$ *is given by* (2.18).

Proof. Note that

$$|\epsilon\langle x - x_e, \mathcal{A}(x - x_e)\rangle|^2 + |\mathcal{B}^*(x - x_e)|^2(\langle x - x_e, \mathcal{A}(x - x_e)\rangle + \alpha|\mathcal{B}^*(x - x_e)|^2)$$

$$= (\epsilon\langle x - x_e, \mathcal{A}(x - x_e)\rangle + \tfrac{1}{2\epsilon}|\mathcal{B}^*(x - x_e)|^2)^2 + (\alpha - \tfrac{1}{4\epsilon^2})|\mathcal{B}^*(x - x_e)|^4.$$

Thus, the theorem follows from (2.18). $\square$

In what follows we assume that $\mathcal{Q} = \alpha\mathcal{B}\mathcal{B}^*$ with $1 \le \alpha \le 1/4\epsilon^2$.

THEOREM 2.6. *Assume that the closed-loop system*

$$\frac{d}{dt}x(t) + \epsilon\mathcal{A}x(t) + \mathcal{F}(x(t)) = -c(x)\mathcal{B}\mathcal{B}^*(x(t) - x_e), \qquad x(0) = \phi \in H, \tag{2.20}$$

has a solution $x(t) \in L^2(0,T;V)\cap C(0,T;H)\cap H^1(0,T;V^*)$. *Then we have for* $\omega \ge 0$

$$\frac{1}{2}e^{\omega t}|x(t) - x_e|_H^2 + \int_0^t(\sqrt{h_1(x) + |\mathcal{B}^*(x - x_e)|^2h_2(x)} - \frac{\omega}{2}|x(s) - x_e|_H^2)e^{\omega s}\,ds \tag{2.21}$$

$$= \frac{1}{2}|x(0) - x_e|_H^2,$$

where

$$h_1(x) = (\epsilon\langle x - x_e, \mathcal{A}(x - x_e)\rangle)^2$$
$$h_2(x) = \langle x - x_e, \mathcal{A}(x - x_e)\rangle + \alpha|\mathcal{B}^*(x - x_e)|^2. \tag{2.22}$$

Proof. The proof follows from Lemma 2.3. $\square$

Next, we obtain the asymptotic stability property of the closed-loop system (2.20).

THEOREM 2.7. *Given* $x(0) = \phi \in H$, *assume that the system* (2.20) *has a global solution. Assume that there exists a positive constant* β *such that*

$$\epsilon\langle x - x_e, \mathcal{A}(x - x_e)\rangle + |\mathcal{B}^*(x - x_e)|^2 \ge \beta|x - x_e|_H^2 \quad \textit{for all } x \in V.$$

Then $|x(t) - x_e|_H \le e^{-\beta t}|x(0) - x_e|_H$ *and* $\lim_{t\uparrow\infty}c(x) = \frac{1}{2\epsilon}$.

Proof. It follows from the estimate (2.21) and the inequality

$$\sqrt{h_1(x) + |\mathcal{B}^*(x - x_e)|^2 h_2(x)} \geq \epsilon \langle x - x_e, \mathcal{A}(x - x_e)\rangle + |\mathcal{B}^*(x(t) - x_e)|^2,$$

that if we set $\omega = 2\beta$ we have

$$e^{2\beta t} |x(t) - x_e|_H^2 \leq |x(0) - x_e|_H^2.$$

The last assertion follows from the formula

$$c(x) = \frac{\langle x - x_e, \mathcal{A}(x - x_e)\rangle + \alpha |\mathcal{B}^*(x - x_e)|^2}{\sqrt{h_1(x)} + \sqrt{h_1(x) + |\mathcal{B}^*(x - x_e)|^2 h_2(x)}},$$

where $h_1(x)$ and $h_2(x)$ are defined in (2.22). □

Throughout the rest of this section, we consider the well-posedness property of the closed-loop control system (2.13) mentioned in the remark. Assume that $\mathcal{F}(x)$ appearing in (2.1) is given by

$$\langle \mathcal{F}(x), \psi\rangle = b(x; x, \psi) \quad \text{for } \psi \in V, \tag{2.23}$$

where b is a bounded trilinear form on $V \times V \times V$ satisfying

$$b(u; v, w) + b(u; w, v) = 0 \quad \text{for all } u, v, w \in V, \tag{2.24}$$

$$|b(u; w, u)| \leq C |u|_H |u|_V |w|_V \quad \text{for all } u, w \in V \tag{2.25}$$

for some constant $C \geq 0$. Suppose that $x_e \in V$ satisfies $\mathcal{A}x_e + \mathcal{F}(x_e) = 0$. Then, from (2.24), we have $\langle \mathcal{A}x_e, x_e\rangle = 0$ and hence $x_e \in \ker(\mathcal{A})$ and $\mathcal{F}(x_e) = 0$. Therefore, assumption (A1) is satisfied if and only if

$$b(\phi; x_e, \phi) = 0 \quad \text{for all } \phi \in V. \tag{2.26}$$

In fact, for any $\phi \in V$,

$$\begin{aligned} \langle \mathcal{F}(x) - \mathcal{F}(x_e),\, \phi\rangle &= b(x - x_e + x_e;\, x - x_e + x_e,\, \phi) \\ &= b(x - x_e;\, x - x_e,\, \phi) + b(x_e;\, x - x_e,\, \phi) + b(x - x_e;\, x_e,\, \phi). \end{aligned} \tag{2.27}$$

LEMMA 2.8. *Under assumptions* (2.23)–(2.26) *on the nonlinear operator* $\mathcal{F}$, *the control system* (2.13) *has a unique global weak solution* $x(\cdot) \in L^2(0, \infty; V) \cap C_{\text{loc}}(0, \infty; H) \cap H^1(0, \infty; V^*)$ *provided that* $\gamma \in L^\infty(0, \infty; \mathbf{R}_+^m)$, *i.e.,* $x(t)$ *satisfies*

$$\left\langle \frac{d}{dt} x(t), \phi \right\rangle + \epsilon \langle \mathcal{A}x(t), \phi\rangle + b(x(t); x(t), \phi) + \left\langle \sum_{i=1}^m \gamma_i(t) b_i \langle b_i,\, x(t) - x_e\rangle,\, \phi \right\rangle = 0 \tag{2.28}$$

for all $\phi \in V$.

Proof. Let $\xi(t) = x(t) - x_e$. Then (2.28) is equivalently written as

$$\left\langle \frac{d}{dt} \xi(t), \phi \right\rangle + \epsilon \langle \mathcal{A}\xi(t), \phi\rangle + \langle \mathcal{F}_e(\xi), \phi\rangle + \left\langle \sum_{i=1}^m \gamma_i b_i \langle b_i, \xi(t)\rangle,\, \phi \right\rangle = 0 \tag{2.29}$$

for all $\phi \in V$, where $\langle \mathcal{F}_e(\xi), \phi \rangle = b(\xi; \xi, \phi) + b(x_e; \xi, \phi)$ from (2.27). Note that the bilinear form on $V \times V$

$$a(t; \phi, \psi) = \epsilon \langle \mathcal{A}\phi, \psi \rangle + \sum_{i=1}^{m} \gamma_i(t)(\langle b_i, \phi \rangle)(\langle b_i, \psi \rangle) \tag{2.30}$$

satisfies, for almost all t,

$$|a(t; \phi, \psi)| \leq M_1 |\phi|_V |\psi|_V \quad \text{for some } M_1 > 0, \tag{2.31}$$

$$a(t; \phi, \phi) \geq \epsilon |\phi|_V^2 - \epsilon |\phi|_H^2 \tag{2.32}$$

and that from inequality (2.25)

$$|\mathcal{F}_e(\xi)|_{V^*} \leq (C |\xi|_H + M_2 |x_e|_V) |\xi|_V$$

for some constant $M_2 > 0$. Hence, it is not difficult to show that (2.29) has a unique solution $\xi \in L^2(0, T; V) \cap C(0, T; H) \cap H^1(0, T; V^*)$ for arbitrary $T \geq 0$ (e.g., see [CF, Lem. 8.4] or [Te, p. 282]). The uniqueness follows from the inequality (2.25). □

3. Application to Burgers equation and Navier–Stokes equations. In this section we apply the dissipative feedback control law (2.18) to the Burgers equation and the two-dimensional Navier–Stokes equations to demonstrate the theoretical results in §2. Several numerical computations are performed to see how this controller affects the global nature of the transient flow, which is not easily characterized analytically. These calculations support the validation of the proposed feedback law not only for achievement of the desired asymptotic behavior but also for time-dependent behavior of the solution. Specifically, throughout our numerical calculations, effects of the feedback control law (2.18) are shown by observing energy dissipation described in Theorem 2.6, convergence of the feedback control gains as in Theorem 2.7, and, most importantly, changes in flow patterns. The computations were carried out on an IBM 3090 and a SUN SPARC Station 2 at the University of Southern California and the Center for Applied Mathematical Sciences. Detailed explanation of the numerical schemes and their convergence results will be reported in a forthcoming paper.

In the first two examples, the Burgers equation with Dirichlet boundary conditions (Example 1) and periodic boundary conditions (Example 2) are considered. Before discussing the details, we derive abstract variational forms for both cases. The governing equation is given by

$$\begin{aligned} &\frac{\partial}{\partial t} y(t, x) - \epsilon \frac{\partial^2}{\partial x^2} y(t, x) + y(t, x) \frac{\partial}{\partial x} y(t, x) \\ &\qquad = -\sum_{i=1}^{m} \gamma_i(t) \cdot b_i(x) \int_\Omega b_i(s)(y(t, s) - y_e(s)) ds, \\ &y(0, x) = y_0(x), \end{aligned} \tag{3.1}$$

with appropriate boundary conditions, where $\epsilon = \frac{1}{\mathrm{Re}}$; Re is the Reynolds number; $b_i(x)$ are control distribution functions; $\gamma_i(t)$ are control laws to be determined; and $y_e(x)$ is a desired equilibrium state.

For simplicity, let the domain for (3.1) be (0,1) for Example 1 and the period in space for Example 2 be 1. Let $\Omega = (0,1)$ and let $H = L^2(\Omega)$, $U = \mathbf{R}^m$, and $V = H_0^1(\Omega)$ (or $H_p^1(\Omega)$=the completion of the set of all C^∞-periodic functions with period 1). To place the control system (3.1) into a variational form, let $y(t)(\cdot) = y(t,\cdot)$, $y_0(\cdot) = y(0,\cdot)$. Define the operator $\mathcal{A}$ and the control input operator $\mathcal{B}$ by

$$\mathcal{A}\phi = -\phi'' \quad \text{and} \quad \mathcal{B}u = \sum_{i=1}^{m} b_i u_i \tag{3.2}$$

for all $\phi \in \mathcal{D}(\mathcal{A}) \subset H$. Then $\mathcal{A}$ is a nonnegative selfadjoint operator on H and $\mathcal{B} \in \mathcal{L}(U,H)$. We now define the bilinear form a on $V \times V$ by

$$a(t;\phi,\psi) = \epsilon \langle \mathcal{A}\phi, \psi\rangle + \sum_{i=1}^{m} \gamma_i(t) \left(\int_\Omega b_i(s)(\phi(s) - y_e)\, ds\right) \cdot \left(\int_\Omega b_i(s)\psi(s)\, ds\right), \tag{3.3}$$

the trilinear form b on $V \times V \times V$ by

$$b(\phi;\psi,\eta) = \frac{1}{3}\int_\Omega \left((\phi\psi)' + \phi\psi'\right)\eta\, dx, \tag{3.4}$$

and the nonlinear operator $\mathcal{F}$ by

$$\langle \mathcal{F}\phi,\ \eta\rangle = b(\phi;\phi,\eta) \tag{3.5}$$

for all ϕ, ψ, $\eta \in V$. Then the variational form, in V^*, for the system (3.1) is given by

$$\begin{aligned} &\left\langle \frac{d}{dt}y(t), \psi \right\rangle + a(t;y(t),\psi) + b(y(t);y(t),\psi) = 0, \quad \text{or} \\ &\left\langle \frac{d}{dt}y(t), \psi \right\rangle + \epsilon\langle \mathcal{A}y(t),\psi\rangle + \langle \mathcal{F}y(t),\psi\rangle = -\langle \gamma(t)\mathcal{B}\mathcal{B}^*(y(t) - y_e),\ \psi\rangle \end{aligned} \tag{3.6}$$

for all $\psi \in V$. It is easy to see that the bilinear form a satisfies inequalities (2.31), (2.32) and the trilinear form b satisfies conditions (2.24), (2.25). Since $0 < \gamma(t) \le \frac{1}{2\epsilon}$ (Theorem 2.5), by Lemma 2.8, the variational form (3.6) has a unique global weak solution $y(t) \in L^2(0,\infty;V) \cap C_{loc}(0,\infty;H) \cap H^1(0,\infty;V^*)$.

Example 1 (Burgers equation with Dirichlet boundary conditions). In this example, two stabilization problems for the Burgers equation are considered. Smoothing effects will be enhanced by two different control laws $\gamma(t) = \frac{1}{2\epsilon}$ (case 1) and $\gamma(t)$ obtained by (2.18) (case 2). The desired equilibrium state $y_e = 0$, the Reynolds number Re=200, the initial condition $y_0(x) = \sin 2\pi x$, and the Dirichlet boundary conditions $y(t,0) = y(t,1) = 0$ are chosen. Two control distribution functions $b_1(x)$ and $b_2(x)$ are chosen as

$$b_1(x) = \begin{cases} 1, & |x - 0.3| \le 0.1, \\ 0, & \text{otherwise}, \end{cases} \qquad b_2(x) = \begin{cases} 1, & |x - 0.7| \le 0.1, \\ 0, & \text{otherwise}. \end{cases} \tag{3.7}$$

For case 2, the control laws $\gamma_i(t) \equiv c(y)$ are given by the formula (2.18) with $x = y$, $x_e = 0$, and $\mathcal{Q} = \mathcal{B}\mathcal{B}^*$, i.e.,

$$c(y) = \frac{-\epsilon\,|y'|^2 + \sqrt{\epsilon^2\,|y'|^4 + g(y)(|y'|^2 + g(y))}}{g(y)}, \tag{3.8}$$

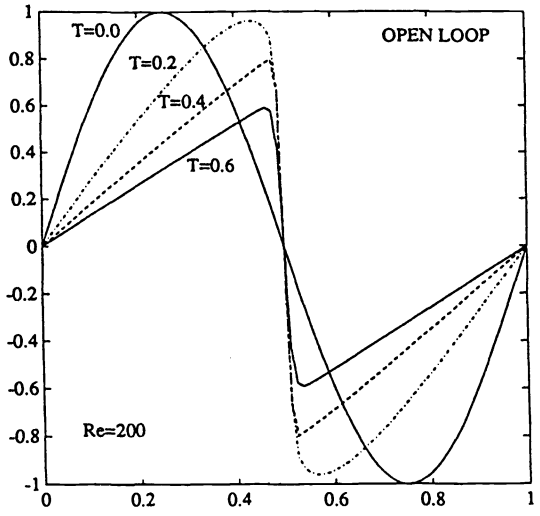


FIG. 1.1. *Open loop (Burgers equation). (Dirichlet B.C. Re* = 200.)

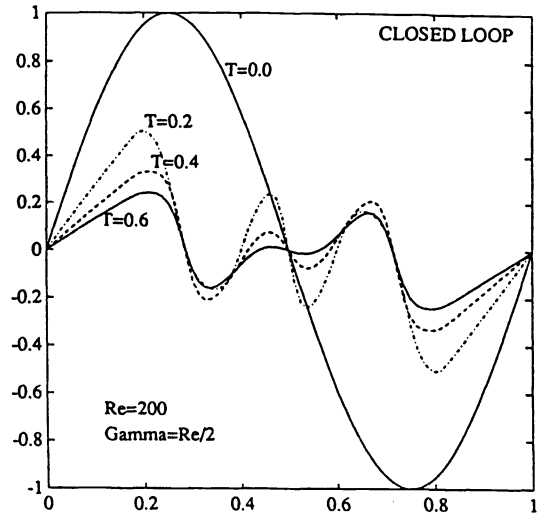


FIG. 1.2. *Closed loop* ($\gamma(t) \equiv 100$).

where $\epsilon = 0.005$ and

$$g(y) = \left(\int_0^1 b_1(x)y(t,x)\,dx\right)^2 + \left(\int_0^1 b_2(x)y(t,x)\,dx\right)^2.$$

In Fig. 1.1, it is easy to see that a steep gradient ("weak shock") of the open-loop solution ($\gamma_i(t) \equiv 0$) is forming in finite time around $x = 0.5$ due to the convective

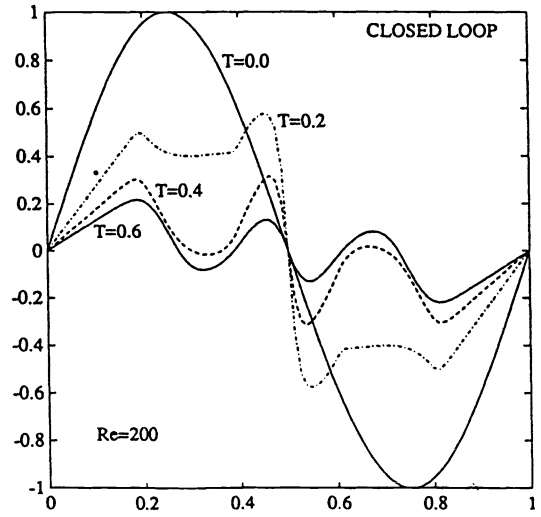


FIG. 1.3. *Closed loop* ($\gamma(t) \equiv 100$).

TABLE 1
Energy comparison (*Burgers equation*).

time	Open	Closed 1 ($\gamma(t) \equiv 100$)	Closed 2 ($\gamma(t) = c(y)$)	
	$\langle y, \mathcal{A}y \rangle$	$\langle y, \mathcal{A}y \rangle$	$\langle y, \mathcal{A}y \rangle$	$\gamma(t)$
$t = 0.0$	19.739	19.739	19.739	0.000
$t = 0.1$	26.346	25.075	18.139	19.736
$t = 0.2$	84.341	19.393	28.154	36.094
$t = 0.3$	108.429	14.639	23.589	59.208
$t = 0.4$	75.098	10.714	12.862	81.897
$t = 0.5$	49.646	7.421	6.901	93.517
$t = 0.6$	33.457	5.081	4.107	96.537

term $y(t,x)\frac{\partial}{\partial x}y(t,x)$. However, the solution is smoothed out eventually due to the diffusion effect $\epsilon(\partial^2/\partial x^2)y(t,x)$. Figures 1.2 and 1.3 show the closed-loop trajectories corresponding to two different feedback laws: $\gamma_i(t) \equiv \frac{1}{2\epsilon}$ (Fig. 1.2) and $\gamma_i(t) \equiv c(y)$ (Fig. 1.3). Recall that, from Theorem 2.7, $\lim_{t\uparrow\infty} \gamma_i(t) = \frac{1}{2\epsilon}$ (see Table 1). Comparing the trajectories (shown in Figs. 1.2 and 1.3) corresponding to the two controllers we observe that smoothing effects of the latter controller are better. For example, local weak shocks created by the latter controller in a neighborhood of control areas are less intensive than the other. Table 1 shows the energy dissipation effects by controllers and the convergence property of $\gamma_i(t)$ to $\frac{1}{2\epsilon}$. It is observed that the energy $\langle y, \mathcal{A}y \rangle$ of the closed-loop solution decay rapidly compared with those of the open-loop solution for both cases. Moreover, one can observe that the energy of the closed-loop in case 2 decays more gradually than the closed-loop in case 1. For numerical computation, we used the standard Galerkin approximation in [SB] with linear spline basis functions.

Example 2 ("Moving shock": Burgers equation with periodic boundary conditions). In this example, two tracking problems for the Burgers equation are considered. The initial condition $y_0(x) = \frac{1}{8} - \frac{1}{4}\sin 2\pi x$, the Reynolds number Re=1000, and

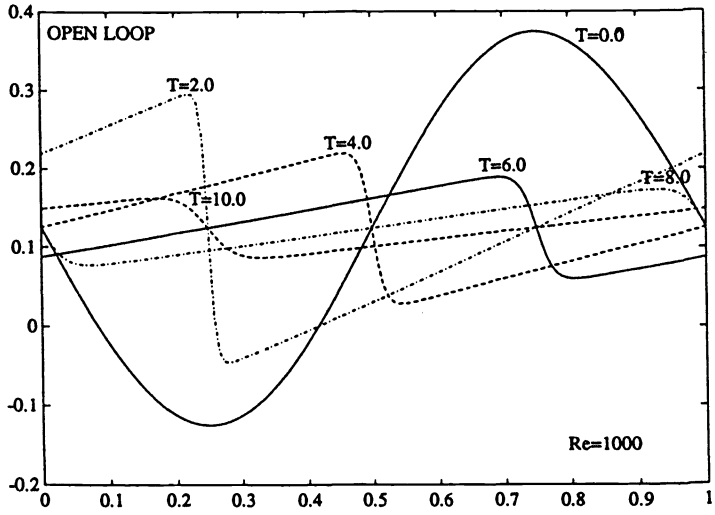


FIG. 2.1. *Open loop (Burgers equation). (Periodic B.C., Re = 1000).*

periodic boundary conditions are chosen. As time increases, a weak shock is formed and propagates with speed $\frac{1}{8}$. For this example, we will derive the solution to two desired equilibrium states $y_e = 0$ and $y_e = 0.5$ by injecting control signals. It is easy to see that the open-loop system (3.1) with $\gamma(t) \equiv 0$ has $\frac{1}{8}$ as the equilibrium state. A locally supported control distribution function $b(x)$ is chosen as

$$b(x) = \begin{cases} 1, & 0.8125 \le x \le 0.8672, \\ 0, & \text{otherwise}, \end{cases} \tag{3.9}$$

and the control law $\gamma(t)$ to achieve the desired state is obtained by formula (2.18) with $x = y$ and $Q = I$, i.e.,

$$\gamma(t) = \frac{-\epsilon\,|y'|^2 + \sqrt{\epsilon^2|y'|^4 + h(y, y_e)(|y'|^2 + |y - y_e|^2)}}{h(y, y_e)}, \tag{3.10}$$

where $\epsilon = 0.001$, $y_e = 0$ (Fig. 2.2) or $y_e = 0.5$ (Fig. 2.3), and

$$h(y, y_e) = \left(\int_0^1 b(x)(y(t, x) - y_e)\, dx \right)^2.$$

Figure 2.1 shows the formation of a weak shock (about time $t = 1.1$, which is not shown in the figure) and the propagation of the formed weak shock with speed $\frac{1}{8}$. Again, due to viscosity effects, the weak shock is smoothed out. The open-loop solution approaches the equilibrium state $y_e = \frac{1}{8} = 0.125$. The only control distribution function $b(x)$ is chosen which is locally supported with width 0.0547. In general, it is very difficult to regulate the flow because of high Reynolds number, 1000, and the very narrow control distribution region (compared with Example 1). As we see in Figs. 2.2 and 2.3, the desired equilibrium states $y_e = 0$ (Fig. 2.2) and $y_e = 0.5$

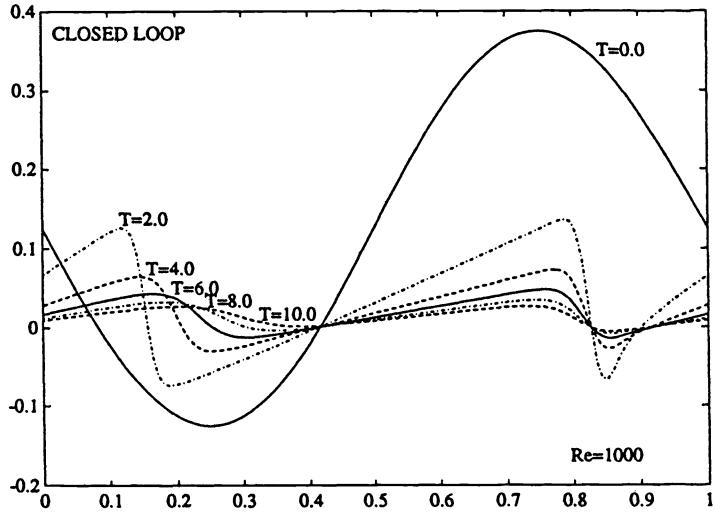


FIG. 2.2. *Closed loop* ($y_e = 0$).

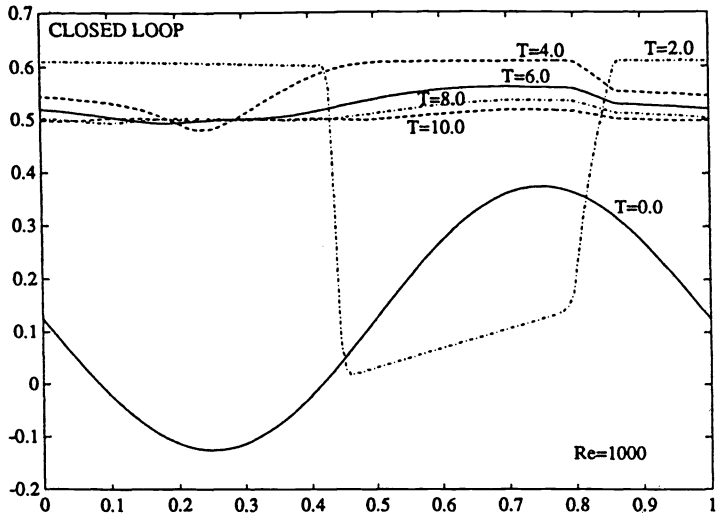


FIG. 2.3. *Closed loop* ($y_e = 0.5$).

(Fig. 2.3) are achieved by the control signals constructed from (2.18). Also, during the transition process, the weak shocks are smoothed out. However, a local weak shock is created by the controller. It is interesting to observe that, in Figure 2.3, huge control actions are needed in the beginning (e.g., up to time $t = 2.0$) to accelerate the flow to the high energy level $y_e = 0.5$ compared with the initial equilibrium state 0.125. Actually, the control signal $-\gamma(t)\mathcal{B}^*(y(t) - y_e)$ strongly depends on local property of

the solution.

For numerical computations, the Fourier-collocation method for space discretization and the two-step implicit method for time integration were used. To initialize the data, the Crank–Nicholson method was used (see [G]).

Example 3 (Navier–Stokes equations). In this example, we consider a velocity field control problem for the two-dimensional Navier–Stokes equations with periodic boundary conditions. For simplicity, the period in space is chosen to be 1. The governing equations are given by

$$
\begin{aligned}
&\frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \nabla p = \begin{pmatrix} b_1(x) \\ b_2(x) \end{pmatrix} f(t), \\
&\nabla \cdot u = 0, \\
&u(t, x + e_i) = u(t, x), \quad x \in \mathbf{R}^2, \quad t > 0, \\
&u(x, 0) = u_0(x),
\end{aligned} \tag{3.11}
$$

where $x = (x_1, x_2) \in \mathbf{R}^2$, $u = u(t,x) = (u_1(t,x), u_2(t,x))$ is the velocity vector, $p = p(t,x)$ is the pressure, $f(t)$ is the control signal to be determined, $e_1 = (1,0)$ and $e_2 = (0,1)$ are the canonical basis elements of $\mathbf{R}^2$. For numerical test, the nondimensional viscosity ν and the control distribution vector $b(x) = (b_1(x_1, x_2), b_2(x_1, x_2))$ are chosen as $\nu = 0.01$ and

$$
b_1(x_1, x_2) = \begin{cases} 5, & \text{if } |x_1 - 0.5| \le 0.1, \text{ and} |x_2 - 0.4| \le 0.1, \\ 0, & \text{otherwise,} \end{cases} \qquad b_2(x_1, x_2) = 0. \tag{3.12}
$$

The initial velocity vector field $u_0(x)$ is induced by the following initial vorticity function $\omega_0(x_1, x_2)$

$$
= \begin{cases} 10(1 + \cos(\frac{\pi(x_1 - 0.5)}{0.2}))\,(1 + \cos(\frac{\pi(x_2 - 0.5)}{0.2})) + c, & |x_i - 0.5| \le 0.2,\ i = 1, 2, \\ c, & \text{otherwise,} \end{cases} \tag{3.13}
$$

where the constant c is chosen so that the discrete compatibility condition is satisfied. That is, the initial velocity field $u_0(x)$ is determined by

$$
u_0(x) = \left(\frac{\partial}{\partial x_2} \psi(x_1, x_2),\ -\frac{\partial}{\partial x_1} \psi(x_1, x_2) \right) + (0.5, 0),
$$

where the stream function $\psi(x)$ satisfies $-\Delta \psi = w_0(x)$ with periodic boundary conditions. This test example is motivated by the numerical study in [BP] in which $(0,0)$ is chosen as constant term in above. Note that the constant velocity field (0.5,0) is an equilibrium state of the Navier–Stokes equations (3.11) with $f(t) \equiv 0$. In this example, the solution will be driven to the zero equilibrium state $u_e = (0,0)$ by the control signal $f(t)$. Also, energy dissipation effects of the feedback law will be demonstrated.

Let $\Omega = (0,1) \times (0,1)$. Consider the following function spaces (see [Te]).

$$
\begin{aligned}
V &= \{ u \in H_p^1(\Omega) \times H_p^1(\Omega) \,:\, \nabla \cdot u = 0 \}, \\
H &= \{ u \in L^2(\Omega) \times L^2(\Omega) \,:\, \nabla \cdot u = 0 \},
\end{aligned} \tag{3.14}
$$

where H_p^1 is the completion of the set of all C^∞ periodic functions with period Ω with respect to the $H^1(\Omega)$ norm.

The Stokes operator $\mathcal{A}$ is defined by

$$
\langle \mathcal{A}u, v \rangle_H = a(u,v) = \sum_{i=1}^{2} \int_\Omega \nabla u_i \cdot \nabla v_i \, dx \tag{3.15}
$$

for $u, v \in V$ and it is given by

$$
\mathcal{A}u = -\Delta u \quad \text{for all } u \in \mathcal{D}(\mathcal{A}) = H^2(\Omega) \times H^2(\Omega) \cap V \tag{3.16}
$$

due to the periodic boundary conditions. For any u, v, $w \in V$, define the trilinear form

$$
b(u; v, w) = \sum_{i=1}^{2} \int_\Omega u_i D_i v_j w_j \, dx \tag{3.17}
$$

and the bilinear continuous operator B from $V \times V$ into V^* by

$$
\langle B(u,v), w \rangle_{V^*,V} = b(u; v, w), \tag{3.18}
$$

where V^* is the dual space of V.

With the bilinear form a and the trilinear form b, the variational form of the control system (3.11) becomes

$$
\begin{aligned}
\frac{d}{dt}\langle u, v \rangle + \nu a(u,v) + b(u,u,v) &= \langle b(x) f(t), v \rangle, \qquad v \in V, \\
u(0) &= u_0.
\end{aligned} \tag{3.19}
$$

It is easy to observe that the pressure term ∇p in (3.11) is dropped in the variational form (3.19) due to the divergence free condition. It is well known [Te] that the variational equation (3.19) with $b(x)f(t) \in V^*$ has a unique global weak solution $u \in L^2(0,T;V) \cap C(0,T;H) \cap H^1(0,T;V^*)$ for $T \geq 0$.

We now define the nonlinear operator $\mathcal{F} \in \mathcal{L}(V, V^*)$ and the control input operator $\mathcal{B} \in \mathcal{L}(\mathbf{R}, H)$ by

$$
\mathcal{F}(u) = \mathrm{P}B(u,u) \quad \text{and} \quad \mathcal{B}f = (\mathrm{P}b(x))f \tag{3.20}
$$

for all $u \in V$ and $f \in \mathbf{R}$, where P is the projection operator from $H_p^0(\Omega) \times H_p^0(\Omega)$ onto the state space H. Then it is easy to see that the operator $\mathcal{F}$ satisfies assumptions (A1) and (A2) in §2. Thus, we can construct the feedback control signal $f(t)$ from (2.18), (2.19), i.e.,

$$
f(t) = -\gamma(t) \int_\Omega b(x) \cdot (u(t,x) - u_e) \, dx, \tag{3.21}
$$

TABLE 2
Energy comparison (Navier–Stokes equations).

time	Open		Closed		
	$\langle u, \mathcal{A}u\rangle$	$\langle u, u\rangle$	$\langle u, \mathcal{A}u\rangle$	$\langle u, u\rangle$	$\gamma(t)$
$t = 0.0$	33.440	0.627	33.440	0.627	0.000
$t = 0.1$	24.376	0.569	18.099	0.431	24.333
$t = 0.2$	18.865	0.526	15.662	0.356	38.484
$t = 0.3$	15.188	0.492	13.736	0.315	44.675
$t = 0.4$	12.573	0.464	11.838	0.284	45.837
$t = 0.5$	10.622	0.441	10.275	0.259	46.392
$t = 1.0$	5.436	0.364	5.854	0.173	48.512
$t = 2.0$	1.979	0.297	2.431	0.093	49.638
$t = 4.0$	0.367	0.259	0.672	0.040	49.498
$t = 6.0$	0.077	0.252	0.273	0.023	49.274
$t = 8.0$	0.016	0.250	0.147	0.015	49.116

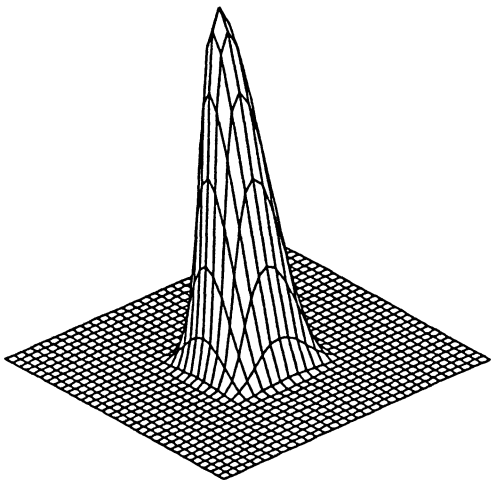


FIG. 3.1. *Vorticity (Navier–Stokes equations). (open loop, $\nu = 0.01$, $T = 0.0$).*

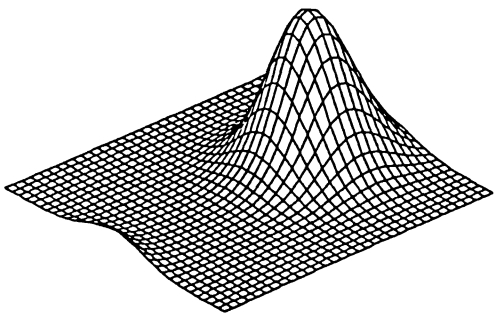


FIG. 3.2. *Vorticity (open loop, $T = 0.5$).*

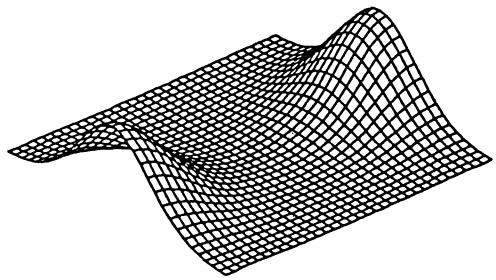


FIG. 3.3. *Vorticity (open loop, $T = 1.0$).*

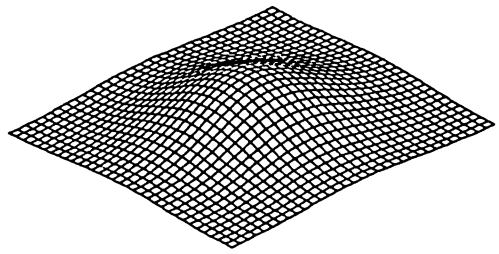


FIG. 3.4. *Vorticity (open loop, $T = 2.0$).*

where

$$\gamma(t) = \frac{-\epsilon\,|\nabla(u-u_e)|_H^2 + \sqrt{\epsilon^2|\nabla(u-u_e)|_H^2 + g(u)(\int_\Omega b(x)\cdot(u(x)-u_e)dx)^2}}{|\int_\Omega b(x)\cdot(u(x)-u_e)dx|^2}$$

with $g(u) = |\nabla(u-u_e)|^2 + (\int_\Omega b(x)\cdot(u(x)-u_e)dx)^2$.

Figures 3.1–3.4 and 3.10–3.12 show vorticity plots of the open-loop ($f(t) \equiv 0$) and the closed-loop solutions from time t=0.0 to t=2.0. Figures 3.5–3.8 and 3.13–3.15 show the corresponding velocity vector fields. In all velocity field plots, vectors are scaled so that the longest vectors are of equal lengths. Recall that the constant term of the initial velocity field $u_0(x)$ was set to be $(0.5, 0)$ and the constant velocity $(.5, 0)$ is an equilibrium state of the open-loop system (3.11). From Figs. 3.1–3.8 and Table 2, it is easy to see the following observations for open-loop trajectories.

(i) The center ("eye") of the initial vortex blob moves along the x_1-axis.

(ii) Due to diffusion effects, the peak of vortex is smoothed out eventually. Here, the viscosity $\nu = 0.01$ is chosen.

(iii) The velocity vector field $u(t,x) = (u_1(x), u_2(x))$ approaches the uniform flow $(0.5, 0)$, the equilibrium state. In Table 2, we can see that $|u(t,\cdot)|_H^2 = \langle u, u\rangle$ converges to $0.25 = |(.5, 0)|_H^2$, as time t increases.

Remark 3.1. We observe from numerical experiments that the open-loop velocity field $u(x) = (u_1(x), u_2(x))$ becomes an "almost" uniform flow $u_e = (0.5, 0)$ after time $t = 5.0$.

Figure 3.9 shows the projected vector field of the control distribution vector $b(x)$

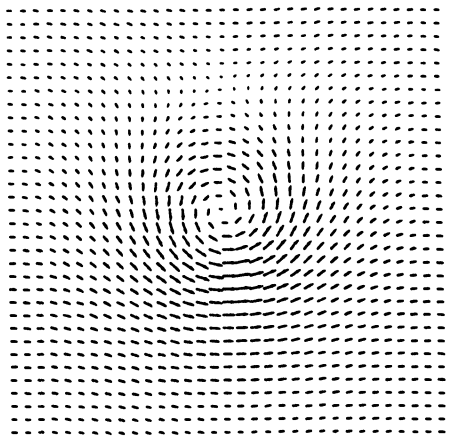


FIG. 3.5. *Velocity fields (open loop, $T = 0.0$).*

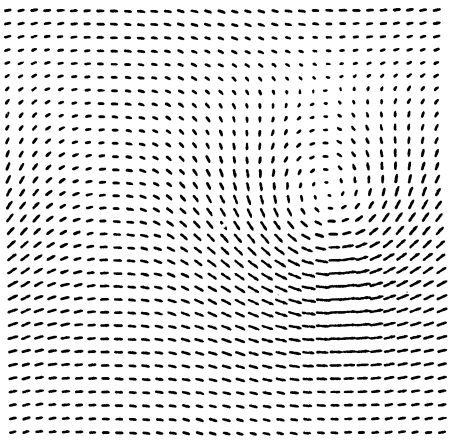


FIG. 3.6. *Velocity fields (open loop, $T = 0.5$).*

onto the divergence free space. The control input vector $b(x)$ is given by (3.12). The velocity vector fields and the corresponding vorticities of the closed-loop system with the control signal $f(t)$ are shown in Figs. 3.10–3.15. We can observe how this controller changes the global nature of the flow. The eye of vortex blob moves to the region $\{(x_1, x_2) \in \Omega : |x_1 - 0.5| \leq 0.1, |x_2 - 0.4| \leq 0.1\}$ where the control distribution vector $b(x)$ is located. Also, the flow changes its direction. "Sucking" actions of the controller for dropping the high energy level to the zero state (desired equilibrium state) are shown in Figs. 3.10–3.12. In the beginning, the energies $\langle u, \mathcal{A}u \rangle$ and $\langle u, u \rangle$ of the closed-loop system are reduced by the controller (Table 2). But, after time $t = 1.0$, the energy $\langle u, \mathcal{A}u \rangle$ goes to zero slowly compared with the open-loop system. This may be due to the fact that a continuing regulation effect of the controller produces a local vorticity in a neighborhood of the controller. The solution itself also approaches the zero state (desired state), i.e., $u(t, x) = (u_1(t, x), u_2(t, x)) \to (0, 0)$ as time t increases, since $\langle u, u \rangle = |u|_H^2 \to 0$ as $t \to \infty$. Note that the control input vector

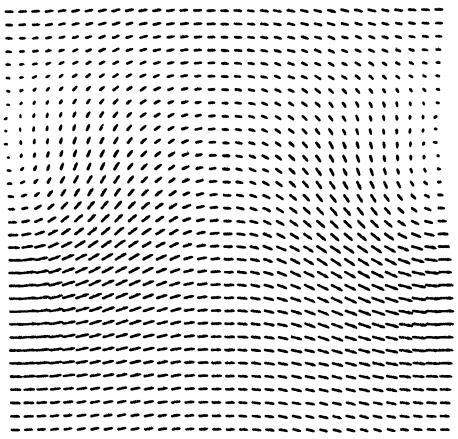


FIG. 3.7. *Velocity fields (open loop, $T = 1.0$).*

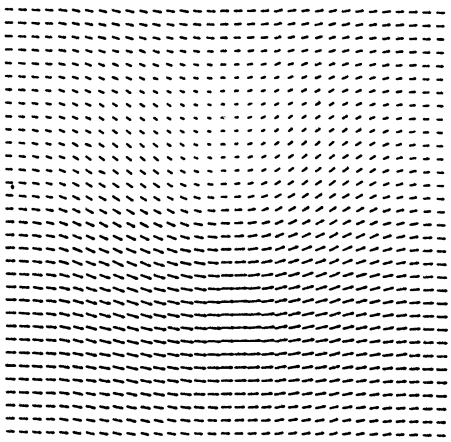


FIG. 3.8. *Velocity fields (open loop, $T = 2.0$).*

$b(x)$ satisfies the condition in Theorem 2.7 with $x_e \equiv 0$. Hence, the control signal $f(t)$ obtained from the formula (3.21) drives the given system (3.11) to the zero state $(0,0)$. On the other hand, as we mentioned before, the open-loop solution approaches the state $(0.5,0)$. Finally, as we expect, the feedback law $\gamma(t)$ converges to $50 = \frac{1}{2\nu}$, $\nu = 0.01$ (see Table 2 and Theorem 2.7). From numerical experiments, we observed that the control action continues "sucking" and "blowing" depending on the sign of $\int_\Omega b(x) \cdot (u(t,x) - u_e)dx$.

For this example, the Fourier–Galerkin scheme for space approximation and the Alternating Directional method for time integration were used [G].

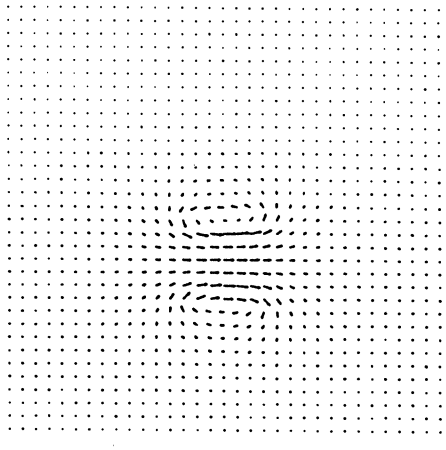


FIG. 3.9. *Control distribution Vector B.*

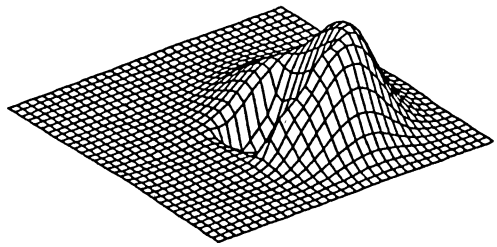


FIG. 3.10. *Vorticity (closed loop, $T = 0.5$).*

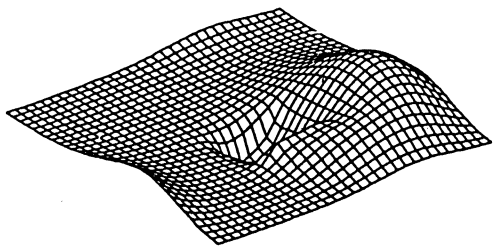


FIG. 3.11. *Vorticity (closed loop, $T = 1.0$).*

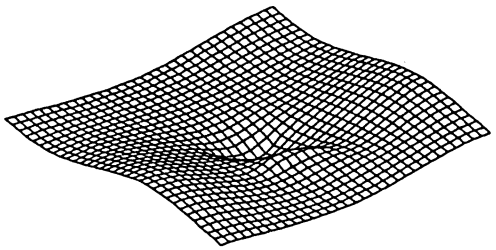


FIG. 3.12. *Vorticity* (*closed loop*, $T = 2.0$).

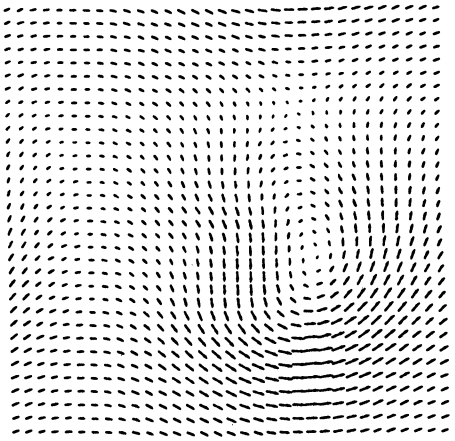


FIG. 3.13. *Velocity fields* (*closed loop*, $T = 0.5$).

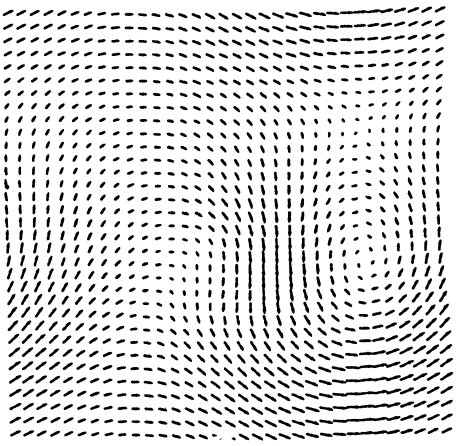


FIG. 3.14. *Velocity fields* (*closed loop*, $T = 1.0$).

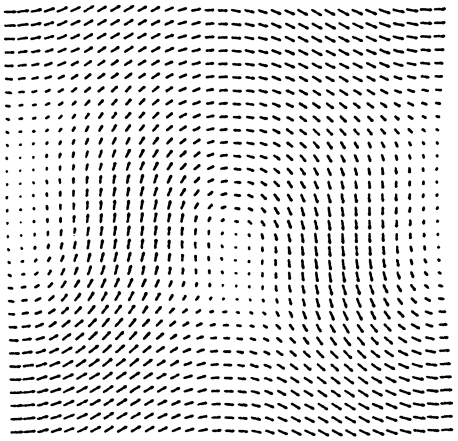


FIG. 3.15. *Velocity fields (closed loop, $T = 2.0$).*

REFERENCES

[B] A. V. BALAKRISHNAN, *Minimum attainable rms attitude error using co-located rate sensors*, in Proc. 5th Annual NASA Spacecraft Control Laboratory Experiment (SCOLE) Workshop, Lake Arrowhead, CA., 1988, pp. 357–367.

[BD] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, London, 1983.

[BI] H. T. BANKS AND K. ITO, *On a Variational Approach to a Class of Boundary Control Problems: Numerical Aanalysis*, preprint, 1993.

[BIK] J. A. BURNS, K. ITO, AND S. KANG, *Unbounded observation and boundary control problems for Burgers equation*, in Proc. 30th IEEE Conference on Decision and Control, December, 1991, Brighton, UK, pp. 2687–2692.

[BK1] J. A. BURNS AND S. KANG, *A control problem for Burgers' equation with bounded input/output*, ICASE Report 90-45, 1990, NASA Langley Research Center, Hampton, VA; Nonlinear Dynamics, 2 (1991), pp. 235–262.

[BK2] ———, *A stabilization problem for Burgers' equation with unbounded control and observation*, in Estimation and Control of Distributed Parameter Systems, C. W. Desch, F. Kappel and K. Kunisch, eds., International Series in Numerical Mathematics 100, Birkhäuser Verlag, Basel, 1991, pp. 51–72.

[BP] C. BÖRGERS AND C. S. PESKIN, *A Lagrangian fractional step method for the incompressible Navier–Stokes equations on a periodic domain*, J. Comput. Phys., 70 (1987), pp. 397–438.

[CF] P. CONSTANTIN AND C. FOIAS, *Navier-Stokes Equations*, The University of Chicago Press, Chicago, 1988.

[CL] M. G. CRANDALL AND P. L. LIONS, *Hamilton–Jacobi equations in infinite dimensions*, Parts I–V, J. Functional Anal., 62 (1985), pp. 379–396; 65 (1986), pp. 368–405; 68 (1986), pp. 214–247; 90 (1990), pp. 237–283; 97 (1991), pp. 417–465.

[FR] W. FLEMING AND R. RISHEL, *Deterministic and Stochastic Control*, Springer-Verlag, Berlin, 1975.

[G] R. GLOWINSKI, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[L] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.

[P] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[PS] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic control problem for infinite-dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[Sr] S. S. SRITHARAN, *Dynamic programming of Navier–Stokes equations*, Systems Control Letts., 16 (1991), pp. 299–307.

[SB] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, New York, 1980.

[Ta] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.

[Te] R. TEMAM, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 855–875, May 1994

012

DECENTRALIZED POLE ASSIGNMENT AND PRODUCT GRASSMANNIANS*

XIAOCHANG WANG†

Abstract. The pole assignment problems of linear systems by decentralized static output feedback are considered in this paper. A compactification of decentralized static feedback space, product Grassmannians, is introduced in this paper. Its degree under Plücker–Segre embedding is computed. Sufficient conditions for arbitrary and almost arbitrary pole assignability are given. It is also proved that the generic $m \times p$ system of McMillan degree n has arbitrary pole assignability by r-channel decentralized static output with m_i inputs and p_i outputs on the ith channel if $\sum_{i=1}^{r} m_i p_i \geq n$ when the degree of the product Grassmannians is odd, or if $\sum_{i=1}^{r} m_i p_i > n$ and the local channels have either the same numbers of inputs or the same numbers of outputs when the degree of the product Grassmannians is even.

Key words. linear systems, decentralized control, pole assignment, product Grassmannians, central projection, degree of variety

AMS subject classifications. 93B55, 93B27

1. Introduction.

In this paper, we investigate the pole assignment problem by decentralized static output feedback. The problem is formulated as follows: Let

$$\sigma: \quad \dot{x} = Ax + \sum_{i=1}^{r} B_i u_i, \quad y_i = C_i x, \qquad i = 1, 2, \ldots, r \tag{1}$$

be an r-channel linear system, where x, u_i, y_i are n, m_i, p_i vectors over $\mathbf{R}$, respectively, and u_i and y_i are the input and output of the ith channel. Under what conditions can the poles of σ be assigned to any self-conjugate set of n complex numbers by applying decentralized feedback

$$u_i = K_i y_i, \qquad i = 1, 2, \ldots, r \tag{2}$$

to the system?

Decentralized control is often applied to large-scale systems such as power systems, socioeconomic systems, large-scale space stations, and so forth. Centralized control of such systems is either uneconomical or unreliable due to long-distance information transfer between local control stations.

The decentralized pole assignment problem is fairly well understood if dynamic compensators are allowed in the feedback loop. Wang and Davison [16] proved that decentralized stabilization using local dynamic feedback is possible if and only if the fixed modes are stable. Corfmat and Morse [6] proved that a strongly connected system can be made controllable and observable through a single channel by local static feedback if and only if the set of fixed modes is empty. Thus the poles of such a system can be assigned freely by applying local dynamic feedback to one channel and local static feedback to all the other channels.

* Received by the editors December 11, 1991; accepted for publication (in revised form) December 3, 1992.

† Department of Mathematics, Texas Tech University, Lubbock, Texas 79409-1042 (mdxia@ttacs1.ttu.edu). This research was supported in part by National Science Foundation grant DMS-9224541 and the Research Enhancement Fund from College of Art and Sciences, Texas Tech University.

It is a standard technique to study the pole assignment problems through so-called "pole assignment map," i.e., the map that assigns each feedback compensator the closed-loop characteristic polynomial. A system has the arbitrary pole assignability if and only if the pole assignment map is onto. It is also a standard technique to extend the pole assignment map continuously to a compact set that contains the compensator space as a subset and study the extended pole assignment map. An example of using these techniques is the (centralized) static output feedback pole assignment problem. In 1981, Brockett and Byrnes [2] explained the problem as an intersection problem in a certain Grassmann variety $\mathrm{Grass}(p, m+p)$. This variety can be considered a compactification of the compensator space. In making the connection to the classical Schubert calculus, they were able to show that there are

$$\deg \mathrm{Grass}(p, m+p) = \frac{1!2!\cdots(p-1)!(mp)!}{m!(m+1)!\cdots(m+p-1)!} \tag{3}$$

complex static output feedback laws, which assign each set of poles for the generic m-input, p-output linear system of McMillan degree $n = mp$. In particular, if (3) is odd, real solutions always exist. It follows that the generic system has the arbitrary pole assignability if $mp \geq n$ and if $\deg \mathrm{Grass}(p, m+p)$ is odd. Using this model, we proved in [19] that the generic system has the arbitrary pole assignability if $mp > n$ and $\deg \mathrm{Grass}(p, m+p)$ is even.

Adopting these ideas to the decentralized pole assignment problems, we introduce a compactification of the decentralized feedback space, a product of Grassmannians in this paper. The degree of product Grassmannians under Plücker–Segre embedding is computed in §2. Sufficient conditions for arbitrary and almost arbitrary pole assignability are given in §3. Generic pole assignability is considered in §4. It is proved in §4 that, when the degree of a product of Grassmannians is odd and $n \leq \sum_{i=1}^{r} m_i p_i$, or when the degree is even and $n < \sum_{i=1}^{r} m_i p_i$, and either $m_1 = m_2 = \cdots = m_r$ or $p_1 = p_2 = \cdots = p_r$, then the generic system of McMillan degree n has arbitrary pole assignability by decentralized static output feedback.

This paper extends the results in [20]. In [20] we considered only pole assignment with local state feedback. The compactification in that case was a product of projective spaces, which was a special kind of product Grassmannian. We only studied the cases in which the degrees of the product projective spaces were odd in [20].

2. Decentralized pole assignment map and product Grassmannians. Let

$$m = \sum_{i=1}^{r} m_i, \qquad p = \sum_{i=1}^{r} p_i \tag{4}$$

and assume that

$$m \leq n, \qquad p \leq n. \tag{5}$$

If we apply the decentralized feedback (2) to (1), then the closed-loop system becomes

$$\dot{x} = \left(A + \sum_{i=1}^{r} B_i K_i C_i\right) x, \quad y_i = C_i x, \qquad i = 1, 2, \ldots, r. \tag{6}$$

Let

$$B = [B_1, B_2, \dots, B_r], \qquad C = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_r \end{bmatrix} \tag{7}$$

and

$$K = \text{ block diag}(K_1, K_2, \dots, K_r). \tag{8}$$

We define the pole assignment map $\phi_\sigma : \mathbf{R}^{m_i p_i} \times \cdots \times \mathbf{R}^{m_r p_r} \to \mathbf{R}^n$ by

$$\phi_\sigma(K_1, \dots, K_r) = \det(sI - A - BKC), \tag{9}$$

where a characteristic polynomial $s^n + a_1 s^{n-1} + \cdots + a_n$ is identified with a point $(a_1, \dots, a_n)$ in $\mathbf{R}^n$. If σ is controllable and observable, then ϕ_σ can be expressed as

$$\phi_\sigma(K_1, \dots, K_r) = \det \begin{bmatrix} I_1 & & & \\ & \ddots & & N(s) \\ & & I_r & \\ K_1 & & & \\ & \ddots & & D(s) \\ & & K_r & \end{bmatrix}, \tag{10}$$

where I_i is the $p_i \times p_i$ identity matrix and $N(s)D^{-1}(s)$ is a right coprime fraction of the transfer function $C(sI - A)^{-1}B$ [2], [3], [17].

Let $\text{Grass}_K(p, m+p)$ be the Grassmannian of all p-dimensional subspaces of an $(m+p)$-dimensional vector space V_{m+p} over a field $\mathbf{K}$. Fix a basis of V_{m+p}; each point $z \in \text{Grass}_K(p, m+p)$ can be represented by an $(m+p) \times p$ full-rank matrix Z such that col sp $Z = z$. Two matrices Z_1 and Z_2 represent the same point if and only if there is a $Q \in GL(p)$ such that

$$Z_1 Q = Z_2. \tag{11}$$

$\text{Grass}_K(p, m+p)$ can be embedded in $\mathbf{P}^N$ by Plücker embedding [8], where

$$N = \binom{m+p}{m} - 1.$$

For each $z \in \text{Grass}_K(p, m+p)$, let Z be a matrix representation of z. The homogeneous Plücker coordinate of z in $\mathbf{P}^N$ is given by $z = (z_{\underline{i}})$, where $\underline{i} = (i_1, i_2, \dots, i_m), 1 \le i_1 < i_2 < \cdots < i_m \le m+p$ ranges over all multi-indices and where $z_{\underline{i}}$ is the p-minor of Z formed by the i_1th, i_2th, $\dots$, i_mth rows. The $\text{Grass}_K(p, m+p) \subset \mathbf{P}^N$ is defined by quadratic equations [8].

Let $\mathbf{P}^n$ and $\mathbf{P}^m$ be two projective spaces. For any $x = (x_0, x_1, \dots, x_n) \in \mathbf{P}^n$ and $y = (y_0, y_1, \dots, y_m) \in \mathbf{P}^m$, the map $S : \mathbf{P}^n \times \mathbf{P}^m \to \mathbf{P}^{mn+m+n}$ defined by $S(x, y) = x^t y$ induces an embedding $\mathbf{P}^n \times \mathbf{P}^m \subset \mathbf{P}^{mn+m+n}$, where the homogeneous Segre coordinate of (x, y) is given by the entries of the $(n+1) \times (m+1)$ matrix $x^t y$. This embedding is called the Segre embedding [12], [15]. If the homogeneous coordinates of $\mathbf{P}^{mn+m+n}$ are written as $(n+1) \times (m+1)$ matrices $\{W\}$, then $\mathbf{P}^n \times \mathbf{P}^m \subset \mathbf{P}^{mn+m+n}$

is defined by rank $(W) = 1$; i.e., all the 2×2 minors are zero. Similarly, for each $x_i = (x_{i0}, x_{i1}, \ldots, x_{in_i}) \in \mathbf{P}^{n_i}$, $i = 1, 2, \ldots, r$, $S(x_1, \ldots, x_r) = (x_{1i_1} x_{2i_2} \cdots x_{ri_r})_{i_j=0}^{n_j}$ defines the Segre embedding $\mathbf{P}^{n_1} \times \cdots \times \mathbf{P}^{n_r} \subset \mathbf{P}^N$, $N = \prod_{i=1}^r (n_i + 1) - 1$.

Let

$$\Pi_K^\alpha = \mathrm{Grass}_K(p_1, m_1 + p_1) \times \cdots \times \mathrm{Grass}_K(p_r, m_r + p_r), \tag{12}$$

where

$$\alpha = (p_1, \ldots, p_r, m_1, \ldots, m_r); \tag{13}$$

then Π_K^α is a projective variety. We can embed each $\mathrm{Grass}_K(p_i, m_i + p_i)$ into projective n_i-space $\mathbf{P}_K^{n_i}$,

$$n_i = \binom{m_i + p_i}{p_i} - 1,$$

using the Plücker embedding, and then embed $\mathbf{P}_K^{n_i} \times \cdots \times \mathbf{P}_K^{n_r}$ into the projective N-space $\mathbf{P}_K^N$, $N = \prod_{i=1}^r \binom{m_i + p_i}{p_i} - 1$, using the Segre embedding.

Following [17], it is possible to write (10) as

$$\phi_\sigma(K_1, \ldots, K_r) = \sum_{i=0}^{M} w_i g_i(s), \tag{14}$$

where $M = \binom{m+p}{p} - 1$ and $(w_0, \ldots, w_M)$ are the Plücker coordinates of

$$\begin{bmatrix} I_1 & & \\ & \ddots & \\ & & I_r \\ K_1 & & \\ & \ddots & \\ & & K_r \end{bmatrix} \in \mathrm{Grass}_R(p, m + p).$$

Note that some of the w_i's are identically zero. A w_i is not identically zero if and only if $w_i = \prod_{j=1}^r z_{ji_j}$, where z_{ji_j} is a $p_j \times p_j$ minor of

$$\begin{bmatrix} I_j \\ K_j \end{bmatrix};$$

namely, w_i is a component of Plücker–Segre coordinate of

$$\left(\begin{bmatrix} I_1 \\ K_1 \end{bmatrix}, \begin{bmatrix} I_2 \\ K_2 \end{bmatrix}, \ldots, \begin{bmatrix} I_r \\ K_r \end{bmatrix} \right) \in \Pi_R^\alpha.$$

Let $z = (z_0, \ldots, z_N)$ be the Plücker–Segre coordinate of

$$\left(\begin{bmatrix} I_1 \\ K_1 \end{bmatrix}, \begin{bmatrix} I_2 \\ K_2 \end{bmatrix}, \ldots, \begin{bmatrix} I_r \\ K_r \end{bmatrix} \right) \in \Pi_R^\alpha.$$

Then (14) becomes

$$\phi_\sigma(z) = \sum_{i=0}^{N} z_i g_i(s), \tag{15}$$

and (15) can be extended to a rational map on $\mathbf{P}_C^N$, where $\Pi_C^\alpha \subset \mathbf{P}_C^N$ by the Plücker–Segre embedding. The extended map is also denoted by ϕ_σ.

Let

$$E_\sigma = \{z \in \mathbf{P}_C^N | \phi_\sigma(z) \equiv 0\}. \tag{16}$$

Then (15) defines a central projection $\mathbf{P}_C^N - E_\sigma \to \mathbf{P}_C^n$ with center E_σ (see [17]), where a polynomial $a_0 s^n + a_1 s^{n-1} + \cdots + a_n$ is identified with a point $(a_0, a_1, \ldots, a_n) \in \mathbf{P}_C^n$. σ has arbitrary (almost arbitrary) pole assignability if

$$\phi_\sigma : \Pi_R^\alpha - E_\sigma \to \mathbf{P}_R^n$$

is onto (almost onto).

PROPOSITION 2.1. *Let X_i be a projective variety of dimension n_i and degree m_i, $i = 1, 2, \ldots, r$. Then the degree of $X_1 \times \cdots \times X_r$ under Segre embedding is*

$$\deg(X_1 \times \cdots \times X_r) = \frac{(n_1 + \cdots + n_r)!}{n_1! \cdots n_r!} \prod_{i=1}^{r} m_i.$$

Proof. We first prove that the Hilbert polynomial (see [9] for definition) of $X_1 \times \cdots \times X_r$ under Segre embedding is the product of the Hilbert polynomials of $\{X_i,\ i = 1, \ldots, r\}$. By induction, we only need to prove it for $r = 2$.

Assume that $X \subset \mathbf{P}^n$ and $Y \subset \mathbf{P}^m$. Let R_l be the $\mathbf{C}$-module consisting of all the homogeneous elements of degree l in the graded ring R. After the substitution $z_{ij} = x_i y_j$, any homogeneous polynomial of degree l in $\mathbf{C}[X \times Y]$ can be considered as a homogeneous polynomial both in $x_0, \ldots, x_n$ and $y_0, \ldots, y_m$ of the same degree of homogeneity l.

Take basis $\{f_1, \ldots, f_s\}$ of $\mathbf{C}[X]_l$ and $\{g_1, \ldots, g_s\}$ of $\mathbf{C}[Y]_l$. Then $\{h_{ij} = f_i g_j\}$ span the $\mathbf{C}[X \times Y]_l$. If

$$\sum_{i,j} a_{ij} f_i(x) g_j(y) \in I(X \times Y),$$

then

$$\sum_i a_{i1} f_i(x) g_1(y) + \cdots + \sum_i a_{im} f_i(x) g_m(y)$$

in $I(Y)$ for each fixed x in X. Since $\{g_j\}$ is a basis of $\mathbf{C}[Y]$, $\sum_i a_{ij} f_i(x) = 0$ for $x \in X$, i.e., $\sum_i a_{ij} f_i(x) \in I(X)$. So $a_{ij} = 0$. Therefore $\{h_{ij} = f_i g_j\}$ is a basis of $\mathbf{C}[X \times Y]_l$ and

$$\dim \mathbf{C}[X \times Y]_l = (\dim \mathbf{C}[X]_l)(\dim \mathbf{C}[Y]_l),$$

so the Hilbert polynomial of $X \times Y$ is the product of the Hilbert polynomials of X and Y.

Let $f_i(z)$ be the Hilbert polynomial of X_i, $i = 1, 2, \ldots, r$. Then

$$f_i(z) = \frac{m_i}{n_i!} z^{n_i} + \cdots.$$

The Hilbert polynomial of $X_1 \times \cdots \times X_r$ under Segre embedding is

$$f(z) = f_1(z) f_2(z) \cdots f_r(z) = \left(\prod_{i=1}^{r} \frac{m_i}{n_i!} \right) z^{n_1 + \cdots + n_r} + \cdots.$$

Therefore

$$\deg(X_1 \times \cdots \times X_r) = \frac{(n_1 + \cdots + n_r)!}{n_1! \cdots n_r!} \prod_{i=1}^{r} m_i. \qquad \square$$

PROPOSITION 2.2. *The* $\deg(X_1 \times \cdots \times X_r)$ *is odd if and only if the degree of all* X_i *are odd and the sets of exponents appearing in the binary representations of* $n_1, n_2, \ldots, n_r$ *are disjoint, where* $n_i = \dim X_i$, $i = 1, 2, \ldots, r$.

Proof. The second condition is equivalent to the fact that

$$\frac{(n_1 + \cdots + n_r)!}{n_1! n_2! \cdots n_r!}$$

is odd [20]. $\square$

COROLLARY 2.1. *It holds that*

$$\deg \Pi_C^{\alpha} = (m_1 p_1 + \cdots + m_r p_r)! \prod_{i=1}^{r} \frac{1!2! \cdots (p_i - 1)!}{m_i!(m_i + 1)! \cdots (m_i + p_i - 1)!}.$$

Proof. Simply note that

$$\deg \operatorname{Grass}_C(p_i, m_i + p_i) = \frac{1!2! \cdots (p_i - 1)!(m_i p_i)!}{m_i!(m_i + 1)! \cdots (m_i p_i - 1)!}$$

(see [10]). $\square$

COROLLARY 2.2. *The* $\deg \Pi_C^{\alpha}$ *is odd if and only if* Π_C^{α} *is in one of the following forms:*

(i) $\operatorname{Grass}_C(2, 2^m + 1) \times \mathbf{P}_C^{n_1} \times \cdots \times \mathbf{P}_C^{n_r}$, *where* $n_i = 2^{m_{i1}} + 2^{m_{i2}} + \cdots + 2^{m_{ir_i}}$, $m_{ij} \neq m_{kl}$ *if* $(i, j) \neq (k, l)$, *and either* $m_{ij} = 1$ *or* $m_{ij} > m$;

(ii) $\mathbf{P}_C^{n_1} \times \cdots \times \mathbf{P}_C^{n_r}$, *where* $n_i = 2^{m_{i1}} + 2^{m_{i2}} + \cdots + 2^{m_{ir_i}}$ *and* $m_{ij} \neq m_{kl}$ *if* $(i, j) \neq (k, l)$.

Here $\operatorname{Grass}(2^h - 1, 2^h + 1)$ *and* $\operatorname{Grass}(n_i - 1, n_i)$ *are identified with* $\operatorname{Grass}(2, 2^h + 1)$ *and* $\mathbf{P}^{n_i}$, *respectively.*

Proof. $\deg \operatorname{Grass}_C(p, m + p)$ is odd if and only if

$$\operatorname{Grass}_C(p, m + p) = \operatorname{Grass}_C(2, 2^h + 1)$$

or

$$\operatorname{Grass}_C(p, m + p) = \mathbf{P}_C^n$$

(see [1]). $\square$

EXAMPLE 2.1. We have that

$$\deg(\mathbf{P}_C^1 \times \mathbf{P}_C^2 \times \mathbf{P}_C^4) = 105 \quad \text{and} \quad \deg(\mathbf{P}_C^1 \times \operatorname{Grass}_C(2, 3)) = 35.$$

3. Sufficient conditions for arbitrary pole assignability. In what follows, d_α denotes the degree of Π_C^α, i.e.,

$$d_\alpha = (m_1p_1 + \cdots + m_rp_r)! \prod_{i=1}^{r} \frac{1!2! \cdots (p_i - 1)!}{m_i!(m_i + 1)! \cdots (m_i + p_i - 1)!}. \tag{17}$$

THEOREM 3.1. *An r-channel system σ of McMillan degree n with $n \leq \sum_{i=1}^{r} m_ip_i$ and d_α odd has arbitrary pole assignability by decentralized static feedback if*

$$\dim(E_\sigma \cap \Pi_C^\alpha) = \sum_{i=1}^{r} m_ip_i - n - 1.$$

In this formula, we assume that the empty set has dimension -1.

Proof. When $\sum_{i=1}^{r} m_ip_i = n$, $E_\sigma \cap \Pi_C^\alpha = \emptyset$ implies that $\phi_\sigma : \Pi_C^\alpha \to \mathbf{P}_C^n$ is a finite morphism and hence is onto and has degree d_α [12], [15]. Since d_α is odd,

$$\phi_\sigma : \Pi_R^\alpha \to \mathbf{P}_R^n$$

is also onto.

When $\sum_{i=1}^{r} m_ip_i > n$, there exists a subspace $H \subset \mathbf{P}_C^N, \bar{H} = H$ of codimension $\sum_{i=1}^{r} m_ip_i - n$ such that

$$E_\sigma \cap H \cap \Pi_C^\alpha = \emptyset.$$

Let $E_1 = E_\sigma \cap H$, ϕ_1 be the projection with center E_1, $E_2 = \phi_1(E_\sigma)$ and $\phi_2 : \mathbf{P}_R^a - E_2 \to \mathbf{P}_R^n$ be the projection with center E_2. Then

$$\phi_1 : \Pi_R^\alpha \to \mathbf{P}_R^a, \qquad a = \sum_{i=1}^{r} m_ip_i$$

is onto. Therefore

$$\begin{aligned} \phi_\sigma(\Pi_R^\alpha - E_\sigma) &= \phi_2(\phi_1(\Pi_R^\alpha - E_\sigma)) \\ &= \phi_2(\phi_1(\Pi_R^\alpha - E_1) - E_2) \\ &= \phi_2(\mathbf{P}_R^a - E_2) = \mathbf{P}_R^n. \end{aligned}$$

□

THEOREM 3.2. *An r-channel system σ of McMillan degree n with $n < \sum_{i=1}^{r} m_ip_i$ and d_α even has arbitrary pole assignability by decentralized static feedback if*

$$\dim E_\sigma \cap \Pi_C^\alpha = \sum_{i=1}^{r} m_ip_i - n - 1,$$

and there is a $z_0 \in E_\sigma \cap \Pi_R^\alpha$ such that

$$\dim E_\sigma \cap T_{z_0} = \sum_{i=1}^{r} m_ip_i - n - 1,$$

where T_{z_0} is the tangent space of Π_C^α at z_0.

Proof. Apply Proposition 3.2 of [19] to ϕ_σ directly. Note that Π_C^α is nonsingular. □

THEOREM 3.3. *An r-channel system σ of McMillan degree n with $n < \sum_{i=1}^{r} m_i p_i$ and d_α even has almost arbitrary pole assignability by decentralized static feedback if*

$$\dim E_\sigma \cap \Pi_C^\alpha = \sum_{i=1}^{r} m_i p_i - n - 1,$$

and there is a $z_0 \in E_\sigma \cap \Pi_R^\alpha$ such that

$$\dim E_\sigma \cap T_{z_0} \leq \sum_{i=1}^{r} m_i p_i - n.$$

Proof. The proof follows from Proposition 3.2 of [19]. □

4. Generic pole assignability. Let $\Sigma_{m,p}^n$ be the set of all m-input, p-output systems of McMillan degree n. Recall that $\Sigma_{m,p}^n$ is a quasi-affine variety of dimension $n(m+p)$. A subset U of $\Sigma_{m,p}^n$ is said to be generic if U is open and dense in $\Sigma_{m,p}^n$. In particular, U is generic if U contains a nonempty Zariski open subset of $\Sigma_{m,p}^n$. Any system in a generic set is called a generic system.

In this section, we consider the conditions under which a generic system has arbitrary pole assignability. Since

$$\dim \Pi_R^\alpha = \sum_{i=1}^{r} m_i p_i,$$

a necessary condition is

$$\sum_{i=1}^{r} m_i p_i \geq n. \tag{18}$$

It is also sufficient if d_α is odd (Theorem 4.1) and is not sufficient if d_α is even (Theorem 4.2). We also prove in this section that

$$\sum_{i=1}^{r} m_i p_i > n \tag{19}$$

is a sufficient condition if d_α is even and all local channels have either the same numbers of inputs or the same numbers of outputs.

In what follows, it will be convenient to switch rows of determinant in (10) and write

$$\phi_\sigma(M_1, M_2, \ldots, M_r) =$$

$$\det \left[\begin{array}{cccc|cccc} M_{11} & 0 & \cdots & 0 & N_{11} & N_{12} & \cdots & N_{1r} \\ M_{12} & 0 & \cdots & 0 & D_{11} & D_{12} & \cdots & D_{1r} \\ 0 & M_{21} & \cdots & 0 & N_{21} & N_{22} & \cdots & N_{2r} \\ 0 & M_{22} & \cdots & 0 & D_{21} & D_{22} & \cdots & D_{2r} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & M_{r1} & N_{r1} & N_{r2} & \cdots & N_{rr} \\ 0 & 0 & \cdots & M_{r2} & D_{r1} & D_{r2} & \cdots & D_{rr} \end{array}\right], \tag{20}$$

where

$$M_i = \begin{bmatrix} M_{i1} \\ M_{i2} \end{bmatrix} \in \mathrm{Grass}_C(p_i, m_i + p_i)$$

is an $(m_i + p_i) \times p_i$ full-rank matrix and where M_{i1} and M_{i2} are $p_i \times p_i$ and $m_i \times p_i$ submatrices, respectively.

PROPOSITION 4.1. *Let $n \leq \sum_{i=1}^{r} m_i p_i$. The set of systems in $\Sigma_{m,p}^n$ such that*

$$\dim(E_\sigma \cap \Pi_C^\alpha) = \sum_{i=1}^{4} m_i p_i - n - 1$$

is a nonempty Zariski open set, where dimension -1 means that the intersection is empty.

Proof. For a proof, see the Appendix. □

By Theorem 3.1 and Proposition 4.1, we obtain the following result.

THEOREM 4.1. *When d_α is odd, the generic system in $\Sigma_{m,p}^n$ has arbitrary pole assignability by decentralized static feedback if and only if*

$$\sum_{i=1}^{r} m_i p_i \geq n.$$

The same result is not true if d_α is even.

THEOREM 4.2. *For any $\sigma \in \Sigma_{2,2}^2$ with*

$$E_\sigma \cap \Pi_C^{(1,1,1,1)} = \emptyset,$$

the set of closed-loop characteristic polynomials that cannot be achieved by decentralized static feedback with $p_1 = p_2 = m_1 = m_2 = 1$ is a nonempty open subset of $\mathbf{R}^2$ (in the classical topology).

Proof. The condition implies that $\phi_\sigma : \Pi_C^{(1,1,1,1)} \to \mathbf{P}_C^2$ is onto, and $\phi_\sigma^{-1}(y)$ contains two points (counted with multiplicity) for any $y \in \mathbf{P}_C^2$. The set of $y \in \mathbf{R}^2$, such that $p_\sigma^{-1}(y)$ contains two complex conjugate points, is an open set of $\mathbf{R}^2$. We only need to show that it is nonempty.

$\Pi_C^{(1,1,1,1)}$ is defined in $\mathbf{P}_C^3$ by

$$Q(z) = z_0 z_3 - z_1 z_2 = 0,$$

where $z_0 = x_0 y_0$, $z_1 = x_0 y_1$, $z_2 = x_1 y_0$, $z_3 = x_1 y_1$, and $(x_0, x_1), (y_0, y_1)$ are the homogeneous coordinates of $\mathbf{P}_C^1$'s. We have

$$\phi_\sigma(z) = \det \begin{bmatrix} x_0 & 0 & \\ & & N(s) \\ 0 & y_0 & \\ x_1 & 0 & \\ & & D(s) \\ 0 & y_1 & \end{bmatrix}$$

or

$$\phi_\sigma(z) = \sum_{i=0}^{3} z_i g_i(s), \qquad Q(z) = 0.$$

By assumption, E_σ contains only one real point. Assume that $E_\sigma = \{e\}, e = (0, e_1, e_2, e_3)$, where $e_0 = 0$, because $g_0(s) = \det D(s)$ is the only polynomial of degree 2. Since $Q(e) \neq 0$, $e_1 e_2 \neq 0$, and it can be easily checked that the real polynomial

$$g_0(s) - \frac{1 + e_3^2}{4 e_1 e_2} g_3(s)$$

cannot be achieved by any real feedback. □

Remark 4.1. This theorem is an analogue of Corollary 6.4 of [2]. Note that $E_\sigma \cap \Pi_C^{(1,1,1,1)} = \emptyset$ for a generic system σ in $\Sigma_{2,2}^2$.

PROPOSITION 4.2. *Assume that d_α is even and that*

$$S = \{\sigma \in \Sigma_{m,p}^n \mid E_\sigma \cap \Pi_R^\alpha \neq \emptyset\}$$

is a generic set. Then the pole assignment map $\phi_\sigma : \Pi_R^\alpha - E_\sigma \to \mathbf{P}_R^n$ is onto for the generic system σ in $\Sigma_{m,p}^n$.

Remark 4.2. $E_\sigma \cap \Pi_C^\alpha = \emptyset$ for the generic system if $n \geq \sum_{i=1}^r m_i p_i$. So the second condition implies that $n < \sum_{i=1}^r m_i p_i$.

Proof. We must prove that

$$\dim E_\sigma \cap T_z = \sum_{i=1}^{r} m_i p_i - n - 1$$

for some $z \in E_\sigma \cap \Pi_R^\alpha$ for the generic system σ. The set of such systems is certainly open. Consider in $\Sigma_{m,p}^n \times \Pi_R^\alpha$ the incidence set

$$X_R = \{(\sigma, z) \mid z \in E_\sigma \cap \Pi_{\mathbf{R}}^\alpha\}.$$

Define

$$X_1 = \left\{(\sigma, z) \in X_R \mid \dim E_\sigma \cap T_z = \sum_{i=1}^{r} m_i p_i - n - 1\right\}.$$

X_1 is a nonempty Zariski open set of X_R by the Appendix. So $\pi_1(X_1)$ is dense in $\pi_1(X_R)$ and therefore is dense in $\Sigma_{m,p}^n$ by assumption, where π_1 is the projection defined by

$$\pi_1(\sigma, z) = \sigma. \qquad \square$$

PROPOSITION 4.3. *$E_\sigma \cap \Pi_R^\alpha \neq \emptyset$ for all systems if*

$$n > \sum_{i=1}^{r} m_i p_i$$

and either $m_1 = m_2 = \cdots = m_r$ or $p_1 = p_2 = \cdots = p_r$.

Proof. Without loss of generality, assume that $p_1 = \cdots = p_r$. Let $\{v_1, \ldots, v_m\}$ be the controllability indices of σ with

$$v_1 \leq v_2 \leq \cdots \leq v_m.$$

Then

$$p_1 m = \sum_{i=1}^{r} p_i m_i > n = v_1 + v_2 + \cdots + v_m \geq v_1 m,$$

which means that $p_1 > v_1$. Let

$$\begin{bmatrix} d_1 \\ n_1 \\ \vdots \\ d_r \\ n_r \end{bmatrix}$$

be the first column of

$$\begin{bmatrix} N_{11} & \cdots & N_{1r} \\ D_{11} & \cdots & D_{1r} \\ \vdots & & \vdots \\ N_{r1} & \cdots & N_{rr} \\ D_{r1} & \cdots & D_{rr} \end{bmatrix}$$

in (20) with column degree v_1 (Such a coprime fraction always exists by [7]). The column degree of

$$\begin{bmatrix} d_i \\ n_i \end{bmatrix}$$

is less than p_i, so there is an $(m_i + p_i) \times p_i$ full-rank matrix

$$M_i = \begin{bmatrix} M_{i1} \\ M_{i2} \end{bmatrix}$$

such that

$$\begin{bmatrix} d_i \\ n_i \end{bmatrix} \in \text{col sp} \begin{bmatrix} M_{i1} \\ M_{i2} \end{bmatrix}, \qquad i = 1, 2, \ldots, r,$$

which implies that

$$\phi_\sigma(M_1, M_2, \ldots, M_r) \equiv 0,$$

i.e.,

$$(M_1, M_2, \ldots, M_r) \in E_\sigma \cap \Pi_R^\alpha. \qquad \square$$

By Proposition 4.2 and 4.3, we obtain the next theorem.

THEOREM 4.3. *When d_α is even, the generic system in $\Sigma_{m,p}^n$ has arbitrary pole assignability by decentralized static feedback if*

$$\sum_{i=1}^{r} m_i p_i > n$$

and either $p_1 = p_2 = \cdots = p_r$ or $m_1 = m_2 = \cdots = m_r$.

Appendix. Let Z be an $(m+p)\times p$ full-rank matrix in $\mathrm{Grass}(p,m+p)$ and $z=(z_{\underline{i}})$ be its Plücker coordinate, where $\underline{i}=(i_1,\dots,i_p)$ and $z_{\underline{i}}$ is the $p\times p$ minor of M formed by the i_1th through i_pth rows.

NOTATION A.1. The following hold:

1. $\underline{i}\le\underline{j}\Leftrightarrow i_l\le j_l,\ l=1,\dots,p$;
2. $|\underline{i}|=i_1+i_2+\cdots+i_p-p(p+1)/2$.

DEFINITION A.1. *Define*

$$S(\underline{i})=\{z\in\mathrm{Grass}(p,m+p)|z_{\underline{j}}=0 \text{ if } \underline{j}\not\le\underline{i}\},$$

$$C(\underline{i})=\{z\in S(\underline{i})|z_{\underline{i}}\neq 0\},$$

and

$$S^*(\underline{i})=\{z\in\mathrm{Grass}(p,m+p)|z_{\underline{j}}=0 \text{ if } \underline{i}\not\le\underline{j}\}.$$

Then $S(\underline{i})$ and $S^*(\underline{i})$ are Schubert varieties of dimension $|\underline{i}|$ and $mp-|\underline{i}|$, respectively, $C(\underline{i})$ is a Schubert cell, and

$$S(\underline{i})=\bigcup_{\underline{j}\le\underline{i}}C(\underline{j}). \tag{21}$$

For any $z=(z_{\underline{j}})\in C(\underline{i})$,

$$z_{\underline{j}}=0 \quad \text{if} \quad |\underline{j}|\ge|\underline{i}| \text{ and } \underline{j}\neq\underline{i}. \tag{22}$$

Let $m=m_1+\cdots+m_r$ and $p=p_1+\cdots+p_r$. Then from (20) we have that

$$\Pi_C^{\alpha}=S^*(\underline{k})\cap S(\underline{l})\subset\mathrm{Grass}(p,m+p),$$

where

$$k_t=m_1+\cdots+m_{s-1}+t \quad \text{and} \quad l_t=k_t+m_s \tag{23}$$

$$\text{for } p_1+\cdots+p_{s-1}<t\le p_1+\cdots+p_s, \quad p_0=m_0=0.$$

LEMMA A.1. *Let*

$$H_t=\left\{z=(z_{\underline{j}})\in\mathbf{P}^M\,\middle|\,\sum_{|\underline{j}|=i}z_{\underline{j}}=0,\ i=t+1,\dots,mp\right\}.$$

Then

$$H_t\cap\mathrm{Grass}(p,m+p)=\bigcup_{|\underline{j}|=t}S(\underline{j}).$$

Proof. Note that

$$\bigcup_{|\underline{j}|=t}S(\underline{j})\subset H_t\cap\mathrm{Grass}(p,m+p).$$

Take any $z = (z_{\underline{i}}) \in H_t \cap \text{Grass}(p, m+p)$. Then $z \in C(\underline{i})$ for some $\underline{i}$ and

$$\sum_{|\underline{j}|=|\underline{i}|} z_{\underline{j}} = z_{\underline{i}} \neq 0$$

by (22). Therefore $|\underline{i}| \leq t$, and

$$x \in \bigcup_{|\underline{j}|\leq t} C(\underline{j}) = \bigcup_{|\underline{j}|=t} S(\underline{j}). \qquad \square$$

LEMMA A.2. *For any Schubert variety $S(\underline{i}) \subset \text{Grass}(p, m+p)$, the set*

$$\left\{ G(s) = N(s)D^{-1}(s) \in \Sigma_{m,p}^{|\underline{i}|} | \det \begin{bmatrix} Z_1 & N(s) \\ Z_2 & D(s) \end{bmatrix} \not\equiv 0 \text{ for all } \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \in S(\underline{i}) \right\}$$

is a nonempty Zariski open set of $\Sigma_{m,p}^{|\underline{i}|}$.

Remark A.1. Brockett and Byrnes [2] proved this for the special Schubert variety, $\text{Grass}(p, m+p)$ itself. The proof given here is similar to Rosenthal's proof [13] of the Brockett–Byrnes result.

Proof. We only need to prove that such a set is nonempty. Take an element $a = (a_{\underline{j}}) \in C(\underline{i})$ such that

$$a_{\underline{j}} \neq 0 \quad \text{for all } \underline{j} \leq \underline{i}$$

and define a curve $\sigma(s) = (\sigma_{\underline{j}}(s)) \subset \text{Grass}(m, m+p)$ by

$$\sigma_{\underline{j}}(s) = a_{\underline{j}} s^{|\underline{i}|-|\underline{j}|},$$

where $\sigma_{\underline{j}}(s)$ is the $m \times m$ minor of an $(m+p) \times m$ matrix $P(s)$ formed by eliminating the j_1th through $\ldots, j_p$th rows. The σ has the following properties:

1. $\sigma_{\underline{j}} = 0$ if $|\underline{j}| > |\underline{i}|$;
2. $\sigma_{\underline{i}} = a_{\underline{i}} \neq 0$;
3. $\sigma_{(1,2,\ldots,p)} = a_{(1,2,\ldots,p)} s^{|\underline{i}|}$ is the only $s^{|\underline{i}|}$ term.

Therefore $\sigma(s)$ can be realized as a polynomial matrix $\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right]$ such that $N(s)D^{-1}(s) \in \Sigma_{m,p}^{|\underline{i}|}$. For any

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \in S(\underline{i}),$$

$Z \in C(\underline{j})$ for some $\underline{j} \leq \underline{i}$, and

$$\begin{aligned} \det \begin{bmatrix} Z_1 & N(s) \\ Z_2 & D(s) \end{bmatrix} &= \sum_{\underline{t}} (-1)^{|\underline{t}|} z_{\underline{t}} \sigma_{\underline{t}}(s) \\ &= (-1)^{|\underline{j}|} z_{\underline{j}} a_{\underline{j}} s^{|\underline{i}|-|\underline{j}|} + \text{higher-power terms} \\ &\not\equiv 0 \end{aligned}$$

by (22). $\square$

Let Π^α be embedded in $\text{Grass}(p, m+p) \subset \mathbf{P}^M$, as in (20).

LEMMA A.3. *Define the subspace* $L_t \subset \mathbf{P}^M$ *for* $|\underline{k}| \le t < |\underline{l}|$ *by*

$$L_t = \left\{ z = (z_{\underline{j}}) \in \mathbf{P}^M \middle| \sum_{|\underline{j}|=i} z_{\underline{j}} = 0,\ i = t+1, \dots, |\underline{l}| \right\},$$

where $\underline{k}$ *and* $\underline{l}$ *are defined by* (23). *Then*

$$L_t \cap \Pi^\alpha = \bigcup_{|\gamma_1|+\cdots+|\gamma_r|=t-|\underline{k}|} S(\gamma_1) \times \cdots \times S(\gamma_r),$$

where $\gamma_j = (i_1, \dots, i_{p_j})$, $1 \le i_1 < \cdots < i_{p_j} \le m_j + p_j$, *and the Schubert variety* $S(\gamma_j)$ *is defined by Definition* A.1 *for*

$$\begin{bmatrix} M_{j1} \\ M_{j2} \end{bmatrix} \in \mathrm{Grass}(p_j, m_j + p_j).$$

Proof. It holds that

$$\begin{aligned} L_t \cap \Pi^\alpha &= L_t \cap S(\underline{k}) \cap S^*(\underline{l}) \\ &= H_t \cap S(\underline{k}) \cap S^*(\underline{l}) \\ &= \bigcup_{|\underline{j}|=t,\ \underline{j} \le \underline{l}} S(\underline{j}) \cap S^*(\underline{k}) \\ &= \bigcup_{|\underline{j}|=t,\ \underline{k} \le \underline{j} \le \underline{l}} C(\underline{j}) \cap S^*(\underline{k}) \\ &= \bigcup_{|\gamma_1|+\cdots+|\gamma_r| \le t-|\underline{k}|} C(\gamma_1) \times \cdots \times C(\gamma_r) \\ &= \bigcup_{|\gamma_1|+\cdots+|\gamma_r|=t-|\underline{k}|} S(\gamma_1) \times \cdots \times S(\gamma_r). \qquad \square \end{aligned}$$

LEMMA A.4. *For any* $S(\gamma_1) \times \cdots \times S(\gamma_r) \subset \Pi^\alpha$, *let* $n = |\gamma_1| + \cdots + |\gamma_r|$. *Then*

$$\{\sigma \in \Sigma^n_{m,p} | E_\sigma \cap S(\gamma_1) \times \cdots \times S(\gamma_r) = \emptyset\}$$

is a nonempty Zariski open set.

Proof. We only need to prove that it is nonempty. Take $N_j(s)D_j^{-1}(s) \in \Sigma^{|\gamma_j|}_{m_j,p_j}$ such that

$$\det \begin{bmatrix} M_{j1} & N_j(s) \\ M_{j2} & D_j(s) \end{bmatrix} \not\equiv 0$$

for all

$$\begin{bmatrix} M_{j1} \\ M_{j2} \end{bmatrix} \in S(\gamma_j)$$

and define $\sigma = N(s)D^{-1}(s) \in \Sigma^n_{m,p}$ with

$$N(s) = \text{block diag.}(N_1(s), \dots, N_r(s)),$$

$$D(s) = \text{block diag.}(D_1(s), \dots, D_r(s)).$$

Then

$$E_\sigma \cap S(\gamma_1) \times \cdots \times S(\gamma_r) = \emptyset. \qquad \square$$

PROPOSITION A.1. *Let $n \le \sum_{i=1}^r m_i p_i$. The set of systems in $\Sigma_{m,p}^n$ such that*

$$\dim(E_\sigma \cap \Pi_C^\alpha) = \sum_{i=1}^4 m_i p_i - n - 1$$

is a nonempty Zariski open set where dimension -1 means that the intersection is empty.

Proof. We only need to prove that the set is nonempty. When $\sum_{i=1}^r m_i p_i = n$, this follows from Lemma A.4. When $\sum_{i=1}^r m_i p_i > n$, take any σ in

$$\bigcap_{|\gamma_1|+\cdots+|\gamma_r|=n} \{\sigma \in \Sigma_{m,p}^n | E_\sigma \cap S(\gamma_1) \times \cdots \times S(\gamma_r) = \emptyset\}.$$

Then

$$\begin{aligned} E_\sigma \cap \Pi_C^\alpha \cap L(n + |\underline{k}|) &= E_\sigma \cap \bigcup_{|\gamma_1|+\cdots+|\gamma_r|=n} S(\gamma_1) \times \cdots \times S(\gamma_r) \\ &= \bigcup_{|\gamma_1|+\cdots+|\gamma_r|=n} E_\sigma \cap S(\gamma_1) \times \cdots \times S(\gamma_r) \\ &= \emptyset. \end{aligned}$$

Since

$$\text{codim } L(n + |\underline{k}|) = |\underline{l}| - |\underline{k}| - n = \sum_{i=1}^r m_i p_i - n,$$

we see that

$$\dim E_\sigma \cap \Pi_C^\alpha \le \sum_{i=1}^r m_i p_i - n - 1$$

by the projective dimension theorem [9]. On the other hand,

$$\dim E_\sigma \cap \Pi_C^\alpha \ge \sum_{i=1}^r m_i p_i - n - 1,$$

because $\dim \Pi_C^\alpha = \sum_{i=1}^r m_i p_i$ and $\text{codim } E_\sigma \le n + 1$. $\square$

PROPOSITION A.2. *Let*

$$X_R = \{(\sigma, z) \in \Sigma_{m,p}^n \times \Pi_{\mathbf{R}}^\alpha \mid z \in E_\sigma \cap \Pi_R^\alpha\}$$

and

$$X_1 = \left\{ (\sigma, z) \in X_R \mid \dim E_\sigma \cap T_z = \sum_{i=1}^r m_i p_i - n - 1 \right\}.$$

Then X_1 is a nonempty Zariski open subset of the X_R if $\sum_{i=1}^{r} m_i p_i > n$, $m \leq n$, and $p \leq n$.

Proof. We only need to prove that X_1 is nonempty. We switch rows of the determinant in (10) and write

$$\phi_\sigma(M_1, \ldots, M_r) = \det \left[\begin{array}{ccc|c} M_{12} & & & \\ & \ddots & & D(s) \\ & & M_{r2} & \\ \hline M_{11} & & & \\ & \ddots & & N(s) \\ & & M_{r1} & \end{array}\right],$$

where M_{i1} and M_{i2} are the same as in (20). For any

$$(\sigma, z) \in X_R, \qquad z = (M_1, \ldots, M_r),$$

$$\phi_\sigma(z) = \phi_\sigma(M_1, \ldots, M_r) \equiv 0,$$

and there exists a $Q_i \in GL(m_i + p_i, \mathbf{R})$ such that

$$Q_i \begin{bmatrix} M_{i2} \\ M_{i1} \end{bmatrix} = \begin{bmatrix} 0 \\ I_i \end{bmatrix}$$

for each

$$\begin{bmatrix} M_{i2} \\ M_{i1} \end{bmatrix},$$

where I_i is the $p_i \times p_i$ identity matrix. Let

$$L = Q_1 \times Q_2 \times \cdots \times Q_r.$$

Then $L(z) = z_0$, where

$$z_0 = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

If we write

$$L(\sigma) = L\left(\begin{bmatrix} D(s) \\ N(s) \end{bmatrix}\right) = \begin{bmatrix} P_1(s) \\ P_2(s) \end{bmatrix},$$

then $\det P_1(s) \equiv 0$, and $(\sigma, z) \in X_1$ if and only if

$$\dim E_{L(\sigma)} \cap T_{z_0} = \sum_{i=1}^{r} m_i p_i - n - 1. \tag{24}$$

Let $a_{ij}(s), 1 \leq i \leq m, j > m$ be the $m \times m$ minor of

$$\begin{bmatrix} P_1(s) \\ P_2(s) \end{bmatrix}$$

formed by the jth row and the first m rows, except the ith row. It is easy to check that (24) is true if and only if

$$\dim sp\{a_{ij}(s)\} = n+1, \tag{25}$$

where $i = 1, 2, \ldots, m$ and

$$m+p_1+\cdots+p_k < j \le m+p_1+\cdots+p_{k+1} \text{ if } m_1+\cdots+m_k < i \le m_1+\cdots+m_{k+1}.$$

Hence we need to construct an $(m+p)\times m$ polynomial matrix such that

(i) The minor formed by the first m rows is zero,

(ii) The matrix satisfies (25),

(iii) It can be changed into $\left[\begin{smallmatrix} D(s) \\ N(s) \end{smallmatrix}\right]$ by a row operation of the form $Q_1 \times Q_2 \times \cdots \times Q_r$.

By the assumptions, p_i and $m_i, i = 1, \ldots, r$ cannot all be 1. Without loss of generality, assume that $\max_i p_i > 1$. Let

$$n = p_1m_1 + \cdots + p_km_k + ap_{k+1} + b,$$

where $k < r$,

$$p_1 \le p_2 \le \cdots \le p_{k+1}, \qquad p_{k+1} > 1,$$

$$0 \le a < m_{k+1},$$

and

$$0 \le b < p_{k+1}.$$

Define

$$P(s) = \left[\begin{array}{ccccccccc}
A_1 & & & & & & & & \\
B_1 & A_2 & & & & & & & \\
 & B_2 & \ddots & & & & & & \\
 & & \ddots & A_k & & & & & \\
 & & & B_k & A_{k+1} & & & & \\
 & & & & 0 & I & & & \\
 & & & & & \ddots & \ddots & & \\
 & & & & & & 0 & I & \\
\hline
0 & \cdots & \cdots & \cdots & C_1 & \cdots & \cdots & 0 & \\
\vdots & & & & \vdots & & & \vdots & \\
\vdots & & & & C_{k+1} & \cdots & \cdots & 0 & \\
\vdots & & & & 0 & \cdots & \cdots & 0 & \\
\vdots & & & & \vdots & & & \vdots & \\
0 & \cdots & \cdots & \cdots & 0 & \cdots & \cdots & 0 &
\end{array}\right],$$

where A_i, B_i, and C_i are $m_i \times m_i, m_{i+1} \times m_i$, and $p_i \times m_i$ matrices defined by the following case.

Case 1. $a > 0$ and $b > 0$. It holds that

$$A_i = \begin{bmatrix} s^{p_i} & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & s^{p_i} \end{bmatrix}, \qquad B_i = \begin{bmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & 0 \end{bmatrix},$$

and

$$C_i = \begin{bmatrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & s \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & s^{p_i-1} \end{bmatrix}$$

for $i = 1, 2, \dots, k$,

$$A_{k+1} = \left[\begin{array}{cccc|ccc} \overbrace{\qquad\qquad\qquad}^{a} & & & & & & \\ s^{p_{k+1}} & & & & & & \\ 1 & \ddots & & & & & \\ & \ddots & s^{p_{k+1}} & & & & \\ & & 1 & s^b & & & \\ & & & 1 & 0 & & \\ & & & & \ddots & \ddots & \\ & & & & & 1 & 0 \end{array}\right]$$

and

$$C_{k+1} = \begin{bmatrix} 0 & \cdots & 1+s \\ 0 & \cdots & s+2s^2 \\ \vdots & & \vdots \\ 0 & \cdots & s^{p_{k+1}-1} + p_{k+1}s^{p_{k+1}} \end{bmatrix}.$$

Case 2. $b = 0$ and $a > 1$. Everything is the same as in Case 1, except that

$$A_{k+1} = \left[\begin{array}{ccccc|ccc} \overbrace{\qquad\qquad\qquad\qquad}^{a} & & & & & & & \\ s^{p_{k+1}} & & & & & & & \\ 1 & \ddots & & & & & & \\ & \ddots & s^{p_{k+1}} & & & & & \\ & & 1 & s^{p_{k+1}-1} & & & & \\ & & & 1 & s & & & \\ & & & & 1 & 0 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & 1 & 0 \end{array}\right].$$

Case 3. $b = 0$, $a = 1$, and $k \neq 0$. Everything is the same as in Case 1, except that

$$A_k = \begin{bmatrix} s^{p_k} & & & \\ 1 & \ddots & & \\ & \ddots & s^{p_k} & \\ & & 1 & s^{p_k-1} \end{bmatrix}, \qquad A_{k+1} = \begin{bmatrix} s & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}.$$

Case 4. $b = 0$, $a = 1$, and $k = 0$. Everything is the same as in Case 3, except that

$$C_1 = C_{k+1} = \begin{bmatrix} 0 & \dots & 1 \\ 0 & \dots & s \\ \vdots & & \vdots \\ 0 & \dots & s^{p_1-1} \end{bmatrix}.$$

Case 5. $a = 0$ and $b \geq p_k$. Everything is the same as in Case 1, except that

$$A_{k+1} = \begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}, \qquad C_{k+1} = \begin{bmatrix} 0 & \dots & 1 \\ 0 & \dots & s \\ \vdots & & \vdots \\ 0 & \dots & s^b \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

Case 6. $a = 0$ and $0 < b < p_k$. Everything is the same as in Case 5, except that

$$A_k = \begin{bmatrix} s^{p_k} & & & \\ 1 & \ddots & & \\ & \ddots & s^{p_k} & \\ & & 1 & s^b \end{bmatrix}, \qquad C_{k+1} = \begin{bmatrix} 0 & \dots & s^{p_k-b} \\ 0 & \dots & s^{p_k-b+1} \\ \vdots & & \vdots \\ 0 & \dots & s^{k_k} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

Case 7. $a = 0$, $b = 0$, $m_k > 1$, and $p_k > 1$. Everything is the same as in Case 5, except that

$$A_k = \begin{bmatrix} s^{p_k} & & & & \\ 1 & \ddots & & & \\ & \ddots & s^{p_k} & & \\ & & 1 & s^{p_k-1} & \\ & & & 1 & s \end{bmatrix}, \qquad C_{k+1} = \begin{bmatrix} 0 & \dots & s \\ 0 & \dots & s^2 \\ \vdots & & \vdots \\ 0 & \dots & s^{p_k} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

Case 8. $a = 0$, $b = 0$, $p_k > 1$, $m_k = 1$, and $k > 1$. Everything is the same as in Case 5, except that $C_{k+1} = 0$,

$$A_{k-1} = \begin{bmatrix} s^{p_{k-1}} & & & \\ 1 & \ddots & & \\ & \ddots & s^{p_{k-1}} & \\ & & 1 & s^{p_{k-1}-1} \end{bmatrix}, \qquad A_k = \begin{bmatrix} s & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix},$$

and

$$C_k = \begin{bmatrix} 0 & \dots & s \\ 0 & \dots & s^2 \\ \vdots & & \vdots \\ 0 & \dots & s^{p_k} \end{bmatrix}.$$

Case 9. $a = 0$, $b = 0$, $p_k > 1$, $m_k = 1$, and $k = 1$,

$$B_1 = \begin{bmatrix} 0 & \dots & 1 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} s & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix},$$

$$C_1 = \begin{bmatrix} 0 & \dots & 1 \\ 0 & \dots & s \\ \vdots & & \vdots \\ 0 & \dots & s^{p_1-2} \\ 0 & \dots & 0 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} 0 & \dots & 1 \\ 0 & \dots & s \\ \vdots & & \vdots \\ 0 & \dots & s^{p_1-1} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

Case 10. $a = 0$, $b = 0$, $p_k = 1$, and $m_1 + m_2 + \cdots + m_k > 1$. Everything is the same as in Case 5, except that

$$A_1 = \begin{bmatrix} 1 & & & \\ 1 & s & & \\ & \ddots & \ddots & \\ & & 1 & s \end{bmatrix}, \qquad C_{k+1} = \begin{bmatrix} 0 & \dots & 1 \\ 0 & \dots & s \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}.$$

Case 11. $a = 0$, $b = 0$, $k = 1$, and $p_1 = m_1 = 1$,

$$A_1 = 0, \quad A_2 = \begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}, \qquad C_1 = 0, \quad C_2 = \begin{bmatrix} 0 & \dots & 1 \\ 0 & \dots & s \\ \vdots & & \vdots \\ 0 & & 0 \end{bmatrix}.$$

It is not difficult to check that $P(s)$ satisfies the requirements. □

Acknowledgments. The author thanks Professor K. A. Grasse and the referees for their help in identifying some errors and potentially confusing points in the original manuscript, and also Professor J. Rosenthal for his suggestions.

REFERENCES

[1] I. BERSTEIN, *On the Lusternick–Schnirelmann category of real Grassmannians*, Proc. Cambridge Philosophy Soc., 79 (1976), pp. 129–239.

[2] R. W. BROCKETT AND C. I. BYRNES, *Multivariable nyquist criteria, root Loci and pole placement: a geometric viewpoint*, IEEE Trans. Automat. Control, 26 (1981), pp. 271–284.

[3] C. I. BYRNES, *Algebraic and geometric aspects of the analysis of feedback systems*, in Geometric Methods in Linear System Theory, C. I. Byrnes and C. F. Martin, eds., D. Reidel, Dordrecht, the Netherlands, 1980, pp. 85–124.

[4] ———, *Pole assignment by output feedback*, in Lecture Notes in Control and Information Sciences, No. 135, Springer-Verlag, New York, 1989, pp. 31–78.

[5] ———, *Stabilizability of multivariable systems and the Ljusternik–Šnirel'mann category of real Grassmannians*, Systems Control Lett., 3 (1983), pp. 255–262.

[6] J. P. CORFMAT AND A. S. MORSE, *Decentralized control of linear multivariable systems*, Automatica, 12 (1976), pp. 479–496.

[7] G. D. FORNEY, *Minimal bases of rational vector spaces with applications to multivariable linear systems*, SIAM J. Control Optim., 13 (1975) pp. 493–520.

[8] P. GRIFFITHS AND J. ADAMS, *Topics in Algebraic and Analytic Geometry*, Princeton University Press, Princeton, NJ, 1974.

[9] R. HARTSHORNE, *Algebraic Geometry*, Springer-Verlag, New York, 1977.

[10] W. V. D. HODGE AND D. PEDOE, *Methods of Algebraic Geometry*, Vol. II, Cambrige University Press, Cambridge, UK, 1952.

[11] C. F. MARTIN AND R. HERMANN, *Application of algebraic geometry to system theory: The McMillan degree and Kronecker indices as topological and holomorphic invariants*, SIAM J. Control Optim., 16 (1978), pp. 743–755.

[12] D. MUMFOR, *Algebraic Geometry* I: *Complex Projective Varieties*, Springer-Verlag, New York, 1976.

[13] J. ROSENTHAL, *Geometric Methods for Feedback Stabilization of Multivariable Linear Systems*, Ph.D. thesis, Dept. of Math., Arizona State University, Tempe, AZ, 1990.

[14] ———, *On dynamic feedback compensation and compactification of systems*, SIAM J. Control Optim., 32 (1994), pp. 279–296.

[15] I. R. SHAFAREVICH, *Basic Algebraic Geometry*, Springer-Verlag, Berlin, New York, 1974.

[16] S. H. WANG AND E. J. DAVISON, *On the stabilization of decentralized control systems*, IEEE Trans. Automat. Control, 18 (1973) pp. 473–478.

[17] X. WANG, *On output feedback via Grassmannians*, SIAM J. Control Optim., 29 (1991), pp. 926–935.

[18] ———, *On compactifications of decentralized output feedback spaces*, in Computation and Control II, K. Bowers and J. Lund, eds., Birkhäuser, Boston, 1991, pp. 351–358.

[19] ———, *Pole placement by static output feedback*, J. Math. Sys. Estim. Control, 2 (1992), pp. 205–218.

[20] X. WANG, C. F. MARTIN, D. GILLIAM, AND C. I. BYRNES, *On decentralized feedback pole placement of linear systems*, Internat. J. Control., 55 (1992), pp. 511–518

[21] J. C. WILLEMS AND W. H. HESSELINK, *Generic properties of the pole placement problem*, in Proc. 1978 IFAC, Helsinki, Finland, pp. 1725–1729.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 876–889, May 1994

013

THE OUTPUT-NULLING SPACE, PROJECTED DYNAMICS, AND SYSTEM DECOMPOSITION FOR LINEAR TIME-VARYING SINGULAR SYSTEMS*

WILLIAM J. TERRELL†

Abstract. A decomposition of a given linear time-varying singular control system is developed by defining the output-nulling space with respect to a given output structure. Relevant subspaces and the dynamics on them are described by computable projection operators obtained from information in the system's derivative array. The relevant projectors are generated as solutions of a homogeneous linear matrix-differential equation. An algorithm for obtaining the system decomposition is outlined in a pointwise manner.

Key words. output-nulling/unobservable subspace, system decomposition, linear time-varying system, implicit, singular, descriptor, differential-algebraic

AMS subject classifications. 34A09, 34A30, 93B07, 93B10, 93B11, 34A46

1. Introduction.

Research in the last decade has established the wide applicability of implicit differential systems of the form

$$E(t)x' + F(t)x = B(t)u, \tag{1a}$$

where E, F are square matrices, E is identically singular on the interval $\mathcal{I}$, $x \in R^n$, and u is a smooth real-valued input function [17], [2], [1], [18]. Equation (1a) is often called a differential-algebraic equation (DAE). The fundamental existence and uniqueness theory for such linear time-varying equations is now fairly complete and is developed by Campbell [3], [5]. More recently, work has appeared on linear time-varying systems that include an output structure

$$y = C(t)x, \tag{1b}$$

where $C(t)$ is a smooth $l \times n$ matrix function. Observability and controllability concepts for system (1) are developed in [10], [8], respectively, and the fundamental duality statement relating these concepts is established in [8]. This recent work on linear time-varying systems is an important step in the development of a general theory of control for fully nonlinear implicit differential systems involving inputs and outputs. For the control theory of the linear time-invariant version of system (1) (see, for example, [16], [11], [12], [15], [22]). The earliest work on observability and controllability for time-varying nonsingular systems appears in [14], [23], [20].

In this paper we extend the work in [10] by developing a computable decomposition of (1) into unobservable subspace $\oplus$ observable complement. Section 2 surveys the solvability conditions (Theorem 2.1) and observability conditions (Theorem 2.2) used in [10]. The unobservable (or output-nulling) space is described in §3. In §4 we define observable complements and describe the dynamics on the unobservable subspace and an observable complement by means of smooth projection operators that are computable from information in the system's derivative array (§2). Section 5 provides a simple example of our results and illustrates an algorithm by which the system decomposition may be obtained in a pointwise manner.

* Received by the editors February 3, 1992; accepted for publication (in revised form) December 15, 1992.

† Department of Mathematical Sciences, Virginia Commonwealth University, Richmond, Virginia 23284.

2. Solvability and observability. We assume that (1a) is solvable on the closed and bounded interval $\mathcal{I}$. Solvability means that solutions exist on $\mathcal{I}$ for every sufficiently differentiable input u, and solutions depend uniquely on their value at any t_0 in $\mathcal{I}$ [5]. When the coefficient matrices E, F are constant, solvability corresponds to regularity of the matrix pencil, $E + \lambda F$.

To simplify notation, set $b(t) = B(t)u(t)$. Differentiating (1a) j times and (1b) k times gives the linear system of equations

$$\begin{bmatrix} \mathcal{F}_j & \mathcal{E}_j \end{bmatrix} \begin{bmatrix} x \\ \mathbf{x}_j \end{bmatrix} = \mathbf{b}_j, \tag{2}$$

$$\mathcal{C}_k \begin{bmatrix} x \\ \mathbf{x}_{k-1} \end{bmatrix} = \mathbf{y}_k, \tag{3}$$

where

$$\mathcal{F}_j = \begin{bmatrix} F \\ F' \\ \vdots \\ F^{(j)} \end{bmatrix}, \mathbf{y}_k = \begin{bmatrix} y \\ y' \\ \vdots \\ y^{(k)} \end{bmatrix}, \mathbf{b}_j = \begin{bmatrix} b \\ b' \\ \vdots \\ b^{(j)} \end{bmatrix}, \mathbf{x}_j = \begin{bmatrix} x' \\ x'' \\ \vdots \\ x^{(j+1)} \end{bmatrix}$$

and

$$\mathcal{E}_j = \begin{bmatrix} E & 0 & \cdot & \cdot & 0 \\ E' + F & E & 0 & \cdot & \cdot \\ E'' + 2F' & 2E' + F & E & \ddots & \cdot \\ \vdots & * & * & \ddots & 0 \\ E^{(j)} + jF^{(j-1)} & * & * & * & E \end{bmatrix}.$$

The matrix $\mathcal{C}_k$ is similarly generated by differentiation of the output equation. If (3) is written as $\tilde{C}_k x + \hat{C}_k \mathbf{x}_{k-1} = \mathbf{y}_k$ and $j + 1 \geq k$, the combined array is

$$\mathcal{O}_{j,k} = \left[\begin{array}{c|c} \mathcal{F}_j & \mathcal{E}_j \\ \tilde{C}_k & \begin{bmatrix} \hat{C}_k & 0_{(k+1)m \times (j+1-k)n} \end{bmatrix} \end{array}\right]. \tag{4}$$

The fundamental solvability conditions for (1a) and the results on observability for system (1) obtained in [10] are expressed in terms of the following definition.

DEFINITION 1. *The system of algebraic equations $Ax = b$, written as*

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

is 1-full with respect to x_1, if x_1 is uniquely determined by any consistent vector b.

THEOREM 2.1 ([5]). *Suppose that* (1a) *is solvable on the interval $\mathcal{I}$ and that E, F are $2n$-times continuously differentiable. Then*

$$\mathcal{E}_j \text{ has constant rank on } \mathcal{I} \text{ for } j = n, \tag{5}$$

$$\mathcal{E}_j \text{ is 1-full with respect to } x' \text{ on } \mathcal{I} \text{ for } j = n, \tag{6}$$

$$[\mathcal{F}_j \ \mathcal{E}_j] \text{ has full row rank on } \mathcal{I} \text{ for } 1 \leq j \leq n. \tag{7}$$

Equation (1a) has *index* ν if ν is the smallest j for which conditions (5), (6), and (7) hold. Suppose then that (1a) is solvable and index ν. Since $\mathcal{E}_\nu$ has constant rank and is 1-full with respect to x' on $\mathcal{I}$, there exists a smooth nonsingular $R(t)$ such that [3]

$$R(t)\,\mathcal{E}_\nu(t) = \begin{bmatrix} I_{n\times n} & 0 \\ 0 & H(t) \\ 0_{\rho\times n} & 0 \end{bmatrix}.$$

It follows that the smooth row reduced form of $[\,\mathcal{E}_\nu\ \mathcal{F}_\nu\ \mathbf{b}_\nu\,]$ is

$$\left[\begin{array}{cc|c|c} I_{n\times n} & 0 & Q_1 & \bar{b}_1 \\ 0 & H & Q_2 & \bar{b}_2 \\ 0_{\rho\times n} & 0 & M & \bar{b}_3 \end{array}\right], \tag{8}$$

where H and M have full row rank by Theorem 2.1. From [4] and the rank properties of H and M just mentioned, the equation

$$M(t)x = \bar{b}_3 \tag{9}$$

determines the manifold of consistent initial conditions of (1a) at time t. For the unforced system (1a) with $u = 0$, we write $\mathcal{G}(t)$ for the solution space at time t.

System (1) is *weakly observable* if $u = 0$ and $y = 0$ imply $x = 0$. An important classical type of observability is total observability.

DEFINITION 2 ([10]). *System* (1) *is* totally observable *on $\mathcal{I}$ if knowledge of the output y and input u on any subinterval $\tilde{\mathcal{I}}$ of $\mathcal{I}$ uniquely determines* smooth *solutions x of* (1a) *on $\tilde{\mathcal{I}}$.*

A stronger type of observability is smooth observability.

DEFINITION 3 ([10]). *System* (1) *is* smoothly observable (*of order (k, j)*) *on $\mathcal{I}$ if there exists smooth $K_i(t)$, $L_i(t)$ on $\mathcal{I}$ such that*

$$x = \sum_{i=0}^{k} K_i(t)y^i(t) + \sum_{i=0}^{j} L_i(t)(Bu)^i(t).$$

If $C(t)$ in (1b) is not full column rank on a dense set, then the additional information required to determine x is obtained by differentiating (1b) and (1a).

THEOREM 2.2 ([10]). *System* (1) *is totally observable on $\mathcal{I}$ if and only if there exist j, k with $k \le j + 1$, such that the matrix $\mathcal{O}_{j,k}$ is 1-full with respect to x on a dense set in $\mathcal{I}$. If $\mathcal{O}_{j,k}$ is 1-full on a dense set in $\mathcal{I}$ and has constant rank, then* (1) *is smoothly observable of order (k, j).*

See [8] for a necessary and sufficient condition for smooth observability expressed in terms of matrix ranks. An analysis of observability for linear time-invariant singular systems is given in [22]. Here we merely note that if we are working with the linear time-invariant version of system (1), and we rewrite the system using the canonical form of [9] (or [22]), then the 1-full condition of Theorem 2.2 is equivalent to the matrix rank condition for observability given in [22, Cor. 2].

3. The output-nulling space. We work with the control system (1) with (1a) solvable on the interval $\mathcal{I}$.

DEFINITION 4. *The* output-nulling space $\mathcal{N}_{\mathcal{I}}(t)$ *for system* (1) *is the largest time-varying space invariant under the homogeneous equation* $E(t)x' + F(t)x = 0$ *and contained pointwise in* $\ker C(t)$.

The following lemma is easily established from Definition 4.

LEMMA 3.1. *For each* t_0, $\mathcal{N}(t_0)$ *consists of those consistent conditions* $x(t_0) \in \mathcal{G}(t_0)$ *whose corresponding solution to the homogeneous equation produces zero output on all of* $\mathcal{I}$.

Since our interval $\mathcal{I}$ is fixed, we drop the subscript $\mathcal{I}$ on $\mathcal{N}$. We always have $\{0\} \subset \mathcal{N}(t)$. Solvability implies that $\mathcal{N}(t_0) = \{0\}$ for *some* t_0 if and only if $\mathcal{N}(t) = \{0\}$ for all t in $\mathcal{I}$, because a solution of the homogeneous equation is either identically zero or never zero on $\mathcal{I}$. If t_0, t_1 are in $\mathcal{I}$, it follows that linearly independent vectors in $\mathcal{N}(t_0)$ are taken to linearly independent vectors in $\mathcal{N}(t_1)$ under the flow of the differential equation. Thus, we arrive at the following lemma.

LEMMA 3.2. $\mathcal{N}(t)$ *has constant dimension on* $\mathcal{I}$.

Total observability or smooth observability as defined in [10] implies $\mathcal{N} = \{0\}$, but simple examples in [21] show that the converse does not hold.

We now show that for systems with analytic coefficients, the output-nulling space $\mathcal{N}$ can be characterized using information in the derivative array $\mathcal{O}_{j,k}$. We need the following lemma.

LEMMA 3.3. *Suppose that* $H(t)$ *is an* $m \times n$ *real analytic matrix function defined on an open interval containing the closed bounded interval* $\mathcal{I}$. *Let*

$$H[k] = \begin{bmatrix} H \\ H' \\ \vdots \\ H^{(k)} \end{bmatrix}$$

and suppose $q = \max\{\operatorname{rank} H[k](t) : k \geq 0, t \in \mathcal{I}\}$. *Then there exists* k^* *such that* $H[k^*]$ *has rank* q *everywhere on* $\mathcal{I}$. *Consequently,*

$$\bigcap_{k \geq 0} \ker H[k] = \ker H[k^*]$$

has constant dimension $n-q$ *on* $\mathcal{I}$. *Moreover,* v *is a constant vector such that* $H(t)v = 0$ *on* $\mathcal{I}$ *if and only if* $v \in \ker H[k^*](t))$ *for all* t *in* $\mathcal{I}$, *or equivalently,* $v \in \ker(H[k^*](t))$ *for some* t *in* $\mathcal{I}$.

Proof. For every t and k we have $\ker H[k+1](t) \subset \ker H[k](t)$, so that the space

$$Z(t) = \bigcap_{k \geq 0} \ker H[k](t)$$

is well defined. Let $\bar{k}$ and t_0 be such that $\operatorname{rank} H[\bar{k}](t_0) = q$. By real analyticity, $H[\bar{k}]$ has rank q at all but a finite number of points $t_1, \ldots, t_r$ in $\mathcal{I}$. Let $v_1, \ldots, v_{n-q}$ be independent vectors in $\ker H[\bar{k}](t_0)$. By choice of q we have $\ker H[k](t_0) = \ker H[\bar{k}](t_0)$ for all $k \geq \bar{k}$. Then $\frac{d^k}{dt^k} H(t)v_i|_{t_0} = 0$ for all $k \geq 0$, and real analyticity implies $H(t)v_i = 0$ for all t in $\mathcal{I}$. But then $v_1, \ldots, v_{n-q}$ are all in $Z(t)$ for every t. Thus, $Z(t)$ has dimension at least $n-q$ for all t. However, by choice of q we cannot have more than $n-q$ independent vectors in any $Z(t)$; hence, $Z(t)$ has constant dimension $n-q$.

Now, by finite dimensionality there exist integers $k_1, \ldots, k_r \geq \bar{k}$ such that

$$\bigcap_{k \geq 0} \ker H[k](t_p) = \ker H[k_p](t_p)$$

for $p = 1, \ldots, r$. Let $k^* = \max\{\bar{k}, k_1, \ldots, k_r\}$. Then we have

$$Z(t) = \ker H[k^*](t)$$

for every t in $\mathcal{I}$.

Finally, note that since H is analytic, the vector v satisfies $H(t)v = 0$ on $\mathcal{I}$ if and only if $v \in Z(t)$ for every t. □

THEOREM 3.4. *Let ν be the index of the* DAE (1a). *Suppose the coefficients E, F, B, C of system* (1) *are real analytic. Then there exists a positive integer k^* such that the following statement is true. For any t_0 in $\mathcal{I}$, $x_0 \in \mathcal{N}(t_0)$ if and only if the equation*

$$\mathcal{O}_{j,k}(t_0) \begin{bmatrix} x_0 \\ \mathbf{x}_j \end{bmatrix} = 0 \tag{10}$$

is consistent for one (and hence any) pair j, k satisfying $k \geq k^*$ and $j \geq \nu + k - 1$.

Proof. The consistency condition (10) is clearly a necessary condition for x_0 to be in $\mathcal{N}(t_0)$. The way in which the derivative array $\mathcal{O}_{j,k}$ transforms under coordinate change [3] allows us to prove sufficiency for the decoupled canonical form for analytic systems [9]

$$\begin{aligned} z_1' &= B_1(t)u, \\ N(t)z_2' + z_2 &= B_2(t)u, \\ y &= C_1(t)z_1 + C_2(t)z_2. \end{aligned}$$

Note that the top $n \times n$ block of the $\mathcal{F}_j$ matrix is

$$\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix},$$

the lower blocks being zero. Using the notation of Lemma 3.3, the consistency condition expressed with the new derivative array has the form

$$\tilde{\mathcal{O}}_{j,k} \begin{bmatrix} z_0 \\ \mathbf{z}_j \end{bmatrix} = \begin{bmatrix} \mathcal{F}_j & \mathcal{E}_j \\ [C_1[k](t)\ C_2[k](t)] & [*\ 0] \end{bmatrix} \begin{bmatrix} z_{10} \\ z_{20} \\ \mathbf{z}_j \end{bmatrix} = 0 \tag{11}$$

for $j \geq \nu + k - 1$. Note that the starred entry in (11) has kn columns. By Lemma 3.3 there exists k^* such that $C_1[k](t)$ has constant rank on $\mathcal{I}$ for $k \geq k^*$. Let $k \geq k^*$ and let $j \geq \nu + k - 1$. Since the array $[\mathcal{F}_j\ \mathcal{E}_j]$ determines the solution manifold of the DAE, consistency of (11) implies that z_{20} must be zero, since $N(t)z_2' + z_2 = g$ is totally singular [9]. Thus we have

$$\mathcal{F}_j \begin{bmatrix} z_{10} \\ z_{20} \end{bmatrix} = 0.$$

Equation (11) now implies that $\mathbf{z}_{k-1} = 0$, because $\mathcal{E}_j$ uniquely determines $z', \ldots, z^{(k)}$ [10]. Therefore, consistency of (11) at t_0 and the form of * in (11) implies

$$C_1[k](t_0)z_{10} = 0. \tag{12}$$

The initial condition z_{10} is a constant solution of the canonical homogeneous equation $z_1' = 0$. Thus, from Lemma 3.3, if z_0 is consistent, then $C_1^{(k)}(t_0)z_{10} = 0$ for all k, and by analyticity $C_1(t)z_{10} = 0$ for all t in $\mathcal{I}$. So $z_0 = [z_{10}^T, 0^T]^T \in \mathcal{N}(t_0)$. □

Remark. Careful consideration of $\tilde{\mathcal{O}}_{j,k}$ in (11) shows that the nullspace of $\tilde{\mathcal{O}}_{j,k}$ has constant dimension. Therefore, $\mathcal{O}_{j,k}$ has constant rank on $\mathcal{I}$ and we have the next corollary.

COROLLARY 3.5. *For systems with analytic coefficients, $\mathcal{N} = \{0\}$ on $\mathcal{I}$ implies smooth observability on $\mathcal{I}$. Thus, $\mathcal{N} = \{0\}$ on $\mathcal{I}$ is equivalent to smooth observability on $\mathcal{I}$ for these systems.*

Proof. From Theorem 3.4 we see that $\mathcal{N} = \{0\}$ implies $\mathcal{O}_{j,k}$ is 1-full on $\mathcal{I}$ for a pair j, k, and by the Remark above, $\mathcal{O}_{j,k}$ has constant rank on $\mathcal{I}$. By Theorem 2.2, the system is smoothly observable on $\mathcal{I}$. □

While Theorem 3.4 assumes analyticity, the subspace $\mathcal{N}$ is characterized by the derivative array for many nonanalytic systems as well. However, it appears difficult to make a general statement that improves on Theorem 3.4. The following example shows that the derivative array may not characterize the output-nulling space for nonsingular systems with C^∞ coefficients.

Example 1. Consider the following nonsingular system which has a zero $F(t)$ coefficient:

$$x_1' = u_1, \qquad x_2' = u_2$$

with output

$$y_1 = \phi(t)x_1, \qquad y_2 = \phi(-t)x_2,$$

where ϕ is defined on an interval around zero and $\phi^{(i)}(t) \neq 0$ for $t < 0$ and all $i \geq 0$, while $\phi(t) = 0$ for $t \geq 0$. Using the notation of Lemma 3.3, the output matrix $C(t)$ has the property that $\operatorname{rank} C[k](t)$ is one for all t except $t = 0$, where $C[k](0) \equiv 0$. But for nonzero v, if v is in $\ker C[k](t_1)$ for $t_1 > 0$, then v cannot be in $\ker C[k](t_2)$ for $t_2 < 0$. □

By considering further the rank and nullity structure of $\tilde{\mathcal{O}}_{j,k}$ in (11), it is possible to state a condition on $\mathcal{O}_{j,k}$ that is necessary if the array is to characterize $\mathcal{N}$ by the consistency of (10).

DEFINITION 5. *The* nullity *of a sequence of matrices $H_k(t)$ defined on $\mathcal{I}$* stabilizes on $\mathcal{I}$ at k^* *if* $\dim \ker H_k(t) = \dim \ker H_{k^*}(t)$ *for all $k \geq k^*$ and* $\dim \ker H_{k^*}(t)$ *is constant on $\mathcal{I}$.*

If we maintain the relation $j = \nu + k - 1$, and define $\mathcal{O}_k \equiv \mathcal{O}_{j,k}$, then it is not difficult to prove the following corollary.

COROLLARY 3.6 ([21]). *If $\mathcal{N}(t_0)$ is characterized for every t_0 in $\mathcal{I}$ by the consistency equation*

$$\mathcal{O}_{k^*}(t_0) \begin{bmatrix} x_0 \\ \mathbf{x}_{\nu+k^*} \end{bmatrix} = 0 \tag{13}$$

for some k^*, then the nullity of $\mathcal{O}_k$ necessarily stabilizes on $\mathcal{I}$ at k^*.

4. System decomposition and dynamics. One difficulty in obtaining decompositions of a singular system with respect to observability is that the solution manifold $\mathcal{M}_b(t)$ of $E(t)x' + F(t)x = b$ depends on b as well as t. However, the space $\mathcal{M}_b(t)$ is a translate of $\mathcal{G}(t)$, the solution manifold at t for the unforced equation $E(t)x' + F(t)x = 0$. From the analysis of observability in [10], $\mathcal{G}(t)$ is the main object of interest with regard to control-theoretic structure of the system (cf. also [8] on

controllability). Moreover, $\mathcal{G}(t)$ is computable. From the smooth row reduced form (8) of

$$\mathcal{E}_j \mathbf{x}_j + \mathcal{F}_j x = \mathbf{b}_j, \tag{14}$$

$\mathcal{G}(t)$ is determined as the solution space of the equation $M(t)x = 0$, where $M(t)$ is $\rho \times n$ and rank $M(t) = \rho =$ nullity $\mathcal{E}_j(t)$ for all t. We seek descriptions of the dynamics on $\mathcal{G}(t)$ and on a complement of $\mathcal{G}(t)$ in R^n.

4.1. Natural projections associated with a natural completion. As motivation for the following discussion, we note that the natural way to solve for x' at t_0 (when it is uniquely determined) is to apply a row-reduction procedure to (14), and the row-reduction procedure is equivalent to premultiplication by a matrix R. A *completion* of a differential-algebraic equation is an ordinary differential equation whose solution set includes the solutions of the DAE. We discuss the existence of completions after stating the next definition.

DEFINITION 6 ([6]). *Let j have a value such that the conditions* (5), (6), *and* (7) *hold with the "for j" phrase omitted. Let $R(t)$ be smooth and nonsingular with*

$$R(t)\,[\mathcal{E}_j(t)\ \mathcal{F}_j(t)] = \left[\begin{array}{cc|c} I_n & 0 & G \\ 0 & H(t) & K \\ 0 & 0 & M \end{array}\right]. \tag{15}$$

Let the $n \times (j+1)n$ matrix $[R_0(t), \ldots, R_j(t)]$ be the first n rows of $R(t)$, with each $R_i(t)$ an $n \times n$ matrix function. Then the algebraic system (14) *implies that*

$$(D + G)x = \sum_{i=0}^{j} R_i(t)(D^i b)(t) \qquad \left(D = \frac{d}{dt}\right). \tag{16}$$

The ordinary differential equation (16) *is called a* natural completion *of $E(t)x' + F(t)x = b$.*

There are natural completions corresponding to any value of j for which the solvability conditions (5), (6), and (7) hold. One such natural completion is given by

$$x' = -\pi_1(\mathcal{E}_j^{\dagger})\mathcal{F}_j x + \pi_1(\mathcal{E}_j^{\dagger})\mathbf{b}_j, \tag{17}$$

where $A^{\dagger}$ is the Moore–Penrose generalized inverse of A [7] and $\pi_1 : R^{(j+1)n} \mapsto R^n$ denotes projection onto the first n components. The completion (17) is called a *least squares completion* [6]. In many cases it may be possible to use a smaller set of equations than that provided by a given sufficient j [6].

Given a natural completion (16), suppose there exists a projection-valued operator $\overline{P}(t)$ such that

(P1) $\mathcal{R}(\overline{P}(t)) = \mathcal{G}(t)$ for all t, and

(P2) $\overline{P}(D + G) = (D + G)\overline{P}$, i.e., the operator $\overline{P}$ commutes with the operator $D + G$ of the natural completion (16).

Consider the differential equation

$$(D + G(t))x = \overline{P}(t) \sum_{i=0}^{j} R_i(t) D^i b(t). \tag{18}$$

If $x(t)$ is a solution of (18) with $x(t_0) \in \mathcal{G}(t_0)$, then $\overline{P}(t)x(t)$ is also a solution of (18) with the same initial condition at t_0. Thus, $\overline{P}(t)x(t) = x(t)$ and $x(t) \in \mathcal{G}(t)$. The time-varying manifold $\mathcal{G}(t)$ is therefore invariant under (18). In particular, if $x(t)$ is any solution of $E(t)x' + F(t)x = b$, then $\overline{x} \equiv \overline{P}x$ satisfies (18). Equation (18) therefore provides a description of the dynamics of the DAE on $\mathcal{G}(t)$. Our previous arguments show that if the commutativity property (P2) holds and $\mathcal{R}(\overline{P}(t_0)) = \mathcal{G}(t_0)$ for some t_0, then $\mathcal{R}(\overline{P}(t)) \supset \mathcal{G}(t)$ for all t. But if $\overline{P}(t)$ is a smooth projector on the connected interval $\mathcal{I}$, then $\dim \mathcal{R}(\overline{P}(t))$ is constant [13, p. 35], hence $\mathcal{R}(\overline{P}(t)) = \mathcal{G}(t)$.

The next proposition shows that projectors $\overline{P}$ are generated as solutions of the linear matrix differential equation

$$\frac{d}{dt}X(t) = [X(t), G(t)], \tag{19}$$

where $[A(t), B(t)] = A(t)B(t) - B(t)A(t)$.

PROPOSITION 4.1. *Let $D \equiv d/dt$ be the operator of differentiation with respect to t operating on C^∞ functions. Let $G(t)$ be a continuous square matrix function, and let $P(t)$ be a differentiable square matrix function. Then*

1. *$P(t)$ commutes with $D + G(t)$ if and only if $P(t)$ is a solution of* (19).
2. *If P_1, P_2 are solutions of* (19), *then so are P_1P_2 and $P_1 + P_2$.*
3. *If $P(t)$ is a solution of* (19) *and $P^2(t_0) = P(t_0)$ for some t_0, then $P^2(t) = P(t)$ for all t.*

Proof. As operators on C^∞ functions, we have

$$P(D+G) - (D+G)P = (-P' + PG - GP) = (-P' + [P, G]),$$

from which 1 follows. A direct calculation shows that 2 holds, and 3 follows from 2 and uniqueness of solutions of linear differential equations. □

From 2, the set of solutions of (19) is also closed under the bracket [,]. We apply Proposition 4.1 as follows. Choose an initial condition $P_0 = P(t_0)$ for the differential equation (19) that is a projection with range $\mathcal{G}(t_0)$. The solution of (19) will then be a smooth projection-valued operator, having range $\mathcal{G}(t)$ for all t, and satisfying properties (P1), (P2).

DEFINITION 7. *A projection-valued operator $\overline{P}(t)$ satisfying* (P1), (P2) *is called a* natural projection *associated with the given natural completion* (16).

4.2. Decomposition with respect to observability. A system decomposition with respect to observability should include these features:

(1) an identification of the output-nulling space, and a consequent splitting of the state space into "unobservable part ⊕ a complement";

(2) the splitting should satisfy an "invariance" property under the system dynamics, so that the unobservable part (the output-nulling part) may be suitably divided out; and the system modulo the unobservable part should be (at least weakly) observable.

In this section we show that the decomposition described in the next definition meets these requirements.

DEFINITION 8. *Given a natural completion*

$$(D+G)x = \sum_{i=0}^{j} R_i D^i b \tag{20}$$

of a solvable DAE, $E(t)x' + F(t)x = b$, *and given a natural projection* $\overline{P}$ *for* (20), *an* observability resolution *for* (20) *with respect to the output* $y = C(t)x$ *is a set of projections* $\{P_1(t), P_2(t), P_3(t)\}$ *such that*

1. $P_iP_j = 0$ *for* $i \neq j$.
2. $P_i(D + G) = (D + G)P_i$ *for* $i = 1, 2, 3$.
3. $P_1 + P_2 + P_3 = I$.
4. $P_1 + P_2 = \overline{P}$, *the given natural projection.*
5. $\mathcal{R}(P_1(t)) = \mathcal{N}(t)$, *the output-nulling space.*

The existence of such P_i is established in the proof of Theorem 4.2. Since the P_i are computable from a linear matrix ordinary differential equation (ODE), an observability resolution is a type of canonical form that is expressed more closely in terms of the original coefficients and identifies a smooth observable complement of $\mathcal{N}(t)$ in $\mathcal{G}(t)$. If $\{P_1(t), P_2(t), P_3(t)\}$ is an observability resolution for (20) relative to the output matrix $C(t)$, then $\{Q^{-1}P_1Q, Q^{-1}P_2Q, Q^{-1}P_3Q\}$ is an observability resolution for the transformed completion

$$Q^{-1}(D + G)Qx = Q^{-1}\sum_{i=0}^{j} R_iD^ib,$$

relative to the output matrix $C(t)Q(t)$.

Given the output-nulling space $\mathcal{N}(t)$, there is no "distinguished" observable complement in $\mathcal{G}(t)$, but there is a unique observability resolution for given initial data as described in the next theorem.

THEOREM 4.2. *Choose* t_0 *in* $\mathcal{I}$ *and choose a complement* $\mathcal{D}(t_0)$ *of* $\mathcal{N}(t_0)$ *in* $\mathcal{G}(t_0)$. *Let* $\overline{P}$ *be a natural projection for* (20). *Then there is a unique observability resolution for* (20) *with*

$$\ker(P_1(t_0)) = \mathcal{D}(t_0) \oplus \mathcal{R}(I - \overline{P}(t_0)).$$

Proof. Let P_0 be the unique projection such that

$$\mathcal{R}(P_0) = \mathcal{N}(t_0), \qquad \ker(P_0) = \mathcal{D}(t_0) \oplus \mathcal{R}(I - \overline{P}(t_0)). \tag{21}$$

Let $P_1(t)$ be the unique solution of

$$X' = [X, G], \qquad X(t_0) = P_0. \tag{22}$$

Then property 2 holds for $i = 1$ and $P_1(t)$ is a projection for all t. To get an observability resolution, we necessarily define

$$P_2(t) = \overline{P}(t) - P_1(t),$$
$$P_3(t) = I - \overline{P}(t).$$

$P_3(t)$ is a projection because $\overline{P}(t)$ is. Once we establish that $P_2(t)$ is indeed a projection and that $\{P_1, P_2, P_3\}$ is an observability resolution, the uniqueness statement follows.

Since P_1 and $\overline{P}$ satisfy the differential equation in (22), so do P_2 and P_3, and P_1, P_2, P_3 are as smooth as G. We have thus established properties 2, 3, and 4 of Definition 8 for these P_i.

We prove next that P_2 is a projection and that property 1 holds. Since P_1 and $\overline{P}$ satisfy (19), Proposition 4.1 implies that

$$(P_1\overline{P})' = [P_1\overline{P}, G],$$
$$(\overline{P}P_1)' = [\overline{P}P_1, G].$$

Consider the initial condition for these equations at time t_0. From (21) and the fact that $\overline{P}(t_0)$ projects onto $\mathcal{G}(t_0)$ and $\ker(\overline{P}(t_0)) \subset \ker(P_1(t_0))$, we get $P_1(t)\overline{P}(t) = P_1(t)$ for all t and $\overline{P}(t)P_1(t) = P_1(t)$ for all t. It follows that $P_2(t)$ is a projection for all t. Also, $P_1P_3 = P_3P_1 = 0$ because $P_1P_3 = P_1(I - \overline{P})$ and $P_3P_1 = (I - \overline{P})P_1$. And $P_1P_2 = P_2P_1 = 0$ because $P_1P_2 = P_1(\overline{P} - P_1)$ and $P_2P_1 = (\overline{P} - P_1)P_1$. Using $\overline{P} = P_1 + P_2$, we have $P_2\overline{P} = \overline{P}P_2 = P_2$ since P_2 is a projection. Thus, $P_3P_2 = (I - \overline{P})P_2 = P_2 - P_2 = 0$ and $P_2P_3 = P_2(I - \overline{P}) = 0$. This completes the proof of property 1.

To prove property 5, note that we have $\mathcal{R}(P_1(t_0)) = \mathcal{N}(t_0)$. By properties 1 and 2, $\mathcal{R}(P_i)$ is a reducing subspace for $D + G$ for $i = 1, 2, 3$. Therefore, by property 3, $D + G$ has a representation of the form

$$\begin{bmatrix} D+G_1 & 0 & 0 \\ 0 & D+G_2 & 0 \\ 0 & 0 & D+G_3 \end{bmatrix}$$

in the coordinates $z_i \in \mathcal{R}(P_i)$ on C^∞ functions. We now show that $\mathcal{N}(t) \subset \mathcal{R}(P_1(t))$ for all t. Suppose $z(t) \in \mathcal{N}(t)$ is an output-nulling solution. By (21) and property 1, we know that $z_2(t_0) = P_2(t_0)z(t_0) = 0$ and $z_3(t_0) = P_3(t_0)z(t_0) = 0$, so $z(t_0) = P_1(t_0)z(t_0) = z_1(t_0)$. Any solution of the homogeneous equation must satisfy the differential equations

$$\begin{aligned} (D+G_1)z_1(t) &= 0, \\ (D+G_2)z_2(t) &= 0, \\ (D+G_3)z_3(t) &= 0. \end{aligned} \tag{23}$$

Since $z_2(t_0) = 0$ and $z_3(t_0) = 0$, we must have $z_2(t) = 0$, $z_3(t) = 0$ for all t. Therefore, $z(t) = z_1(t) \in \mathcal{R}(P_1(t))$ for all t. Since $z(t)$ was an arbitrary output-nulling solution, we have $\mathcal{N}(t) \subset \mathcal{R}(P_1(t))$ for all t. Since $\operatorname{rank} P_1(t)$ is constant, $\mathcal{R}(P_1(t)) = \mathcal{N}(t)$ for all t. This proves property 5 of the definition. □

We now use the resolution $\{P_1, P_2, P_3\}$ to *generate* a coordinate transformation to a system form whose new coordinates explicitly exhibit a decoupling of the three system components. It is possible to do this *locally*, in a neighborhood of each t_0 in $\mathcal{I}$, by the following procedure:

Let $n_j = \dim \mathcal{R}(P_j)$ for $j = 1, 2, 3$. Let t_0 be in $\mathcal{I}$.

1. Choose n_j independent columns of $P_j(t_0)$ for $j = 1, 2, 3$. These columns remain independent on some interval $(t_0 - \epsilon, t_0 + \epsilon)$.

2. Form the matrix $Q(t) = [\hat{P}_1(t)\, \hat{P}_2(t)\, \hat{P}_3(t)]$, where $\hat{P}_j(t)$ is an $n \times n_j$ matrix whose columns are the columns of $P_j(t)$ chosen in 1. By construction, $Q(t)$ is smooth and nonsingular on $(t_0 - \epsilon, t_0 + \epsilon)$.

3. Define new coordinates w by

$$x = Q(t)w = [\hat{P}_1(t)\, \hat{P}_2(t)\, \hat{P}_3(t)] \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}. \tag{24}$$

In the w coordinates, the resolution $\{P_j\}$ becomes $\{\tilde{P}_j\}$, where, for $j = 1, 2, 3$, $\tilde{P}_j = Q^{-1}P_jQ = \operatorname{diag}[\delta_{1j}I_{n_1}\ \delta_{2j}I_{n_2}\ \delta_{3j}I_{n_3}]$, where δ_{ij} is the Kronecker delta.

COROLLARY 4.3. *Given a natural completion*

$$(D+G)x = \sum_{i=0}^{j} R_i D^i(Bu)$$

and given an observability resolution $\{P_1, P_2, P_3\}$ *relative to a given* $\overline{P}$ *and output matrix* $C(t)$*, the transformation* $x = Q(t)w$ *defined in* (24) *produces a new completion, defined on* $(t_0 - \epsilon, t_0 + \epsilon)$*, which has the form*

$$w_1' + \tilde{G}_1(t)w_1 = \tilde{P}_1 Q^{-1} \sum_{i=0}^{j} R_i D^i(Bu), \tag{25}$$

$$w_2' + \tilde{G}_2(t)w_2 = \tilde{P}_2 Q^{-1} \sum_{i=0}^{j} R_i D^i(Bu), \tag{26}$$

$$w_3' + \tilde{G}_3(t)w_3 = \tilde{P}_3 Q^{-1} \sum_{i=0}^{j} R_i D^i(Bu). \tag{27}$$

Given an input $u(t)$*, let* x_0 *be consistent for the original* DAE *at* t_0*, and let* $w_0 = Q^{-1}(t_0)x_0$*. Then, with this restriction on initial conditions* w_{30} *in* (27)*, the system for* w_2 *and* w_3*, with output*

$$y = C(t)Q(t)w = C(t)\hat{P}_2(t)w_2 + C(t)\hat{P}_3(t)w_3$$

is (at least weakly) observable on any closed subinterval of $(t_0 - \epsilon, t_0 + \epsilon)$*.*

Proof. The only part of the statement that requires proof is the fact that the coefficient of w, namely, $\tilde{G} \equiv Q^{-1}Q' + Q^{-1}GQ$, has a block diagonal form as stated. The $\tilde{P}_j$ are constant, as shown above, and satisfy $\tilde{P}_j' = [\tilde{P}_j, \tilde{G}]$, since they commute with $(D + \tilde{G})$. Thus, $[\tilde{P}_j, \tilde{G}] = 0$, so $\tilde{P}_j$ commutes with $\tilde{G}$ for $j = 1, 2, 3$. Since the $\tilde{P}_j$ themselves commute, $\tilde{G}$ is block diagonal. □

The original derivation of the general form for solvable systems [5] uses a matrix of independent solutions of the homogeneous equation as part of a globally defined coordinate transformation to the new form. The approach of Corollary 4.3 is local, but does not require explicit knowledge of system solutions. By this approach, finding the desired transformation requires computation of P_1, P_2, and P_3 from a linear ODE (equivalent to a vector system of size n^2), once $G(t)$ is itself known. If a splitting of $\mathcal{G} = \mathcal{N} \oplus \mathcal{D}$ according to observability is not needed, then a computation of $n_1 + n_2 = \dim \mathcal{G}$ independent columns of $\overline{P}$, plus a complementary set of columns from P_3, is required. We note that the decomposition of Corollary 4.3 is similar to one given in [19] but does not follow from the results of that paper, since we do not assume constant rank of our coefficient matrices and we do not require coordinate changes to reach the decomposition.

For linear time-varying nonsingular systems, the paper [24] establishes structural forms with respect to observability (and controllability) by proving the existence of appropriate coordinate transformations. The new features in Corollary 4.3 are (i) the *generation* of an appropriate transformation, using computable projections onto relevant subspaces, and (ii) applicability to singular as well as nonsingular systems. We now outline an algorithm for the pointwise computation of the system decomposition.

(1) The solution manifold $\mathcal{G}(t)$ is pointwise computable from array $[\mathcal{F}_j\ \mathcal{E}_j]$.

(2) The matrix $G(t)$ of the natural completion is computable from $[\mathcal{F}_j\ \mathcal{E}_j]$.

(3) The projections P_i are computable from a linear ODE involving $G(t)$ and knowledge of the solution space $\mathcal{G}(t)$ and output-nulling space $\mathcal{N}(t)$.

(4) The transformation $Q(t)$ may be computed pointwise from the P_i by choosing independent columns.

(5) The remaining element required to compute the matrix $\tilde{G}(t)$ of the transformed completion is the derivative Q'. But $P_i' = [P_i, G]$, so Q' is pointwise computable by choosing the appropriate columns of $[P_i, G]$.

5. An example. We end the paper with a simple example to illustrate the algorithm just outlined for the calculation of the projections, and the decoupled completion of Corollary 4.3.

Example 2. Consider the system with $B(t) = I$ and

$$E(t) = \begin{bmatrix} 1 & -t & 0 & 0 \\ 0 & 1 & t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad F(t) = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The output is given by $y = x_2 + x_3 + x_4$. This system is index one. It is possible to backsolve from the array $[\mathcal{E}_1 \ \mathcal{F}_1 \ \mathbf{u}_1]$ to get x' for a natural completion. One may take the matrices R_0 and R_1 to be

$$R_0 = \begin{bmatrix} 1 & t & -t^2 & 0 \\ 0 & 1 & -t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad R_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and $G(t)$ is given by

$$G(t) = \begin{bmatrix} 0 & -1 & t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The manifold $\mathcal{G}$ is defined by $x_4 = 0$. We take $\overline{P} = \operatorname{diag}[1\ 1\ 1\ 0]$, since this projection has range $\mathcal{G}$ and commutes with $D+G(t)$. If the output is identically zero for a solution of the homogeneous equation, then necessarily we have $x_2 = -x_3 = -c_3$, c_3 constant. But the second system equation (or the second equation of the completion) implies that $x_2' = -c_3$, hence $c_3 = 0$. The remaining equation $x_1' = 0$ is satisfied for any $x_1 = c_1$, c_1 constant. Thus, the space $\mathcal{N}$ is given as $\mathcal{N}(t) = \{[c_1\ 0\ 0\ 0]^T : c_1 \textit{ arbitrary}\}$. As the initial condition for a projection P_1 onto $\mathcal{N}$ we take $P_1(0) = \operatorname{diag}[1\ 0\ 0\ 0]$. Integration of the equation $P_1' = [P_1, G]$ yields

$$P_1(t) = \begin{bmatrix} 1 & -t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then $P_1(t)x \in \ker[0\ 1\ 1\ 1] = \ker C(t)$. The projection $P_2(t)$ is calculated as

$$P_2 = \overline{P} - P_1 = \begin{bmatrix} 0 & t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

while P_3 is then

$$P_3 = I - \overline{P} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

We may define a coordinate transformation using the given resolution $\{P_1, P_2, P_3\}$ by setting

$$Q(t) = \begin{bmatrix} 1 & t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad x = Q(t)w,$$

which, in this case, is a global transformation. The new transformed completion in w coordinates is given by

$$\tilde{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

together with

$$\tilde{R}_0 = Q^{-1}R_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -t & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \tilde{R}_1 = Q^{-1}R_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The output is $y = [0\ 1\ 1\ 1]Q(t)w = [0\ 1\ 1\ 1]w$. The observable dynamics and output equation are given by

$$w_2' + w_3 = u_2 - tu_3, \tag{28a}$$

$$w_3' = u_3, \tag{28b}$$

$$y = w_2 + w_3 + u_4. \tag{28c}$$

The dynamics for w_1 and w_4 may also be obtained from Corollary 4.3; in this case, $w_1' = u_1$, and $w_4' = u_4'$.

As a final remark, we note that system (28) in Example 2 is smoothly observable. In fact, two differentiations of the output equation (28c) and substitution for w_1', w_2' from (28a), (28b) is sufficient to solve for the observable state in terms of input, output, and their derivatives.

6. Conclusion. In this paper we defined the output-nulling space of a linear time-varying singular control system. We showed how to decompose the system, by means of projection operators, into (unobservable subspace) $\oplus$ (observable complement) relative to a given output structure. The relevant subspaces and projectors were shown to be computable for a large class of control systems by pointwise linear algebra from information in the system's derivative array. We described the dynamics on these spaces by the projection operators. Finally, we illustrated the dynamic system decomposition by applying our algorithm to a simple example.

Acknowledgment. I thank my thesis advisor, Dr. Stephen L. Campbell, for helpful comments in the preparation of this paper. In addition, I thank the two referees for thorough comments and suggestions that resulted in an improved exposition.

REFERENCES

[1] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *The Numerical Solution of Initial Value Problems in Ordinary Differential-Algebraic Equations*, Elsevier, New York, 1989.

[2] S. L. Campbell, *Singular Systems of Differential Equations* II, Research Notes in Mathematics No. 61, Pitman, Boston, MA, 1982.

[3] ———, *The numerical solution of higher index linear time varying singular systems of differential equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 334–348.

[4] ———, *Consistent initial conditions for linear time varying singular systems*, in Frequency Domain and State Space Methods for Linear Systems, C. I. Byrnes and A. Lindquist, eds., Elsevier–North Holland, Amsterdam, 1986, pp. 313–318.

[5] ———, *A general form for solvable linear time varying singular systems of differential equations*, SIAM J. Math. Anal., 18 (1987), pp. 1101–1115.

[6] ———, *Uniqueness of completions for linear time-varying differential algebraic equations*, Linear Algebra Appl., 161 (1992), pp. 55–67.

[7] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Pitman, Boston, MA, 1979; Reprinted by Dover, New York, 1990.

[8] S. L. Campbell, N. Nichols, and W. J. Terrell, *Duality, observability and controllability for linear time-varying descriptor systems*, Circuits, Systems Signal Process., 10 (1991), pp. 455–470.

[9] S. L. Campbell and L. R. Petzold, *Canonical forms and solvable singular systems of differential-algebraic equations*, SIAM J. Algebraic Discrete Meth., 4 (1983), pp. 517–521.

[10] S. L. Campbell and W. J. Terrell, *Observability of linear time varying descriptor systems*, SIAM J. Matrix Anal. Appl., 3 (1991), pp. 484–496.

[11] M. A. Christodoulou and P. N. Paraskevopoulos, *Solvability, controllability, and observability of singular systems*, J. Optim. Theory Appl., 45 (1985), pp. 53–72.

[12] D. Cobb, *Controllability, observability, and duality in singular systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1076–1082.

[13] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1982.

[14] E. Kriendler and P. E. Sarachik, *On the concepts of observability and controllability*, IEEE Trans. Automat. Control, 64 (1964), pp. 129–136.

[15] F. L. Lewis, *Fundamental, reachability, and observability matrices for discrete descriptor systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 502–505.

[16] B. G. Mertzios, M. A. Christodoulou, B. L. Syrmos, and F. L. Lewis, *Direct controllability and observability time domain conditions for singular systems*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 788–790.

[17] R. W. Newcomb, *The semistate description of nonlinear time-variable circuits*, IEEE Trans. Circuits and Systems, CAS-28 (1981), pp. 62–71.

[18] R. W. Newcomb and B. Dziurla, *Some circuits and systems applications of semistate theory*, Circuits, Systems Signal Process., 8 (1989), pp. 235–260.

[19] A. J. van der Schaft, *Representing a nonlinear state space system as a set of higher-order differential equations in the inputs and outputs*, Systems Control Lett., 12 (1989), pp. 151–160.

[20] L. M. Silverman and H. E. Meadows, *Controllability and observability in time-variable linear systems*, SIAM J. Control, 5 (1967), pp. 64-73.

[21] W. J. Terrell, *Observability and External Description of Linear Time Varying Singular Control Systems*, Ph.D. thesis, Department of Mathematics, North Carolina State University, Raleigh, NC, 1990.

[22] E. L. Yip and R. F. Sincovec, *Solvability, controllability, and observability of continuous descriptor systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 702–707.

[23] L. Weiss, *The concepts of differential controllability and differential observability*, J. Math. Anal., 10 (1965), pp. 442–449.

[24] L. Weiss, *On the structure theory of linear differential systems*, SIAM J. Control, 6 (1968), pp. 659–680.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 3, pp. 890–916, May 1994

014

NEW EXISTENCE RESULTS FOR OPTIMAL CONTROLS IN THE ABSENCE OF CONVEXITY: THE IMPORTANCE OF EXTREMALITY*

ERIK J. BALDER†

Abstract. A new approach to "existence without convexity" is presented. Instead of applying a Weierstrass–Tonelli-type existence result to a convexified (i.e., relaxed) form of the control problem, which is standard, it uses a new extremum principle. This principle guarantees the existence of an optimal relaxed control function that is a convex combination of relaxations of ordinary control functions. The same linear relationship is maintained for the corresponding trajectories, and this is of much benefit in the final phase, which consists of a deconvexification by a global subgradient argument and Lyapunov's theorem. The extremum principle is based on extreme point features of the relaxed control problem. It can be stated in an abstract form and extends the classical extremum principle of Bauer. The new approach applies both to optimal control problems of suitable type (dynamics linear in the state variable) and to abstract variational problems without dynamics, well-known in economics. With a different interpretation, it can also give bang-bang existence results. Several new applications serve to illustrate the power of the approach.

Key words. optimal control, existence without convexity, Bauer extremum principle, extreme points, Young measures, relaxed controls

AMS subject classifications. 49J15, 49J27, 49J45, 46A55

1. Introduction.

This paper introduces a new, quite general approach to the subject of "existence without convexity" for optimal control. It may be recalled that control problems without the usual convexity conditions for the orientor field do not lend themselves directly to the traditional approach in the spirit of Tonelli, because the corresponding integral functionals fail to have suitable lower semicontinuity properties. Yet such problems may have optimal solutions, as was first demonstrated by Neustadt [27]; see [15, Chap. 16] for an extensive account. In recent years, existence problems of this kind have gained increasing prominence in mechanics (e.g., see [11]).

The approach presented here differs notably from the usual ones by the use of a new extremum principle, which can be stated in an abstract form and generalizes the well-known extremum principle of Bauer [19, 13.A-B]. As particular cases, the main existence result of this paper contains the traditional existence results of this kind (e.g., [27], [28], [15], [24]), the recent nontraditional results of Raymond [29], [31], Cellina and Colombo [14], and Mariconda [25], as well as the abstract existence results for variational problems without dynamics, as given by Aumann and Perles [3], Berliocchi and Lasry [12], Artstein [2], and Balder [5], [6]. In addition to these connections with the existing literature, several new applications are produced, even within classical contexts.

To illustrate some of the key ideas, let us tackle in this section the existence question for the following special optimal control problem (B_0):

$$\text{minimize } I(u) := \int_0^1 [(u^2(t)-1)^2 - y_u^2(t)]dt - \exp\left(-\sup_{t\in[0,1]} |y_u(t)|\right)$$

* Received by the editors December 31, 1990; accepted for publication (in revised form) January 26, 1993.

† Mathematical Institute, University of Utrecht, P.O. Box 80.010, 3508 TA Utrecht, the Netherlands (balder@math.ruu.nl).

over all integrable functions $u : [0,1] \to \boldsymbol{R}$. Here y_u, the absolutely continuous trajectory corresponding to u, is given by

$$y_u(t) = \int_0^t u(\tau)\, d\tau.$$

We can easily show that, if the term $-y_u^2(t)$ were to be changed to $+y_u^2(t)$, the resulting optimal control problem does not have an optimal solution. In fact, if also—next to this change of sign—the exponential term in $I(u)$ is omitted, the resulting problem constitutes a classical counterexample to existence, which is due to Bolza (see [21, pp. 304–305]). Nevertheless, (B_0) has an optimal solution, and much of this section is devoted to proving this. Even though (B_0) is a very special problem, this constitutes a new existence result. It cannot be obtained from the current literature, either directly or indirectly. However, without the exponential term, the existence result would be of the recent nontraditional type found in [14], [29], [31]; this type is characterized by the presence of a trajectory cost integrand that is *concave* in the state variable. (The same kind of of concavity was also studied in the unpublished paper [4], but only for control problems having a compact space of control points: In fact, under the additional constraint $|u(t)| \leq 1$ almost everywhere, the above existence result for (B_0)—with the exponential term as stated—would also follow from [4].)

The existence result for (B_0) follows from Theorem 2.2, which contains an existence result "without convexity" for a control problem (B) that is much more general than (B_0). The proof of Theorem 2.2 is based on the three phases of the *convexification-deconvexification scheme* given below:

(i) *Phase* 1 (*convexification*). Determine the relaxed form (B_{rel}) of (B).

(ii) *Phase* 2 (*extremum principle*). Apply the abstract extremum principle (Theorem 3.1) to (B_{rel}); this guarantees the existence of a special optimal relaxed solution of (B_{rel}) (having the so-called *Minkowski form*).

(iii) *Phase* 3 (*deconvexification*). Use the special optimal relaxed solution of Phase 2 to obtain an optimal solution for (B).

Standard approaches to existence without convexity use a theorem of the Weierstrass–Tonelli type in Phase 2 and therefore fail to take advantage of the geometric—in particular extremal—features of the relaxed problem (B_{rel}).

A short-cut to the key ideas of this paper is obtained when we first work out the convexification-deconvexification scheme for the simple problem (B_0), instead of (B).

Phase 1. *Convexification via relaxation.* The motivation for relaxation comes from the extremum principle: It only applies to minimization problems with a (partly) concave (or quasi-concave) objective function. Since the integrand $v \mapsto (v^2-1)^2$ of I is neither convex nor concave, relaxation provides the necessary concavity for the objective function. Incidentally, it should be observed that practically all approaches to "existence without convexity" initially convexify toward existence anyway; in essence, this is done to create a setting in which the lower semicontinuity and compactness needs of the Tonelli–Weierstrass approach can be met.

The relaxed version $(B_{0,\text{rel}})$ of the simple problem (B_0) has the following objective function:

$$J(x) := \int_0^1 \left[\int_{\boldsymbol{R}} (v^2-1)^2\, x(t)(dv)\right] dt - \int_0^1 y_x^2(t)dt - \exp\left(-\sup_{0\leq t\leq 1} |y_x(t)|\right).$$

Here x is a transition probability (alias Young measure) from [0,1] into $M_1^+(\boldsymbol{R})$, with $M_1^+(\boldsymbol{R})$ denoting the set of all Borel probability measures on $\boldsymbol{R}$; see §4 for more

details about Young measures. Let us denote the set of all Young measures from $[0,1]$ into $M_1^+(\boldsymbol{R})$ by C. Above, the relaxed trajectory corresponding to $x \in C$ is denoted by y_x. It is given by

$$y_x(t) := \int_0^t \left[\int_{\boldsymbol{R}} \upsilon \, x(\tau)(d\upsilon) \right] d\tau,$$

provided that these integrals exist. Formally, for trajectories corresponding to an "original" control function u, we should now write y_{ϵ_u} instead of y_u, but this distinction will be neglected, as it is obvious what is meant. Clearly, $x \mapsto y_x(t)$ is affine for every fixed t. Hence J is concave.

By itself, concavity of the objective is not enough to bring the extremum principle of Phase 2 to bear. Also essential is the presence of a constraint $b_0(x) \leq 0$, with $b_0 : C \to (-\infty, +\infty]$ an affine *inf-compact* function (that is, the level sets $\{x \in C : b_0(x) \leq \gamma\}$ must be compact for every $\gamma \in \boldsymbol{R}$). It is not immediately obvious how such a constraint could be found for (B_0). Consider, however, the following reasoning ad hoc. For the constant function $u \equiv 0$, we evidently have $I(0) = 0$. Using the Cauchy–Schwarz inequality, it is easy to see that $I(u) \leq 0$ implies that $\int_0^1 (u^4(t) - 3u^2(t))dt \leq 0$. In turn, this gives $\int_0^1 u^4(t)\, dt \leq 9$. We conclude that this constraint can be added to (B_0) without altering the optimization problem in any essential way, i.e.,

$$\inf\{I(u) : u \in L_1([0,1])\} \;=\; \inf\left\{ I(u) : u \in L_1([0,1]), \int_0^1 u^4(t)\, dt \leq 9 \right\}.$$

To the relaxed problem $(B_{0,\mathrm{rel}})$ we now simply add the relaxation of this constraint: Let $(B_{0,\mathrm{rel}})$, the relaxed version of (B_0), be given by

$$\text{minimize } J(x) := \int_0^1 \left[\int_{\boldsymbol{R}} (\upsilon^2 - 1)^2 \, x(t)(d\upsilon) \right] dt - \int_0^1 y_x^2(t) dt - \exp\left(- \sup_{0 \leq t \leq 1} |y_x(t)| \right)$$

over all relaxed control functions $x : [0,1] \to M_1^+(\boldsymbol{R})$ satisfying

$$b_0(x) := \int_0^1 \left[\int_{\boldsymbol{R}} (\upsilon^4 - 9) \, x(t)(d\upsilon) \right] dt \;\leq\; 0.$$

From well-known facts about Young measures, it follows that b_0 is inf-compact (see §4). A beneficial side-effect of introducing the constraint is that the integral in the definition of $y_x(t)$ certainly exists when $x \in C$ satisfies $b_0(x) \leq 0$.

Phase 2. Application of the extremum principle. After the aforegoing relaxation, the new extremum principle (see §3) can be applied to the problem $(B_{0,\mathrm{rel}})$. In this phase, the approach of this paper deviates from the standard in an essential way. While the latter merely apply some result of the Weierstrass–Tonelli variety to the relaxed problem, the extremum principle guarantees the existence of an optimal solution $x_* \in C$ for $(B_{0,\mathrm{rel}})$ that has the following special *Minkowski form*:

$$x_*(t) \;=\; \lambda \epsilon_{u_1}(t) + (1 - \lambda)\epsilon_{u_2}(t) \quad \text{a.e.}$$

Here $\lambda \in [0,1]$ is a constant, and u_1, u_2 are two measurable—in fact, integrable—functions from [0,1] into $\boldsymbol{R}$. (Recall the notation involved: $\epsilon_u(t)$ denotes the Dirac (probability) measure concentrated at the point $u(t)$, $t \in [0,1]$.) Optimal solutions

in Minkowski form were also investigated in [4]. For the remainder of the paper, it is very important that the trajectories inherit an equally simple linear relationship,

$$y_{x_*} = \lambda y_{u_1} + (1-\lambda) y_{u_2}. \tag{1}$$

This differs completely from what the standard approaches can deliver; e.g., see [15], [14], [31]. At best, they prove the existence of an optimal relaxed control function $\tilde{x} \in C$ of the form

$$\tilde{x}(t) = \sum_k \lambda_k(t) \epsilon_{u_k}(t),$$

with *t-dependent* convex coefficients $\lambda_k(t)$. However, it is obvious that any linear relationship like (1) between the corresponding trajectories is lost in the process.

Phase 3. *Deconvexification.* First, let us check that $\inf(B_{0,\mathrm{rel}}) \leq \inf(B_0)$. For any ordinary control function $u : [0,1] \to \boldsymbol{R}$ with $\int_0^1 u^4 \leq 9$, the relaxation $x := \epsilon_u$ satisfies $b_0(x) \leq 0$, and, evidently, $y_x = y_u$ and $J(x) = I(u)$. So $\inf(B_{0,\mathrm{rel}}) \leq \inf_{u \in L_1([0,1])}\{I(u) : \int_0^1 u^4 \leq 9\} = \inf(B_0)$, by what we concluded in Phase 1. By optimality of x_*, found in Phase 2, it follows that a sufficient condition for a measurable control function $u : [0,1] \to \boldsymbol{R}$ to be optimal for (B_0) is that (a) $b_0(\epsilon_u) \leq 0$ and (b) $I(u) \leq J(x_*)$; let us see how (a) and (b) can be ensured for u. Define the trajectory cost functional $\bar{r} : C([0,1]) \to \boldsymbol{R}$ by

$$\bar{r}(y) := -\int_0^1 y^2(t)\, dt - \exp(-\|y\|),$$

where the supremum norm $\|y\| := \sup_{t \in [0,1]} |y(t)|$ is used. We already saw that $\bar{r}$ is concave on $C([0,1])$. Clearly, $\bar{r}$ is also continuous in the supremum norm on $C([0,1])$. Hence, there exists a subgradient $y^* \in (C([0,1]))^*$ to the convex function $-\bar{r}$ in the point y_{x_*}, i.e.,

$$\bar{r}(y_{x_*}) - \bar{r}(y) \geq \langle y - y_{x_*}, y^* \rangle \quad \text{for all } y \in C([0,1]).$$

Let the control cost functional r be given by

$$r(x) := \int_0^1 \left[\int_{\boldsymbol{R}} (v^2 - 1)^2 \, x(t)(dv) \right] dt.$$

An obvious sufficient condition to guarantee that (a) and (b), above, hold, hence to guarantee u's optimality for (B_0), is given by the following system (A_0) of three equations:

$$\begin{aligned} b_0(\epsilon_u) &= b_0(x_*), \\ r(\epsilon_u) &= r(x_*), \\ \langle y_u, y^* \rangle &= \langle y_{x_*}, y^* \rangle. \end{aligned}$$

Indeed, the first equation implies that $b_0(\epsilon_u) = b_0(x_*) \leq 0$, and the third one causes $\bar{r}(y_u) \leq \bar{r}(y_{x_*})$ by the subgradient property. Together with the second equality, this gives $I(u) = J(\epsilon_u) \leq J(x_*)$, establishing (a) and (b).

Now $y^* \in (C([0,1]))^*$ has a Riesz representation [37, I.5.8]. That is, there exists a bounded signed Borel measure θ on $[0,1]$ such that $\langle y, y^* \rangle = \int_{[0,1]} y(t)\, \theta(dt)$ for all $y \in C([0,1])$. By the Minkowski form of x_* and Fubini's theorem, we can rewrite

$$\langle y_u - y_{x_*}, y^* \rangle = \int_0^1 [u(t) - \lambda u_1(t) - (1-\lambda) u_2(t)] \theta((t,1])\, dt,$$

thanks to the linear relationship in (1). Similar substitutions can also be performed directly in the two other equalities. This leads to the following equivalent form of (A_0):

$$\int_0^1 u^4(t)\, dt = \lambda \int_0^1 u_1^4(t)\, dt + (1-\lambda) \int_0^1 u_2^4(t)\, dt,$$

$$\int_0^1 (u^2(t) - 1)^2\, dt = \lambda \int_0^1 (u_1^2(t) - 1)^2\, dt + (1-\lambda) \int_0^1 (u_2^2(t) - 1)^2\, dt,$$

$$\int_0^1 u(t) \theta((t,1])\, dt = \lambda \int_0^1 u_1(t) \theta((t,1])\, dt + (1-\lambda) \int_0^1 u_2(t) \theta((t,1])\, dt.$$

Clearly, by Lyapunov's theorem, this system can be solved for some measurable u, which is then the optimal control for (B_0), by the properties exhibited for (A_0).

As was already observed, the linear relationship (1) between the trajectories is lost in the approaches that have appeared in the literature. Fortunately, in the traditional existence results without convexity [15], the trajectory cost integrand is *linear* in the state variable. Therefore, the entire trajectory cost can be integrated away, so that essentially one is left with only a control cost and—possibly—a final cost term. However, the recent nontraditional existence results mentioned above have the trajectory cost integrand concave in the state variable, just like problem (B_0). Then such an integration device can no longer be used, and the resulting nonlinear relationship between the trajectories of $\tilde{x}$ and of the u_i's leads to quite involved applications of Lyapunov's theorem; see [14, pp. 101–105] and [31, pp. 122–125]. These involve not only integrable selections of subgradients [14], but also an infinite sequence of iterates of an integral operator in the more general—but still quite modest—context of [31] (optimal control of a linear differential equation). The ad hoc character of these arguments is evident; moreover, they display a worrisome tendency to grow more complex as the dynamical system becomes more general. The inclusion of the extremum principle in Phase 2 of the convexification-deconvexification scheme avoids such complications altogether and reaches much further.

The organization of this paper is as follows. Section 2 presents the general optimal control problem (B) and contains in Theorem 2.2 the principal existence result "without convexity" of this paper. A version of this existence result, for variational problems without dynamics, is given in Theorem 2.3. Section 3 can also be read independently from the other sections. It contains the new extremum principle (Theorem 3.1), formulated for an abstract optimization problem in a topological vector space setting. This principle provides sufficient conditions for the existence of an optimal solution that is a (finite) convex combination of extreme points of the optimization domain (this is the natural abstraction of the Minkowski form introduced before). Theorem 3.1 contains the well-known Bauer extremum principle. In §4 Theorems 2.2 and 2.3 are proved by working out the convexification-deconvexification scheme. First, the relaxed version (B_{rel}) of (B) is formulated; it is based on an extra hypothesis that makes it possible to apply Theorem 3.1 directly to (B_{rel}). Phase 3

(deconvexification) then proceeds by solving a system (A) of linear equations, much like we did above with (A_0). Finally, by means of deparametrization, it is shown that, for the existence result to hold, the extra hypothesis is not really needed. In §5 several new applications show the power of Theorems 2.2 and 2.3.

The approach followed in this paper can also be used to obtain new *bang-bang*-type existence results for problems with suitably concave cost functions. These applications are of a different nature, however; refer to [10] for the details.

2. Existence without convexity. Two new, general existence results without convexity are presented in this section. The first and most important one is a quite general existence result for optimal control problems. The second one is actually a special version, formulated for certain variational problems that lack a dynamical system altogether.

Let T be a compact metric space, the *abstract time domain*, equipped with its Borel σ-algebra $\mathcal{B}(T)$ and a *nonatomic* finite measure μ. Let $\mathcal{T}$ denote either $\mathcal{B}(T)$ or its μ-completion. Let U be a metric Lusin or Suslin space of *control points*, equipped with the Borel σ-algebra $\mathcal{B}(U)$. To simplify matters, we may think of U as a Euclidean or a complete separable metric space. Let $t \mapsto U(t)$ be a multifunction from T into U, having a $\mathcal{T} \times \mathcal{B}(U)$-measurable graph M. A *control pair* is a pair (π, u), where π, the *control parameter*, belongs to a compact metric space Π, where $\Gamma(\pi)$ is the abstract *time domain* determined by π, and where $u : \Gamma(\pi) \to U$ is a measurable *control function* such that $u(t) \in U(t)$ for almost every t in $\Gamma(\pi)$. Let $\mathcal{G}$ denote the set of all such control pairs (π, u). As shown in Example 5.1, control parameters can be used not only to specify the control time domain, but also the initial and/or final positions of the dynamical systems, hitting times, and/or positions with a specified target, and so forth (see also [37]). The following continuity condition is supposed to hold for the μ-measure of the symmetric differences of the time domains:

$$\lim_{k\to\infty} \mu(\Gamma(\pi_k)\Delta\Gamma(\pi_0)) = 0 \quad \text{whenever } \pi_k \to \pi_0 \text{ in } \Pi.$$

For instance, in the control problem (B_0) of §1, we had $M := [0,1] \times \boldsymbol{R}$, and the space Π is trivial (say $\Pi := \{0\}$). Thus, we can set there $\Gamma(0) := T := [0,1]$. Of course, T is then equipped with the Lebesgue measure μ.

Let us start the description of the trajectories of the control problem. Let $b : M \to \boldsymbol{R}^\nu$ be a measurable function and let $\kappa : T \times T \to \boldsymbol{R}^{n\times\nu}$ be bounded in absolute value by a constant c_κ, measurable in its second variable, and *continuously hereditary* in the following way [37, II.5]: There exists a continuous functional $\sigma : T \to [0,1]$ such that $\alpha \mapsto \mu(T_\alpha)$ is also continuous, for $T_\alpha := \{t \in T : \sigma(t) \le \alpha\})$, and such that

$$\kappa(t,\tau) = 0 \quad \text{if } \sigma(t) < \sigma(\tau).$$

The following expression defines the *control action integral operator*:

$$\mathcal{A}(\pi,u)(t) := \int_{\Gamma(\pi)} \kappa(t,\tau) b(\tau, u(\tau)) \mu(d\tau).$$

A control pair $(\pi, u) \in \mathcal{G}$ is called *admissible* if it satisfies

$$\int_{\Gamma(\pi)} |b(t,u(t))| \mu(dt) \; < \; +\infty.$$

In §4 it is demonstrated that $\mathcal{A}(\pi, u)$ belongs to $C(T; \boldsymbol{R}^n)$ whenever $(\pi, u) \in \mathcal{G}$ is admissible. The *trajectory* $y_{\pi,u}$ in $C(T; \boldsymbol{R}^n)$, corresponding to an admissible pair $(\pi, u) \in \mathcal{G}$, is defined by

$$y_{\pi,u} := \Phi(\pi, \mathcal{A}(\pi, u)).$$

Here the *representation operator* $\Phi : \Pi \times C(T; \boldsymbol{R}^n) \to C(T; \mathbf{R}^n)$ is supposed to be such that $\Phi(\pi, y)$ is continuous in (π, y) and affine in y. Moreover, constants $c_{\Phi,1}, c_{\Phi,2}$ are supposed to exist such that

$$\|\Phi(\pi, y)\| \leq c_{\Phi,1}\|y\| + c_{\Phi,2} \quad \text{on } \Pi \times C(T; \boldsymbol{R}^n).$$

Similar to the previous section, $\|\cdot\|$ denotes the supremum norm here. For motivation for these concepts, skip to Example 5.1. As is shown in §5, linear ordinary differential equations, linear Volterra equations, and linear functional-integral equations can be covered in this way. For instance, for (B_0) in §1, we had $b(t, v) = v$, $\kappa(t, \tau) = 1$ if $\tau \leq t$ and $\kappa(t, \tau) = 0$ otherwise; σ and Φ were the respective identity functions, and $\mu(T_\alpha) = \alpha$.

Let $g_0, g_1, \ldots, g_m : M \to (-\infty, +\infty]$ be measurable functions. Carrying a notion due to Cesari to its logical extreme, we define the *orientor field* $Q : T \times \boldsymbol{R}^\nu \to 2^{\boldsymbol{R} \times [-\infty, +\infty]^m}$, corresponding to $g_0, \ldots, g_m$ and b, by letting $Q(t, v)$ be the set of all $(w^0, w) \in \boldsymbol{R} \times [-\infty, +\infty]^m$ such that

$$w^0 \geq g_0(t, v), \ldots, w^m \geq g_m(t, v) \quad \text{and} \quad v = b(t, v)$$

for some $v \in U(t)$; here $w := (w^1, \ldots, w^m)$. This orientor field is supposed to have the upper semicontinuity property (K) [15] on $T \times \boldsymbol{R}^\nu$, i.e.,

$$Q(t, v) \;=\; \cap_{\epsilon > 0} cl \cup_{|v' - v| < \epsilon} Q(t, v') \quad \text{for every } (t, v) \in T \times \boldsymbol{R}^\nu.$$

Here "cl" denotes closure in $\boldsymbol{R} \times [-\infty, +\infty]^m$. Also, we suppose that there exists $\chi' : \boldsymbol{R}_+ \to [0, +\infty]$, nondecreasing, lower semicontinuous, and convex, such that $\chi'(0) = 0$,

$$\lim_{\xi \to \infty} \chi'(\xi)/\xi \;=\; +\infty,$$

and such that, for a certain integrable function $\phi_0 \in L_1^+(T)$,

$$g_0(t, v) \geq \chi'(|b(t, v)|) - \phi_0(t) \quad \text{on } M.$$

Furthermore, it is suppposed that, for every $\epsilon > 0$, there exists an integrable $\phi_\epsilon \in L_1^+(T)$ such that

$$g_i(t, v) + \epsilon\chi'(|b(t, v)|) \geq -\phi_\epsilon(t) \quad \text{on } M$$

for $i = 1, \ldots, m$. The *control cost integral functional* (for $i = 0$) and the ith constraint integral functional (for $i = 1, \ldots, m$) are defined by

$$I_{g_i}(\pi, u) := \int_{\Gamma(\pi)} g_i(t, u(t))\mu(dt).$$

Let $\alpha_1, \ldots, \alpha_m : \Pi \to \boldsymbol{R}$ be lower semicontinuous functionals. A control pair $(\pi, u) \in \mathcal{G}$ is called *feasible* if

$$I_{g_1}(\pi, u) + \alpha_1(\pi) \leq 0, \ldots, I_{g_m}(\pi, u) + \alpha_m(\pi) \leq 0.$$

For instance, in the original problem (B_0), we had $g_0(t,v) = (v^2-1)^2$, and there were no constraints of the above kind. Also, in that problem, we can choose $\chi'(\xi) := (\max(\xi^2-1,0))^2$, causing $g_0(t,v) \geq \chi'(|v|) = \chi'(|b(t,v)|)$.

Remark 2.1. A sufficient condition for the above property (K) of Q is as follows. Suppose that, for every t, the following hold:

(i) $\chi'(|b(t,\cdot)|)$ is inf-compact on $U(t)$,
(ii) $g_1(t,\cdot),\dots,g_m(t,\cdot)$ are lower semicontinuous on $U(t)$,
(iii) $b(t,\cdot)$ is continuous on $U(t)$,
(iv) $U(t)$ is closed in U.

Then Q has property (K). To demonstrate this, it is enough to show the following: If, for a sequence $(v_k) \in U(t)$ the limits $\bar{v} := \lim_k b(t,v_k)$ and $(\bar{w}^0,\bar{w}) := \lim_k (g_0(t,v_k),\dots,g_m(t,v_k))$ exist, respectively, in $\boldsymbol{R}^\nu$ and $\boldsymbol{R}\times[-\infty,+\infty]^m$, then there exists $\bar{v}\in U(t)$ with $\bar{v} = b(t,\bar{v})$ and $\bar{w}^j \geq g_j(t,\bar{v})$, $j=1,\dots,m$. Since $\bar{w}^0 < +\infty$ and $\chi'(|b(t,v_k)|) \leq g_0(t,v_k)+\phi_0(t)$, condition (i) implies that a subsequence of (v_k) converges to some $\bar{v}\in U(t)$; by (ii)–(iv), this limit easily solves the above. In view of the properties χ' already has, a particular case of the above situation is obtained when U is Euclidean and when $\lim_{|v|\to+\infty, v\in U(t)} |b(t,v)| = +\infty$ for every $t\in T$.

Let $\bar{g}: T\times\boldsymbol{R}^n \to (-\infty,+\infty]$ be $\mathcal{T}\times\mathcal{B}(\boldsymbol{R}^n)$-measurable such that, for every $t\in T$,

$$\bar{g}(t,\cdot) \text{ is lower semicontinuous and concave on } \boldsymbol{R}^n.$$

We suppose that there exist integrable $\bar{\phi}_1,\dots,\bar{\phi}_{\bar{d}},\bar{\psi}\in L_1^+(T)$ and nondecreasing functions $\bar{m}_1,\dots,\bar{m}_{\bar{d}}:\boldsymbol{R}_+\to\boldsymbol{R}_+$, such that $\bar{\chi}:\xi\mapsto\sum_{i=1}^{\bar{d}}\bar{m}_i(\xi)\int_T\bar{\phi}_i\,d\mu$ satisfies

$$\bar{g}(t,\eta)+\bar{\chi}(|\eta|) \;\geq\; -\bar{\psi}(t) \quad \text{on } T\times\boldsymbol{R}^n.$$

The following growth relationship between ξ' and $\bar{\chi}$, lower bounds of, respectively, g_0 and $\bar{g}$, is supposed to hold:

$$\lim_{\xi\to\infty}[\mu(T)\chi'(\xi)-\bar{\chi}(c_{\Phi,1}c_\kappa\mu(T)\xi+c_{\Phi,2})] \;=\; +\infty.$$

The *trajectory cost integral functional* is defined by

$$I_{\bar{g}}(\pi,y) := \int_{\Gamma(\pi)}\bar{g}(t,y(t))\mu(dt).$$

In (B_0) we encountered $\bar{g}(t,\eta) := -\eta^2$. So if we set there $\bar{m}_1(\eta) := \eta^2$ and $\bar{\phi}_1\equiv 1$, then the above conditions hold.

Let $s_1,\dots,s_d:\Pi\times\mathcal{C}(T;\boldsymbol{R}^n)\to\boldsymbol{R}$ be d *evaluation functionals*, which intervene in the final cost expression, as does the *indirect final cost functional* $e:\Pi\times\boldsymbol{R}^d\to(-\infty,+\infty]$. We suppose that

$$\begin{gathered} s_i(\pi,y) \text{ is continuous in } (\pi,y) \text{ and linear in } y,\\ e \text{ is lower semicontinuous on } \Pi\times\boldsymbol{R}^d,\\ e(\pi,\zeta)\geq -c_e \quad \text{on } \Pi\times\boldsymbol{R}^d \end{gathered}$$

for some constant c_e. With $s := (s_1,\dots,s_d)$, the *indirect final cost* of $(\pi,u)\in\mathcal{G}$ is given by $e(\pi,s(y_{\pi,u}))$.

In addition, we allow for a more direct evaluation of final costs: Let $\bar{e}:\Pi\times\mathcal{C}(T;\boldsymbol{R}^n)\to(-\infty,+\infty]$ be the *direct final cost functional*. It is supposed that

$$\bar{e}(\pi,y) \text{ is lower semicontinuous in } (\pi,y) \text{ and concave in } y$$

(in comparison with e above, concavity is additional) and that

$$\bar{e}(\pi, y) \geq -c_{\bar{e}} \quad \text{on } \Pi \times \mathcal{C}(T; \boldsymbol{R}^n)$$

for some constant $c_{\bar{e}}$. The *direct final cost* of $(\pi, u) \in \mathcal{G}$ is given by $\bar{e}(\pi, y_{\pi,u})$. In problem (B_0), we had $e \equiv 0$ and $\bar{e}(y) = -\exp(-\|y\|)$; it was already observed that this functional is concave.

The main existence result without convexity of this paper can now be stated; several applications can be found in §5.

THEOREM 2.2 (optimal control). *Consider the following problem* (B):

$$\text{minimize } I(\pi, u) := I_{g_0}(\pi, u) + I_{\bar{g}}(\pi, y_{\pi,u}) + e(\pi, s(y_{\pi,u})) + \bar{e}(\pi, y_{\pi,u})$$

over all admissible and feasible control pairs $(\pi, u) \in \mathcal{G}$. *Assume, next to the conditions stated in this section, that* $\inf(B) < +\infty$. *Then there exists an optimal control pair for* (B).

When the dynamical system is absent/trivial (i.e., $b \equiv 0$, $\bar{g} \equiv 0$, $e, \bar{e} \equiv 0$, and so on), the condition that T is compact and metric can be dispensed with. It then suffices to have an abstract nonatomic finite measure space $(T, \mathcal{T}, \mu)$. Observe how, for $b \equiv 0$, the property (K) for Q becomes equivalent to

$$Q(t, 0) \text{ is closed for every } t \in T,$$

and the lower bound for g_0 reads

$$g_0(t, v) \geq -\phi_0'(t) \quad \text{on } M$$

for some integrable $\phi_0' \in L_1(T)$ (indeed $\phi_0' := \phi_0 - \chi'(0)$).

THEOREM 2.3 (variational problem). *Consider the following problem* (B'):

$$\text{minimize } I_{g_m}(\pi, u) + \alpha_m(\pi)$$

over all pairs $(\pi, u) \in \mathcal{G}$ *such that* $I_{g_0}(\pi, u) \leq 0$ *and* $I_{g_i}(\pi, u) + \alpha_i(\pi) \leq 0$, $i = 1, \ldots, m-1$. (*The reversal in the roles of* g_0 *and* g_m, *while not necessary, increases the options for applications.*) *Assume, next to the conditions stated in this section* (*when specialized to the trivial dynamics case* $b \equiv 0$, $\bar{g} \equiv 0$, $e, \bar{e} \equiv 0$) *and the already-announced abstract nature of the nonatomic measure space* $(T, \mathcal{T}, \mu)$, *that* $\inf(B) < +\infty$. *Then there exists an optimal pair for* (B').

This result generalizes a well-known existence result for a variational problem in economics, first stated by Aumann and Perles [3], and later generalized by Berliocchi and Lasry [12], Artstein [2], and Balder [5]; cf. Corollary 5.4, below.

3. Fundamental extremum principle. This section, abstract in nature, can be read independently from the rest of this paper. Let E be a Hausdorff locally convex topological vector space and let C be a subset of E. A new, abstract extremum principle is now stated for an optimization problem, where a partly quasi-concave objective function is minimized over C under *finitely many* affine constraints. It guarantees the existence of a special optimal solution, viz., one that is a *convex combination of extreme points of* C.

THEOREM 3.1 (extremum principle). *Consider the following abstract optimization problem* (P):

$$\text{minimize } J(x) := f(x) + g(L(x))$$

over all $x \in C$ satisfying the constraints

$$b_0(x) \leq 0, b_1(x) \leq 0, \ldots, b_m(x) \leq 0.$$

Here, $b_0, b_1, \ldots, b_m : C \to (-\infty, +\infty]$ are functions such that

$$b_0 \text{ is inf-compact}^1 \text{ and affine,}$$
$$b_1, \ldots, b_m \text{ are lower semicontinuous and affine on } C_0,$$

where $C_0 := \{x \in C : b_0(x) \leq 0\}$. Also, C is supposed to be such that

$$C \text{ is a convex extremal subset of some compact convex subset of } E$$

(i.e., $\frac{1}{2}(x + x') \in C$ implies that $x, x' \in C$ for all x, x' in the compact convex subset), and $f : C \to (-\infty, +\infty]$ and $g : L(C) \to (-\infty, +\infty]$ are objective functions such that

$$f \text{ is lower semicontinuous and quasiconcave on } C_0,$$
$$g \text{ is lower semicontinuous on } L(C_0),$$

where $L : C \to \boldsymbol{R}^d$ is an operator such that

$$L \text{ is continuous and affine on } C_0.$$

Then (P) has an optimal solution that is a convex combination of at most $2m+4d+2$ extreme points of C, provided that it has at least one feasible solution. Moreover, if C_0 happens to be extremal in C, then $m + 2d + 1$ extreme points already suffice.

Remarkably, no compactness or even closedness (in E) is required for C itself. It should be stressed already that the extremality condition for C is *automatically* satisfied when C is the set of all probability measures on a suitable (but noncompact) topological space or, more generally, when C is the set of all Young measures on such a space. Indeed, the extremality condition reflects the extremal position that the set of Young measures holds with respect to its compactification; cf. §4. Similar results for probability measures are classical [17, III].

The classical extremum principle of Bauer (see [16, 25.9] for concave f and [19, 13.A-B] for quasi-concave f) is contained in Theorem 3.1.

COROLLARY 3.2 (Bauer extremum principle). *Consider the optimization problem*

$$\text{minimize } f(x)$$

over all $x \in K$. Here K and $f : K \to (-\infty, +\infty]$ are such that

$$K \text{ is a nonempty compact convex subset of } E,$$
$$f \text{ is lower semicontinuous and quasi-concave.}$$

Then the above optimization problem has an optimal solution that is an extreme point of K.

Proof. Apply Theorem 3.1 to the case where $C := K$ by setting g, $b_0 \equiv 0$ and m, $d := 0$. Then $C = C_0 = K$, so the additional clause of Theorem 3.1 applies. □

We now work in the converse direction and prove Theorem 3.1 by a combination of Corollary 3.2 and the following new characterization of the extreme points of an

[1] i.e., $\{x \in C : b_0(x) \leq \gamma\}$ is compact for every $\gamma \in R$.

affinely constrained subset of E, which extends certain well-known results of this type (e.g. [12, p. 145]). Such characterizations, with applications in the theory of moments, seem to have started with Rosenbloom [33]; cf. [34], [35] for linear-programming-oriented applications to optimal control. A very nice, general exposition can be found in [38].

PROPOSITION 3.3. *Under the conditions of Theorem* 3.1, *every extreme point of the feasible set for* (P) *is a convex combination of at most* $2m+2$ *extreme points of* C. *Moreover, if* C_0 *happens to be extremal in* C, *then* $m+1$ *extreme points of* C *already suffice.*

Proof. By the properties of b_0, the set C_0 is compact and convex. So by [38, Thm. 2.1, Ex. 2.1] every extreme point of the feasible set for (P) is a convex combination of at most $m+1$ extreme points of C_0 (here we use the fact that the feasible set is precisely the set of all $x \in C_0$ with $b_i(x) \leq 0$, $i = 1, \ldots, m$). Let $\hat{C}$ denote the compact convex set in which C is supposed to be a convex extremal subset. Now observe that the extension $\hat{b}_0 : \hat{C} \to (-\infty, +\infty]$ of b_0, obtained by setting $\hat{b}_0 \equiv +\infty$ on the convex set $\hat{C} \backslash C$, is affine. It is also inf-compact (hence lower semicontinuous) on $\hat{C}$, by the inf-compactness of b_0 on C (indeed, for every $\gamma \in \boldsymbol{R}$, the set $\{x \in \hat{C} : \hat{b}_0(x) \leq \gamma\}$, which coincides with $\{x \in C : b_0(x) \leq \gamma\}$, is compact). Since $C_0 = \{x \in \hat{C} : \hat{b}_0(x) \leq 0\}$, the result from [38] can be applied once more, giving that every extreme point of C_0 is a convex combination of at most two extreme points of $\hat{C}$. However, $C_0 \subset C$, and extremality of C in $\hat{C}$ cause the two extreme points in the latter combination to be extreme points of C itself. Under the additional clause, this last step can be omitted. □

Proof of Theorem 3.1. By compactness of C_0 and lower semicontinuity of the b_i's on C_0, it follows that the feasible set of (P) is compact. Also, the objective function J is easily seen to be lower semicontinuous on the feasible set. So, by the Weierstrass theorem, there exists a feasible $\bar{x}$ such that $J(x) \geq J(\bar{x})$ for all other feasible x. Consider the auxiliary optimization problem $(\bar{P})$, which runs as follows:

$$\text{minimize } f(x)$$

over the set K of all $x \in C$ satisfying both $b_0(x) \leq 0, b_1(x) \leq 0, \ldots, b_m(x) \leq 0$ and $L(x) = L(\bar{x})$. Define $b_{m+j}(x) := L_j(x) - L_j(\bar{x})$ and $b_{m+d+j} := -b_{m+j}$ for $j = 1, \ldots, d$. Here L_j denotes the jth component function of L. Then K is precisely the set of all $x \in C_0$ satisfying $b_i(x) \leq 0$, $i = 1, \ldots, m+2d$. All of the b_i's are lower semicontinuous and affine on C_0. In particular, K is a compact convex subset of C_0. By Corollary 3.2 [19, 13.A-B], there exists an optimal solution x_* for $(\bar{P})$ that is an extreme point of K. By the above, it is clear that the characterization of Proposition 3.3 applies to x_*. This gives x_*, precisely as stated. Finally, note that x_* is also an optimal solution of (P), since optimality of $\bar{x}$ for (P) and optimality of x_* for $(\bar{P})$ give, when combined,

$$J(x) \;\geq\; J(\bar{x}) = f(\bar{x}) + g(L(x_*)) \;\geq\; J(x_*)$$

for every x that is feasible for (P).

The following example illustrates the role of the unusual extremality condition for C. Here we see in essence what was already announced immediately following Theorem 3.1: The extremality condition always holds when we take for C the set of all probability measures on a suitable topological space. See [17, III] for topological background material.

Example 3.4. Consider $C := M_1^+((0,1])$, the set of all Borel probability measures on the interval $(0,1]$. Canonically, a probability measure on $(0,1]$ can also be regarded

as a probability measure on the larger set $[0,1]$. Hence, C is a convex subset of $\hat{C} := M_1^+([0,1])$, and it is easy to see that C is extremal in $\hat{C}$. In turn, $\hat{C}$ is a convex subset of the set $E := M([0,1])$ of all bounded signed measures on $[0,1]$. Equip E with the usual narrow topology (this coincides with the weak star—alias vague—topology $\sigma(E, C([0,1]))$. This topology makes E Hausdorff [17, III] and locally convex. By [17, III.60], the set $\hat{C}$ is compact in E. Let $b_0 : C \to (-\infty, +\infty]$ be given by $b_0(x) := \int_{(0,1]} v^{-1}\, x(dv) - 2$. By [17, III.55], b_0 is certainly lower semicontinuous for the narrow topology on C (which is indeed the relative topology induced by the narrow topology on E [17, III]). Furthermore, for any $x \in C$, we see that $b_0(x) \leq \gamma < +\infty$ implies that $x([\epsilon, 1]) \geq 1 - \epsilon(2+\gamma)$ for any $\epsilon > 0$. Therefore, it follows by Prohorov's theorem [17, III.59] that b_0 is inf-compact on C. Define $\bar{x} := \frac{1}{2}\epsilon_{1/3} + \frac{1}{2}\epsilon_1$, where ϵ_ξ denotes the Dirac (probability) measure concentrated in the point ξ, $\xi \in (0,1]$. It is easy to verify that $\bar{x}$ is an extreme point of $C_0 = \{x \in M_1^+((0,1]) : \int_{(0,1]} v^{-1}\, x(dv) \leq 2\}$. The form of $\bar{x}$ is as predicted by Proposition 3.3 (with $m = 0$), for it is well known [16, Prob. 25.1] that the Dirac measures on $(0,1]$ form precisely the extreme points of C. Finally, note that, for the artifical problem of minimizing $-b_0$ over C_0, $\bar{x}$ is also of the form predicted by Theorem 3.1.

We can now somewhat explain what went on in Phase 2 in §1 (more details are forthcoming): In (B_0), we had $m = 0$, the only constraint being $b_0(x) \leq 0$. Also, $f := J$, $g \equiv 0$, and $d = 0$. This would give an optimal solution x_*, which is a convex combination of at most two extreme points of the set C of all transition probabilities from $[0,1]$ into $M_1^+(\boldsymbol{R})$. By analogy with the previous example, it would seem obvious that the extreme points of C are formed by the Dirac transition probabilities, i.e., the transition probabilities of the form ϵ_u, hence the expression for x_* given in §1.

4. Proof of the results in §2. We follow the lines of the convexification-deconvexification scheme. These were explained in §1, but they were only worked out there for the simple problem (B_0). An additional feature of our present march toward the optimal control pair is as follows: Phase 1 is concluded by applying the Weierstrass theorem for the purpose of establishing the optimal control *parameter*. The optimal control *function* is then determined through Phases 2 and 3. Another complication is formed by the great degree of generality incorporated in property (K) for the orientor field Q. For this reason, if we are prepared to accept the only marginally more stringent conditions (i)–(iv) of Remark 2.1, we will find Proposition 4.9 a useful substitute for Theorem 2.2. Actually, the convexification-deconvexification scheme leads straight to that proposition; thereafter, deparametrization and measurable regularization ideas are used to deduce Theorem 2.2 from Proposition 4.9.

4.1. Phase 1: Convexification via relaxation. Let us formulate a relaxed version (B_{rel}) of problem (B). The present choice of relaxation only makes sense under a working hypothesis, which we make from here until Proposition 4.9.

WORKING HYPOTHESIS. *The sufficient conditions* (i)–(iv) *for property* (K), *as stated in Remark* 2.1, *hold.*

Let G be the set of all relaxed control pairs (π, x), with $\pi \in \Pi$ and $x : \Gamma(\pi) \to M_1^+(U)$ a transition probability such that $x(t)(U(t)) = 1$ almost everywhere. Let $r : G \to (-\infty, +\infty]$ be defined by

$$r(\pi, x) := \int_{\Gamma(\pi)} \left[\int_{U(t)} g_0(t, v)\, x(t)(dv) \right] \mu(dt).$$

The relaxed trajectory can easily be stated; it is given by

$$y_{\pi,x} := \Phi(\pi, A(\pi, x)),$$

where

$$A(\pi,x)(t) := \int_{\Gamma(\pi)} \kappa(t,\tau) \left[\int_{U(t)} b(\tau,v)\, x(\tau)(dv) \right] \mu(d\tau).$$

Of course, this presupposes that (π, x) is *admissible*, i.e.,

$$a_{m+1}(\pi,x) := \int_{\Gamma(\pi)} \left[\int_{U(t)} |b(t,v)|\, x(t)(dv) \right] \mu(dt) < +\infty.$$

Just as for the original control functions, such admissibility implies that the trajectory $y_{\pi,x}$ belongs to $C(T; \boldsymbol{R}^n)$ (see Lemma 4.3, below). The constraints are relaxed as follows: Define $a_i : G \to (-\infty, +\infty]$ by

$$a_i(\pi,x) := \int_{\Gamma(\pi)} \left[\int_{U(t)} g_i(t,v)\, x(t)(dv) \right] \mu(dt);$$

then $(\pi, x) \in G$ is called *feasible* if $a_i(\pi, x) + \alpha_i(\pi) \leq 0$ for $i = 1, \ldots, m$. Let $\bar{r} : \Pi \times C(T; \boldsymbol{R}^n) \to (-\infty, +\infty]$ be defined by

$$\bar{r}(\pi,y) := \int_{\Gamma(\pi)} \bar{g}(t, y(t)) \mu(dt) + \bar{e}(\pi, y).$$

Clearly, this functional is well defined (see also §4) and concave in the y-variable. The relaxed version (B_{rel}) of (B) is defined as follows:

$$\text{minimize } J(\pi,x) := r(\pi,x) + \bar{r}(\pi, y_{\pi,x}) + e(\pi, s(y_{\pi,x}))$$

over all admissible and feasible pairs $(\pi, x) \in G$.

Next, we prepare (B_{rel}) for a treatment by means of Theorem 3.1. In the course of this, we present a brief summary of Young measure theory. By the presence of the variable time domains, this expands slightly on the more standard expositions [6], [36]. Let R stand for the set of all (equivalence classes of) transition *sub*probabilities $x : T \to M^+_{\leq 1}(U)$. Here $M^+_{\leq 1}(U)$ denotes the set of all Borel subprobabilities (i.e., measures with mass at most 1) on U. For every $(\pi, x) \in G$, the control part x is by definition a transition probability from $\Gamma(\pi)$ into $M^+_1(U)$. Note that x can be extended so as to belong to R, simply by setting $x(t)$ equal to the null measure on U for almost every t in $T \backslash \Gamma(\pi)$. We call this *extension by nullity* of Young measures. This device, to make the relaxed control functions—in particular, the relaxations of the original control functions—"null and void" outside the time domain over which control is exercised, can already be found in McShane's work [26], but seems not to have received the attention that it deserves (for instance, the brief description in [37, VI.5.2] diverges, and is considerably less general). The fact that R consists of (equivalence classes of) transition *sub*probabilities causes no essential differences with the usual treatment of Young measure topology: It is well known [7], [6] that R can be identified with a subset in the closed unit ball $\hat{R}$ of the space $E := L_\infty(T, \mathcal{T}, \mu; M(\hat{U}))$ of (equivalence classes of) essentially bounded, scalarly (i.e., with respect to $C(\hat{U})$, the set of all continuous

functions on $\hat{U}$) measurable functions from T into $M(\hat{U})$. Here $\hat{U}$ denotes the Hilbert cube compactification of U (for U a Suslin (Lusin) space, this makes U a universally measurable [Borel measurable] subset of $\hat{U}$; cf. [17]). Also, $M(\hat{U})$ denotes the set of all bounded signed Borel measures on $\hat{U}$. By the Alaoglu–Bourbaki theorem, $\hat{R}$ is compact for the weak star topology on E. Define the functional $a_0 : G \to [0, +\infty]$ by

$$a_0(\pi, x) := \int_{\Gamma(\pi)} \left[\int_{U(t)} \chi'(|b(t,v)|)\, x(t)(dv) \right] \mu(dt).$$

LEMMA 4.1. (i) a_0 *is* inf-*compact on* $G \subset \Pi \times R$.

(ii) r *and* a_{m+1} *are lower semicontinuous on* G.

(iii) *For every* $\gamma \in \boldsymbol{R}$, $a_1, \ldots, a_m$ *are lower semicontinuous on* $\{(\pi, x) \in G : a_0(\pi, x) \leq \gamma\}$.

Proof. (i) Observe that, by L_1-norm separability of E's predual $L_1(T, \mathcal{T}, \mu; C(\hat{U}))$, the weak star topology on $\hat{R}$ is metrizable [19, 12.F]. So let us now prove for an arbitrary $\gamma \in \boldsymbol{R}$ that the set G', consisting of all $(\pi, x) \in G$ for which $a_0(\pi, x) \leq \gamma$, is sequentially compact. Given any sequence $((\pi_k, x_k))$ in G', we may suppose without loss of generality that $(\pi_0, x_0) := \lim_k (\pi_k, x_k)$ exists in the englobing compact set $\Pi \times \hat{R}$. Then $\mu(\Gamma(\pi_k) \Delta \Gamma(\pi_0)) \to 0$, by the continuity property of Γ. Define $g \in L_1 := L_1(T, \mathcal{T}, \mu; C(\hat{U}))$ by $g(t,v) := 1$ if $t \in \Gamma(\pi_0)$ and $g(t,v) := 0$, otherwise. Then the weak star convergence $x_k \to x_0$ implies that $\langle g, x_k \rangle \to \langle g, x_0 \rangle$, i.e., $\mu(\Gamma(\pi_0)) = \lim_k \mu(\Gamma(\pi_0) \cap \Gamma(\pi_k)) = \int_{\Gamma(\pi_0)} x_0(t)(\hat{U})\, \mu(dt)$. Since $x_0(t)(\hat{U}) \leq 1$ for almost every t in T, this implies that $x_0(t) \in M_1^+(\hat{U})$ for almost every t in $\Gamma(\pi_0)$. Similar reasoning for the function $1 - g$ leads to $x_0(t) =$ null measure for almost every t in $T \backslash \Gamma(\pi_0)$. Thus, it follows that $\int_{\Gamma(\pi_k)} [\int_{\hat{U}} g(t,v)\, x_k(t)(dv)] \mu(dt) \to \int_{\Gamma(\pi_0)} [\int_{\hat{U}} g(t,v)\, x_0(t)(dv)] \mu(dt)$ for every $g \in L_1$. By well-known arguments, this implies $\liminf_k \int_{\Gamma(\pi_k)} [\int_{\hat{U}} g(t,v)\, x_k(t)(dv)] \mu(dt) \geq \int_{\Gamma(\pi_0)} [\int_{\hat{U}} g(t,v)\, x_0(t)(dv)] \mu(dt)$ for every measurable $g : T \times \hat{U} \to [0, +\infty]$ such that $g(t, \cdot)$ is lower semicontinuous (i.e., inf-compact) on the compact set $\hat{U}$ (e.g., see [6, Lemma A.2]). In particular, the function g has those properties, where g is defined by $g(t,v) := \chi'(|b(t,v)|)$ if $v \in U(t)$ and by $g(t,v) := +\infty$ if $v \in \hat{U} \backslash U(t)$ (note the role played by our working hypothesis in connection with Remark 2.1(i)). This gives $\gamma \geq \liminf_k a_0(\pi_k, x_k) \geq \int_{\Gamma(\pi_0)} [\int_{\hat{U}} g(t,v)\, x_0(t)(dv)] \mu(dt)$. It follows that $\int_{\hat{U}} g(t,v)\, x_0(t)(dv) < +\infty$ for almost every t in $\Gamma(\pi_0)$, which in turn forces $x_0(t)(U(t)) = 1$ for almost every t in $\Gamma(\pi_0)$. We conclude that (π_0, x_0) belongs to G'.

(ii) Since $g_0(t,v) \geq -\phi_0(t)$ and $|b(t,v)| \geq 0$ on M, lower semicontinuity of the two integral functionals follows by [6, Lemma A.2, Thm. 3.3], in view of the fact that both integrands are measurable in (t,v) and lower semicontinuous in v (for g_0, this holds by the working hypothesis).

(iii) By standard arguments [6, Lemma A.2, Thm. 3.3], the hypotheses give that $a_i + \epsilon a_0$ is lower semicontinuous on G for every $\epsilon > 0$ (note the role played by the working hypothesis in connection with Remark 2.1(ii)). For $a_0(\pi, x) \leq \gamma$, we obviously have

$$a_i(\pi, x) \;=\; \sup_{\epsilon > 0} [a_i(\pi, x) + \epsilon a_0(\pi, x) - \epsilon \gamma],$$

so the result follows from the fact that the pointwise supremum of lower semicontinuous functionals is also lower semicontinuous. □

LEMMA 4.2. (i) $a_0(\pi,x) \geq \chi(a_{m+1}(\pi,x))$ *on* G, *with* $\chi(\xi) := \mu(T)\chi'(\xi/\mu(T))$.
(ii) $\bar{r}(\pi,y) \geq -c_{\bar{e}} - \bar{\chi}(\|y\|) - \int_T \bar{\psi} d\mu$ *on* $\Pi \times \mathcal{C}(T;\boldsymbol{R}^n)$.
(iii) $\bar{r}$ *is lower semicontinuous on* $\Pi \times \mathcal{C}(T;\boldsymbol{R}^n)$.

Proof. (i) By Jensen's inequality, we obtain, for every $t \in \Gamma(\pi)$,

$$\int_{U(t)} \chi'(|b(t,v)|)\, x(t)(dv) \;\geq\; \chi'\left(\int_{U(t)} |b(t,v)|\, x(t)(dv)\right),$$

thanks to convexity of χ'. By $\chi'(0) = 0$ and $x(t) =$ null measure for $t \notin \Gamma(\pi)$ (extension by nullity), one more application of Jensen's inequality gives the desired inequality.

(ii) By monotonicity of the $\bar{m}_i$, the definition of the supremum norm entails $\bar{m}_i(y(t)) \leq \bar{m}_i(\|y\|)$ for every t, $i = 1,\ldots,\bar{d}$. The desired inequality then follows with ease.

(iii) By Fatou's lemma (applicable by the lower bound for $\bar{r}$ indicated in the above proof of (ii)), the result follows easily from the lower semicontinuity of $\bar{g}(t,\eta)$ in η, the continuity property of $\Gamma(\pi)$ in π, and the lower semicontinuity of $\bar{e}$. □

LEMMA 4.3. (i) *For every admissible pair* $(\pi,x) \in G$, $A(\pi,x)$ *belongs to* $\mathcal{C}(T;\boldsymbol{R}^n)$.
(ii) *For every* $\gamma \in \boldsymbol{R}$, *the function* A *is continuous on* $\{(\pi,x) \in G : a_0(\pi,x) \leq \gamma\}$.

Proof. (i) The proof is contained in the equicontinuity part of the next proof.

(ii) Denote the subset $\{a_0 \leq \gamma\}$ of G by G'. First, let us check that, for every fixed $t \in T$, the mapping $(\pi,x) \mapsto A(\pi,x)(t)$ is continuous from G' into $\boldsymbol{R}^n$. Observe that, for the jth component,

$$[A(\pi,x)(t)]^j \;=\; \int_{\Gamma(\pi)} \left[\int_{U(t)} g_{t,j}(t,v)\, x(t)(dv)\right] \mu(dt),$$

where $g_{t,j} : T \times U \to \boldsymbol{R}$ is defined by $g_{t,j}(\tau,v) := [\kappa(t,\tau)b(\tau,v)]^j$, $j = 1,\ldots,n$. This integrand is measurable in (τ,v) and continuous in v, so both $g_{t,j,\epsilon}(\tau,v) := g_{t,j}(\tau,v) + \epsilon\chi'(|b(\tau,v)|)$ and $g'_{t,j,\epsilon}(\tau,v) := -g_{t,j}(\tau,v) + \epsilon\chi'(|b(\tau,v)|)$ are measurable in (τ,v), lower semicontinuous in v, and bounded from below by an integrable function (in fact, by a constant). The latter follows directly from the inequality $|g_{t,j}(\tau,v)| \leq c_\kappa |b(\tau,v)|$ and the property $\lim_{\xi\to\infty} \chi'(\xi)/\xi = +\infty$. Reasoning similarly to that in the proof of Lemma 4.1(ii), it follows that the integral functional appearing in the right-hand side of the above expression for $[A(\pi,x)(t)]^j$ is continuous in (π,x) on G'.

The argument is now finished by showing that $\{A(\pi,x) : (\pi,x) \in G'\}$ forms an equicontinuous collection on the compact set T. (In particular, taking for G' the singleton consisting of (π,x), with $\gamma := a_{m+1}(\pi,x)$, settles the proof of part (i).) For t, $t' \in T$, it follows by σ-hereditarity of κ that

$$|A(\pi,x)(t) - A(\pi,x)(t')| \;\leq\; \int_{T_{\sigma(t)}\Delta T_{\sigma(t')}} |\kappa(t,\tau) - \kappa(t',\tau)||b(t,v)|x(\tau)(dv)\mu(d\tau).$$

Let $\epsilon > 0$ be arbitrary; then $-|b(\tau,v)| + \epsilon\chi'(|b(\tau,v)|)$ is bounded from below by some nonpositive constant $-c_\epsilon$. So

$$|A(\pi,x)(t) - A(\pi,x)(t')| \;\leq\; 2c_\kappa[c_\epsilon \mu(T_{\sigma(t)}\Delta T_{\sigma(t')}) + \epsilon\gamma],$$

uniformly in $(\pi,x) \in G'$. The measure of the symmetric difference converges to zero as $t' \to t$ (by the continuity of both σ and $\alpha \mapsto \mu(T_\alpha)$), so equicontinuity has

been proved. By the Arzela–Ascoli theorem [37, I.5.4], it follows that the mapping $(\pi, x) \mapsto A(\pi, x)$ is also continuous. □

Phase 1 concludes with an application of the Weierstrass theorem, which is essentially only used to determine an optimal control parameter (this explains why it was not needed for (B_0) in §1).

LEMMA 4.4. *There exists an optimal solution for the relaxed optimal control problem* (B_{rel}).

Proof. Clearly, $\inf(B_{\mathrm{rel}}) > -\infty$ and $\inf(B_{\mathrm{rel}}) \leq \inf(B) < +\infty$ (the latter by hypothesis). Let F be the nonempty set of all feasible and admissible relaxed control pairs $(\pi, x) \in G$ satisfying $J(\pi, x) \leq \inf(B_{\mathrm{rel}}) + 1$. Clearly, $\inf(B_{\mathrm{rel}}) = \inf_{(\pi,x)\in F} J(\pi, x)$. By Lemma 4.2, every pair $(\pi, x) \in F$ satisfies

$$\mu(T)\chi'(a_{m+1}(\pi, x)/\mu(T))\bar{\chi}(c_{\Phi,1}c_\kappa a_{m+1}(\pi, x) + c_{\Phi,2}) \leq \beta' + \int_T \phi_0 \, d\mu,$$

for $\beta' := \inf(B_{\mathrm{rel}}) + 1 + \int_T \bar{\psi} \, d\mu + c_e + c_{\bar{e}}$. By the limit relationship between χ' and $\bar{\chi}$, it follows that there exists a constant $\alpha \in \boldsymbol{R}$ such that $a_{m+1}(\pi, x) \leq \alpha$ for all $(\pi, x) \in F$. In turn, this implies that

$$r(\pi, x) - \bar{\chi}(c_{\Phi,1}c_\kappa\alpha + c_{\Phi,2}) \leq \beta',$$

by monotonicity of $\bar{\chi}$. Hence, it follows that $r(\pi, x) \leq \beta' + \bar{\chi}(c_{\Phi,1}c_\kappa\alpha + c_{\Phi,2})$ for all $(\pi, x) \in F$. Evidently, this implies that $a_0(\pi, x) \leq \beta := \beta' + \bar{\chi}(c_{\Phi,1}c_\kappa\alpha + c_{\Phi,2}) + \int_T \phi_0 \, d\mu$. Therefore, the constraints

$$a_0(\pi, x) \leq \beta, \qquad a_{m+1}(\pi, x) \leq \alpha$$

hold *implicitly* and can be added to the definition of F without altering (B_{rel}). Thus, $\inf(B_{\mathrm{rel}}) = \inf(Q)$, where (Q) is the problem of minimizing $J(\pi, x)$ over the set F' of all pairs $(\pi, x) \in G$ for which $a_0(\pi, x) \leq \alpha$, $a_i(\pi, x) \leq 0$, $i = 1, \ldots, m$, and $a_{m+1}(\pi, x) \leq \beta$. By Lemmas 4.1–4.3 (with β now taking the place of the γ used there), the feasible set F' of this problem is compact, and its objective functional J is lower semicontinuous. So there exists an optimal relaxed pair for (Q) by the Weierstrass theorem; by the above, this is also optimal for (B_{rel}). □

Remark 4.5. From the proof of Lemma 4.4, it is evident that the coercivity conditions (i.e., the lower bounds for g_0, $\bar{g}$ and the limit property involving χ, $\bar{\chi}$) *only* serve to guarantee the existence of $\alpha, \beta \in \boldsymbol{R}$ such that

$$a_0(\pi, x) \leq \beta, \quad a_{m+1}(\pi, x) \leq \alpha \quad \text{for all } (\pi, x) \in F.$$

In other words, a simpler and more limited alternative would have been to introduce from the beginning a nontrivial constraint $a_0'(\pi, x) \leq 0$ for (B), with a_0' such that (i) a_0' is inf-compact on G (this takes care of Lemma 4.1(i) by hypothesis) and (ii) $a_0'(\pi, \cdot)$ is affine for every $\pi \in \Pi$. The coercivity conditions could then have been omitted.

4.2. Phase 2: Application of the extremum principle. This phase consists simply of an application of Theorem 3.1.

PROPOSITION 4.6. *There exists an optimal solution* $(\pi_*, x_*) \in G$ *for* (B_{rel}) *such that* x_* *is a convex combination of at most* $N = 2m + 4d + 4$ *relaxations of functions* u_k *with* $(\pi_*, u_k) \in \mathcal{G}$, *i.e.,*

$$x_*(t) = \sum_{k=1}^{N} \lambda_k \epsilon_{u_k}(t) \quad \textit{for a.e. } t \textit{ in } \Gamma(\pi_*),$$

for certain $\lambda_1, \ldots, \lambda_N \in [0,1]$, *satisfying* $\sum_{k=1}^N \lambda_k = 1$ (*that is,* x_* *has Minkowski form*).

Proof. In the proof of Lemma 4.4, it was demonstrated that $\inf(Q) = \inf(B_{\mathrm{rel}})$ and that (Q) has an optimal solution, say $(\pi_*, \bar{x}) \in F' \subset G$. Consider the auxiliary optimization problem (Q_{π_*}): Minimize $J(\pi_*, x)$ over all transition probabilities $x : \Gamma(\pi_*) \to M_1^+(U)$ such that $(\pi_*, x) \in F'$. Clearly, $\inf(Q_{\pi_*}) = J(\pi_*, \bar{x}) = \inf(B_{\mathrm{rel}})$. Writing F', we see that (Q_{π_*}) is precisely of the form addressed by Theorem 3.1. For this, we take C to be the set of all transition probabilities $x : \Gamma(\pi_*) \to M_1^+(U)$ such that $(\pi_*, x) \in G$ and we set $f(x) := r(\pi_*, x) + \bar{r}(\pi_*, y_{\pi_*, x})$, $L(x) := s(\pi_*, y_{\pi_*, x})$, and $g(\zeta) := e(\pi_*, \zeta)$). The required concavity and affinity are seen to hold by elementary considerations—cf. §1. The required extremality property of C is seen to hold by observing that the set $\hat{C}$ of all transition probabilities $x : \Gamma(\pi_*) \to M_1^+(\hat{U})$ is compact (this follows from the compactness of $\hat{R}$ observed in the proof of Lemma 4.1) and convex and contains C as an extremal subset. Therefore, the extremum principle applies. Since $J(\pi_*, x_*)$ attains $\inf(Q_{\pi_*})$, the above gives also $J(\pi_*, x_*) = \inf(B_{\mathrm{rel}})$. Finally, it remains to observe that, by [13, IV.15] (see also [23]), the extreme points of C are precisely those (equivalence classes of) transition probabilities $x : \Gamma(\pi_*) \to M_1^+(U)$ that are of the form $x = \epsilon_u$ for some measurable $u : \Gamma(\pi_*) \to U$ with $u(t) \in U(t)$ almost everywhere □

4.3. Phase 3: Deconvexification. The following proposition provides the key to this phase.

LEMMA 4.7. *Let* (π_*, x_*) *be as in Proposition* 4.6. *There exist* $d+1$ *dual elements* $y_1^*, \ldots, y_{d+1}^*$ *in* $(\mathcal{C}(T; \boldsymbol{R}^n))^*$ *such that the following holds true: For any measurable function* $u : \Gamma(\pi_*) \to U$, *with* $u(t) \in U(t)$ *almost everywhere, the system* (A) *of* $m + d + 3$ *equations*

$$I_{g_i}(\pi_*, u) = \sum_k \lambda_k I_{g_i}(\pi_*, u_k), \qquad i = 0, \ldots, m+1,$$

$$\langle \mathcal{A}(\pi_*, u), y_j^* \rangle = \sum_k \lambda_k \langle \mathcal{A}(\pi_*, u_k), y_j^* \rangle, \qquad j = 1, \ldots, d+1$$

implies that (π_*, u) *is an optimal pair for* (B).

Proof. We already observed that $\bar{r}(\pi_*, \cdot)$ is concave. By Lemma 4.2(ii), the convex functional $-\bar{r}(\pi_*, \cdot)$ is locally bounded from above at every point of $\mathcal{C}(T; \boldsymbol{R}^n)$, including the point y_{x_*, π_*} (here the nature of the supremum norm and the monotonicity of $\bar{\chi}$ are used). By [18, I.2.4, I.5.2], this implies that there exists a subgradient $y^* \in (\mathcal{C}(T; \boldsymbol{R}^n))^*$ for $-\bar{r}(\pi_*, \cdot)$ in the point y_{x_*, π_*}, i.e.,

$$\bar{r}(\pi_*, y_{x_*, \pi_*}) - \bar{r}(\pi_*, y) \geq \langle y - y_{x_*, \pi_*}, y^* \rangle \quad \text{for all } y \in \mathcal{C}(T; \boldsymbol{R}^n).$$

Define y_{d+1}^* to be the image of y^* under the adjoint $\Lambda^* : (\mathcal{C}(T; \boldsymbol{R}^n))^* \to (\mathcal{C}(T; \boldsymbol{R}^n))^*$ of the continuous linear mapping $\Lambda : y \mapsto \Phi(\pi_*, y) - \Phi(\pi_*, 0)$. Also, our assumptions for (B) imply that each evaluation map s_i belongs to the dual space $(\mathcal{C}(T; \boldsymbol{R}^n))^*$; let $y_j^* := \Lambda^*(s_j)$, $j = 1, \ldots, d$.

Now take any u as stated, satisfying (A). The first equation $(i = 0)$ implies that (π_*, u) and (π_*, x_*) have the same control cost (i.e., $r(\pi_*, \epsilon_u) = r(\pi_*, x_*)$). The equations for $i = 1$ to $i = m+1$ imply that $a_i(\pi_*, \epsilon_u) = a_i(\pi_*, x_*)$. Hence, by the fact that (π_*, x_*) is admissible and feasible for (B_{rel}), (π_*, u) is admissible and feasible for (B). The final equation $(j = d+1)$ implies that $\bar{r}(\pi_*, y_{\pi_*, x_*}) - \bar{r}(\pi_*, y_{\pi_*, u}) \geq \langle y_{d+1}^*, A(\pi_*, x_*) - A(\pi_*, \epsilon_u) \rangle = 0$, by the above subgradient property of y^*. The

remaining equations (for $j = 1$ to $j = d$) imply that $s(\pi_*, y_{\pi_*,x_*}) = s(\pi_*, y_{\pi_*,u})$. Adding all four cost terms shows that $I(\pi_*, u) \le J(\pi_*, x_*) = \inf(B_{\text{rel}}) \le \inf(B)$. So (π_*, u) is optimal for (B). □

LEMMA 4.8. *There exists a measurable function $u_* : \Gamma(\pi_*) \to U$, with $u_*(t) \in U(t)$ almost everywhere, which solves the system (A) of Lemma 4.7.*

Proof. Each of the $y_j^* \in (C(T; \boldsymbol{R}^n))^*$ has a Riesz representation [37, I.5.8]; that is, there exists a bounded signed Borel vector measure $\theta_j := (\theta_j^1, \dots, \theta_j^n)$ on T such that $\langle y, y_j^* \rangle = \sum_{p=1}^n \int_T y^p(t)\, \theta_j^p(dt)$ for all $y := (y^1, \dots, y^n) \in C(T; \boldsymbol{R}^n)$. By Fubini's theorem

$$\sum_k \lambda_k \langle \mathcal{A}(\pi_*, u_k), y_j^* \rangle \;=\; \sum_k \lambda_k \int_{\Gamma(\pi_*)} \tilde{b}_{j,k}(\tau)\mu(d\tau),$$

where $\tilde{b}_{j,k}(\tau) := \sum_{p,p'} \int_T \kappa_{p,p'}(t,\tau) b_j^{p'}(\tau, u_k(\tau)) \theta_j^p(dt)$ defines an integrable function, $j = 1, \dots, d+1$, $k = 1, \dots, N$. More directly, the remaining right-hand sides in (A) can be written as

$$\sum_k \lambda_k I_{g_i}(\pi_*, u_k) \;=\; \sum_k \lambda_k \int_{\Gamma(\pi_*)} g_{i,k}(\tau)\mu(d\tau),$$

with integrable functions $g_{i,k}(\tau) := g_i(\tau, u_k(\tau))$, $i = 0, 1, \dots, m+1$, $k = 1, \dots, N$. Since μ is nonatomic, it follows by Lyapunov's theorem that there exists a measurable partition $C_1, \dots, C_N$ of $\Gamma(\pi_*)$ such that $\nu(C_k) = \lambda_k \nu(\Gamma_{\pi_*})$ for all k. Here ν denotes the vector measure with component measures having, respectively, densities $\tilde{b}_{j,k}$ and $g_{i,k}$ with respect to μ. Then, however, a solution u_* of (A) is obtained by defining $u_* : \Gamma(\pi_*) \to U$ as follows: For $t \in C_k$, define $u_*(t) := u_k(t)$, $k = 1, \dots, N$. □

PROPOSITION 4.9. *Under the working hypothesis in connection with Remark 2.1, there exists an optimal solution for (B).*

Proof. By Lemmas 4.7 and 4.8, the pair (π_*, u_*) of Lemma 4.8 is an optimal solution for (B). □

This finishes the application of the convexification-deconvexification scheme to problem (B). The proof of Theorem 2.2 would now be finished if it were not for the fact that we used the conditions of Remark 2.1, instead of our original property (K) required for the orientor field (this is where Theorem 2.2 differs from Proposition 4.9, just proved). So it remains to see how the original situation can be reduced to the one that we considered until now. We do so by relying on the device of *deparametrization* via measurable regularization, as introduced in [6].

To begin, let us reformulate (B) in the case where the integrands $g_0, g_1, \dots, g_m$ would have happened to be nonmeasurable. In that case,

$$I_{g_i}^*(\pi, u) := \inf \left\{ \int_{\Gamma(\pi)} \phi \, d\mu : \phi \in L_1(T), \phi(t) \ge g_i(t, u(t)) \quad \text{a.e.} \right\}$$

defines an *outer integral functional* (which coincides with $I_{g_i}(\pi, u)$ if g_i is measurable); e.g., see [8, App. A]. In that imaginary nonmeasurability case, we could define the following optimal control problem (B^*):

$$\text{minimize } I^*(\pi, u) := I_{g_0}^*(\pi, u) + I_{\bar{g}}(\pi, y_{\pi,u}) + e(\pi, s(y_{\pi,u})) + \bar{e}(\pi, y_{\pi,u})$$

over all admissible pairs $(\pi, u) \in \mathcal{G}$ satisfying the revised feasibility constraints $I_{g_i}^*(\pi, u) + \alpha_i(\pi) \le 0$, $i = 1, \dots, m$. The trajectories $y_{\pi,u}$ are exactly defined as in §2,

with no alterations needed, since *only* the integrands g_i are subjected to our present (imaginary) nonmeasurability investigation. The following result extends Proposition 4.9 to the nonmeasurable extension (B^*) of (B).

PROPOSITION 4.10. *Under the conditions of Theorem 2.2 and the working hypothesis, but possibly without the measurability conditions for the integrands $g_0, g_1, \ldots, g_m$, there exists an optimal control pair for (B^*).*

Proof. By the measurable regularization ideas introduced in [6, A.5], the following is true: For each i, there exists a $\mathcal{T} \times \mathcal{B}(U)$-measurable function $\tilde{g}_i : M \to (-\infty, +\infty]$ such that $\tilde{g}_i(t, \cdot)$ is lower semicontinuous on $U(t)$ for almost every t, and the following properties hold: (i) $\tilde{g}_i(t, \cdot) \geq g_i(t, \cdot)$ for almost every t, (ii) $I_{\tilde{g}_i}(\pi, u) = I^*_{g_i}(\pi, u)$ for all $\pi \in \Pi$ and all measurable $u : \Gamma(\pi) \to U$ with $u(t) \in U(t)$ almost everywhere (to see that this follows by [6, A.5], extend each g_i by setting $g_i \equiv +\infty$ on $(T \times U) \backslash M$). Define the auxiliary control problem $(\tilde{B})$ just as (B) in §2, but with the integrands g_i replaced by $\tilde{g}_i$. By property (i) of the measurable regularizations, it can be seen that all conditions used until now for the g_i are transferred to the $\tilde{g}_i$ (the essential inequalities all point in the right way), including those of the working hypothesis. Proposition 4.9 therefore applies to the auxiliary problem $(\tilde{B})$, which must have an optimal control pair (π_*, u_*). By property (ii) of the measurable regularizations, however, it follows that problems $(\tilde{B})$ and (B^*) are the same (the objective and constraint integral functionals coincide). Hence, (π_*, u_*) is also an optimal solution for (B^*). □

Let us now introduce an optimal control problem (B_{red}), which is an equivalent *reduction* [32] of the original problem (B), but which has the nonmeasurability characteristics of (B^*). We do this using the device of deparametrization, following ideas of Cesari and Rockafellar [15], [32], respectively, to which the measurable regularization idea has been added [6].

Let $\ell_0 : T \times \boldsymbol{R}^\nu \times [-\infty, +\infty]^m \to (-\infty, +\infty]$ be defined by

$$\ell_0(t, v, w) := \inf_{v \in U(t)} \{g_0(t, v) : g_1(t, v) \leq w^1, \ldots, g_m(t, v) \leq w^m, b(t, v) = v\}$$

(by convention, $\ell_0(t, v, w) := +\infty$ if the set over which the infimum is taken is empty). Observe that

$$\ell_0(t, v, w) \;=\; \inf\{w^0 : (w^0, w) \in Q(t, v)\}.$$

Therefore, by closedness of $Q(t, v)$ (a direct consequence of Q having property (K)), each of the above two infima on the right side is *attained* as soon as it is merely finite (i.e., if $\ell_0(t, v, w) < +\infty$). From the latter expression, it also follows easily that $\ell_0(t, \cdot, \cdot)$ is lower semicontinuous on $\boldsymbol{R}^\nu \times [-\infty, +\infty]^m$ for every $t \in T$, thanks again to property (K); cf. [15], [6]. Although the graph of the orientor field Q is measurable, this does not imply that ℓ_0 should be measurable (the reason being that projections of measurable sets need not be measurable). Furthermore, for $i = 1, \ldots, m$ we define

$$\ell_i(t, v, w) := \begin{cases} w^i & \text{if } (w^0, w) \in Q(t, v) \text{ for some } w^0 \in \boldsymbol{R}, \\ +\infty & \text{otherwise.} \end{cases}$$

Observe that $\ell_i(t, v, w) \geq w^i$ and that $\ell_i(t, \cdot, \cdot)$ is lower semicontinuous for every $t \in T$, by property (K). We define the optimal control problem (B_{red}) as follows:

$$\text{minimize } I_{\text{red}}(\pi, z) := I^*_{\ell_0}(\pi, z) + I_{\bar{g}}(\pi, y_{\pi,z}) + e(\pi, s(y_{\pi,z})) + \bar{e}(\pi, y_{\pi,z})$$

over all pairs (π, z), where $z : \Gamma(\pi) \to U_{\text{red}}$ is measurable and satisfies

$$\int_{\Gamma(\pi)} |b_{\text{red}}(t, z(t))| \mu(dt) < +\infty \qquad \text{(admissibility)},$$

$$I^*_{\ell_1}(\pi, z) + \alpha_1(\pi) \le 0, \dots, I^*_{\ell_m}(\pi, z) + \alpha_m(\pi) \le 0 \qquad \text{(feasibility)}.$$

Here $U_{\text{red}} := \boldsymbol{R}^\nu \times [-\infty, +\infty]^m$, $b_{\text{red}}(t, v, w) := |v|$, and $y_{\pi,z}$ is given by

$$y_{\pi,z} := \Phi(\pi, \mathcal{A}_{\text{red}}(\pi, z)),$$

with

$$\mathcal{A}_{\text{red}}(\pi, z)(t) := \int_{\Gamma(\pi)} \kappa(t, \tau) b_{\text{red}}(\tau, z(\tau)) \mu(d\tau).$$

Again outer integration is used, as follows:

$$I^*_{\ell_i}(\pi, z) := \inf \left\{ \int_{\Gamma(\pi)} \phi \, d\mu : \phi \in L_1(T), \phi(t) \ge \ell_i(t, z(t)) \quad \text{a.e.} \right\}.$$

PROPOSITION 4.11. *Under the conditions of* §2, *there exists an optimal solution* (π_*, z_*) *for* (B_{red}).

Proof. Let us apply Proposition 4.10. By definition of ℓ_0, the original lower bound for g_0 implies that

$$\ell_0(t, v, w) \ge \chi'(|v|) - \phi_0(t) \ge -\phi_0(t) \quad \text{on } T \times U_{\text{red}}.$$

Likewise, the original inequality $g_i + \epsilon\chi'(|b|) \ge -\phi_\epsilon$ now implies that

$$\ell_i(t, v, w) + \epsilon\chi'(|v|) \ge -\phi_\epsilon(t) \quad \text{on } T \times U_{\text{red}}.$$

Note further that $b_{\text{red}}(t, v, w) := |v|$, defined above, is automatically inf-compact in (v, w) (note the role played by the intrinsic compactness of $[-\infty, +\infty]^m$). Thus condition (i) of Remark 2.1 is valid. Also, above the $\ell_i(t, v, w)$ were seen to be lower semicontinuous in the variable (v, w); thus, condition (ii) of that remark is met. Lastly, we work here with $U_{\text{red}}(t) \equiv U_{\text{red}}$, so condition (iii) of that remark holds, too. The remaining conditions of Proposition 4.10 are easily seen to hold as well for (B_{red}), so we conclude from Proposition 4.10 that (B_{red}) has an optimal solution (π_*, z_*). □

LEMMA 4.12. $\inf(B_{\text{red}}) \le \inf(B)$.

Proof. For any pair $(\pi, u) \in \mathcal{G}$, feasible and admissible for (B), we obtain an admissible pair (π, z) for (B_{red}) by setting $z(t) := (b(t, u(t)), g_1(t, u(t)), \dots, g_m(t, u(t)))$ (recall that for (B) itself the g_i were—and still are—measurable). Clearly, $\ell_i(t, z(t)) = g_i(t, u(t))$ on $\Gamma(\pi)$ for all $i = 1, \dots, m$, so (π, z) is also feasible. Furthermore, $\ell_0(t, z(t)) \le g_0(t, u(t))$ and $y_{\pi,u}(t) = y_{\pi,z}(t)$, so it follows that $I_{\text{red}}(\pi, z) \le I(\pi, u)$. □

We now finish the proof of Theorem 2.2. By Lemma 4.12, Proposition 4.11, and $\inf(B) < +\infty$, it follows that $I_{\text{red}}(\pi_*, z_*)$; hence $I^*_{\ell_0}(\pi_*, z_*)$ is finite. The latter implies that there exists $\phi_* \in L_1(\Gamma(\pi_*))$ with $\ell_0(t, z_*(t)) \le \phi_*(t)$ and $\int_{\Gamma(\pi)} \phi_* d\mu = I^*_{\ell_0}(\pi_*, z_*)$; cf. [8, Prop. A.1]. In particular, this implies that $\ell_0(t, z_*(t)) < +\infty$ almost everywhere. By what was observed about attainment of the infimum in connection with finiteness of the expression defining ℓ_0, it follows that, for almost every t in $\Gamma(\pi_*)$, there exists

$u_t \in U(t)$ such that $\ell_0(t, z_*(t)) = g_0(t, u_t)$, $b(t, u_t) = v_*(t)$, and $g_i(t, u_t) \leq w_*^i(t)$ almost everywhere for all i. Here we write $z_*(t) =: (v_*(t), w_*^1(t), \ldots, w_*^m(t))$. By a standard measurable selection argument, this implies that there exists a measurable function $u_* : \Gamma(\pi_*) \to U$ such that $u_*(t) \in U(t)$, $g_0(t, u_*(t)) \leq \phi_*(t)$, $b(t, u_*(t)) = v_*(t)$, and $g_i(t, u_*(t)) \leq w_*^i(t)$ almost everwhere for all i. By the properties of z_*, it is now easy to see that $(\pi_*, u_*) \in \mathcal{G}$ is feasible and admissible for (B) and satisfies $I(\pi_*, u_*) \leq I_{\text{red}}(\pi_*, z_*) = \inf(B_{\text{red}}) \leq \inf(B)$. This proves that (π_*, u_*) is the desired optimal control pair for (B). The proof of Theorem 2.2 has thus been finished.

Taking Remark 4.5 into account, it is easy to see that all conditions of Theorem 2.2 are fulfilled in Theorem 2.3, except for that concerning the underlying measure space (now T does not have to be compact and metric). Compactness and metrizability of T were only used in the proof of Lemma 4.8, to ensure the existence of a Riesz representation. In all other steps of the proof of Theorem 2.2, an abstract measure space could be used, except for the proof of Lemma 4.1. There the fact was used that the σ-algebra $\mathcal{T}$ (T being there a separable metric space) was countably generated (indeed, this implies separability of $L_1 := L_1(T, \mathcal{T}, \mu; C(\hat{U}))$ and, by [19, 12.F], metrizability of the weak star topology on $\hat{R}$). Yet we need not even retain the condition that $\mathcal{T}$ is countably generated, by concentrating on a minimizing sequence for (B'), and using a conditional expectation argument; see the proof of [6, A.3, p. 592].

5. Applications and discussion. In this section, applications are presented to show that Theorems 2.2 and 2.3 are very flexible and general. We start with an example that describes an important dynamical system used in this section.

Example 5.1. Consider on $T = [0, 1]$ the following differential equation in n dimensions:

$$\dot{y}(t) = B(t)y(t) + b(t, u(t)) \quad \text{a.e. in } [t_0, t_1],$$

with variable initial/final time $t_0, t_1 \in T$, $t_0 \leq t_1$. Here the matrix-valued function $B : [0, 1] \to \boldsymbol{R}^{n \times n}$ is supposed to be Lebesgue-integrable. Also, $b : [0, 1] \times U \to \boldsymbol{R}^n$ is supposed to be product measurable, where U is a Borel-measurable subset of $\boldsymbol{R}^q$.

Suppose that the trajectories y must "hit" a compact space-time target set $H \subset [0, 1] \times \boldsymbol{R}^n$, i.e., must satisfy

$$(t', y(t')) \in H \quad \text{for some } t' \in [t_0, t_1].$$

We choose for the space Π of control parameters the set of all $\pi := (t_0, t_1, t', p')$ with $(t_0, t_1) \in [0, 1]^2$, $t_0 \leq t' \leq t_1$, and $(t', p') \in H$. We set $\Gamma(\pi) := [t_0, t_1]$ and

$$\mathcal{A}(\pi, u)(t) := \int_{[t_0, t_1]} \kappa(t, \tau) b(\tau, u(\tau)) d\tau,$$

where $\kappa(t, \tau) := 1$ for $t \geq \tau$ and $\kappa(t, \tau) = 0$, otherwise. Note that this gives $\mathcal{A}(\pi, u)(t) = \mathcal{A}(\pi, u)(t_1)$ for $t \geq t_1$ and $\mathcal{A}(\pi, u)(t) = 0$ for $t \leq t_0$ (recall that we write $\pi := (t_0, t_1, t', p')$). With $\sigma(\alpha) := \alpha$, continuous hereditarity is obvious. The well-known variation of constants formula for the "real" solution $\tilde{y}_{\pi, u}$, corresponding to an admissible pair $(\pi, u) \in \mathcal{G}$, is as follows [37, II.4.8]:

$$\tilde{y}_{\pi, u}(t) := E(t, t')p' + \int_{t'}^{t} E(t, \tau) b(\tau, u(\tau)) d\tau,$$

for $t_0 \leq t \leq t_1$; this formula remains valid for $t < t'$ by the usual change of sign convention. Here $E(t, \tau) := E(t)E^{-1}(\tau)$, where E and $F := E^{-1}$ are fundamental

matrix solutions of the homogeneous differential equation: $\dot{E} = BE$, $E(0) = I_n$, and $\dot{F} = -FB$, $F(1) = I_n$ (identity matrix). The solutions E and F are absolutely continuous, hence bounded in matrix-norm on $[0,1]$. By partial integration in the above formula, we obtain

$$\tilde{y}_{\pi,u}(t) = E(t,t')p' + \mathcal{A}(\pi,u)(t) - E(t,t')\mathcal{A}(\pi,u)(t') + \int_{t'}^{t} E(t,\tau)B(\tau)\mathcal{A}(\pi,u)(\tau)d\tau$$

for $t \in [t_0,t_1]$ (also this expression is valid for $t < t'$ by the change of sign convention). Therefore, $\Phi : \Pi \times C([0,1];\boldsymbol{R}^n) \to C([0,1];\boldsymbol{R}^n)$ can be chosen as follows:

$$\Phi((t_0,t_1,t',p'),y)(t) := E(t,t')p' + y(t) - E(t,t')y(t') + \int_{t'}^{t} E(t,\tau)B(\tau)y(\tau)d\tau.$$

Observe that $\Phi(\pi,y)$ is continuous in (π,y) and affine in y, as it should be. Note that the constant $c_{\Phi,1} := 1 + \sup_{(t,\tau)\in[0,1]^2, t\geq\tau} |E(t,\tau)|(1 + \int_0^1 |B(t)|\,dt)$ is finite; existence of a suitable finite $c_{\Phi,2}$ follows by compactness of H. It is easy to verify the following facts for the formal solution $y_{\pi,u} := \Phi(\pi,\mathcal{A}(\pi,u))$: On $[t_0,t_1]$, the formal solution $y_{\pi,u}$ coincides with the real solution $\tilde{y}_{\pi,u}$; on $[0,t_0]$ and $[t_1,1]$, the formal solution $y_{\pi,u}$ is the solution of the homogeneous equation $\dot{y} = By$, respectively, with the final condition $y(t_0) = y_{\pi,u}(t_0)$ and the initial condition $y(t_1) = y_{\pi,u}(t_1)$. Thus, we see that the trajectory cost integral functional for (π,u) (defined by integration over $\Gamma(\pi) = [t_0,t_1]$) reflects the costs along the real solution $\tilde{y}_{\pi,u}$. Likewise, for the indirect and direct final cost terms $e(\pi,s(y))$ and $\bar{e}(\pi,y)$ to be meaningful, it is necessary that they, too, are only based on the restriction of y to the time interval $\Gamma(\pi) = [t_0,t_1]$. For instance, to take $s_1(\pi,y) := y(t_0)$ makes sense (by $s_1(\pi,y_{\pi,u}) = s_1(\pi,\tilde{y}_{\pi,u})$), but to set $s_1(\pi,y) := y(0)$ (when $t_0 > 0$) makes only formal sense.

COROLLARY 5.2. *Using the notation of Example* 5.1, *consider the following problem* (B_1):

$$\textit{minimize} \int_{t_0}^{t_1} g_0(t,u(t))dt + \int_{t_0}^{t_1} \bar{g}(t,y(t))dt + c(t_0,t_1,y(t_0),y(t_1)) + \bar{c}(t_0,t_1,y)$$

over all $(t_0,t_1,t',p') \in \Pi$, *all absolutely continuous* $y : [t_0,t_1] \to \boldsymbol{R}^n$ *and all measurable* $u : [t_0,t_1] \to U$, *satisfying*

$$\begin{gathered} \dot{y}(t) = B(t)y(t) + b(t,u(t)) \quad \textit{a.e. in } [t_0,t_1], \\ u(t) \in U(t) \quad \textit{a.e. in } [t_0,t_1], \\ y(t') = p', \\ (t_0,t_1,t',p',y(t_0),y(t_1)) \in D. \end{gathered}$$

Here D *is a closed subset of* $[0,1]^2 \times H \times \boldsymbol{R}^{2n}$; $c : [0,1]^2 \times \boldsymbol{R}^{2n} \to (-\infty,+\infty]$ *and* $\bar{c} : [0,1]^2 \times C([0,1]) \to (-\infty,+\infty]$ *are lower semicontinuous and bounded from below. Suppose also that* $\bar{c}(t_0,t_1,\cdot)$ *is concave on* $C([0,1])$ *for every* t_0,t_1. *Let the assumptions of Example* 5.1 *hold for the dynamical system; let* $g_0 : [0,1] \times U \to (-\infty,+\infty]$ *and* $\bar{g} : [0,1] \times \boldsymbol{R}^n \to (-\infty,+\infty]$ *satisfy the measurability, lower semicontinuity, concavity, growth, and orientor field conditions of Theorem* 2.2 (*the constants* $c_{\Phi,i}$ *were discussed in Example* 5.1). *Assume that* $\inf(B_1) < +\infty$. *Then there exists an optimal control pair for* (B_1).

Proof. In Example 5.1, we already found the representation and integral operators Φ and $\mathcal{A}$ for the dynamical system of (B_1). For control and trajectory integral cost

terms, the conditions of Theorem 2.2 obviously hold. As for the indirect final costs, we define a lower semicontinuous function $e : \Pi \times \boldsymbol{R}^2 \to (-\infty, +\infty]$ by setting

$$e(t_0, t_1, t', p', \zeta_1, \zeta_2) := \begin{cases} c(t_0, t_1, \zeta_1, \zeta_2) & \text{if } (t_0, t_1, t', p', \zeta_1, \zeta_2) \in D, \\ +\infty & \text{otherwise.} \end{cases}$$

Together with this, we set $s_1(t_0, t_1, t', p', y) := y(t_0)$ and $s_2(t_0, t_1, t', p', y) := y(t_1)$; these evaluations are continuous in all variables jointly, and linear in y, as required in Theorem 2.2. The treatment of the direct final cost term is obvious. Therefore, we conclude that all conditions of Theorem 2.2 have been met. The result follows immediately. □

The main results of Raymond [31, Thm. 4.2], Cellina and Colombo [14], and the main existence result without convexity of Cesari [15, 16.5.i] are special cases of Corollary 5.2. This is now demonstrated.

In the notation of the present paper, Raymond works with the dynamical system

$$\begin{gathered} \dot{y}(t) = B(t)y(t) + u'(t) \quad \text{a.e. in } [t_0, t_1], \\ (u(t), u'(t)) \in U'(t) \quad \text{a.e. in } [t_0, t_1], \\ y(t_0) = p \in P. \end{gathered}$$

To understand this, note that he works with a *multifunction* $b : M \to 2^{\boldsymbol{R}^n}$, having measurable graph [31, (H1)]. Although in appearance it would seem more general to work with the differential inclusion

$$\dot{y}(t) - B(t)y(t) \in b(t, u(t)), \qquad u(t) \in U(t) \quad \text{a.e. in } [t_0, t_1],$$

as he does, we observe that this relation is, after all, only equivalent to

$$\dot{y}(t) - B(t)y(t) = u'(t), \quad u'(t) \in b(t, u(t)), \quad u(t) \in U(t) \quad \text{a.e. in } [t_0, t_1],$$

by an implicit measurable function theorem [13, III.38]. Taking for $U'(t)$ the set of all pairs $(v, v') \in \boldsymbol{R}^{q+n}$ with $v' \in b(t, v)$ shows the point (it is no coincidence that in his hypothesis [31, (H2)] precisely the present $U'(t)$ is required to be closed for every t). As for the hypotheses of the above corollary, these hold in Raymond's paper by the following substitutions : If his parameter q equals 1, then take $\chi'(\xi) := c_1\Phi_1$ (c_1, Φ_1 is as in his condition (H12) on p. 125) and $\bar{m}_1(\xi) := \xi^q = \xi, \bar{\phi}_1 \equiv c_2$ (here c_2 is as in his condition (H6) on p. 112). By superlinear growth of his Φ_1, the required limit property follows. If Raymond's parameter q is strictly larger than 1, then use $\chi'(\xi) := c_1\xi^q, \bar{m}_1 := \xi^q, \bar{m}_2 := |\xi|^s$ (his s then satisfies $1 \leq s < q$) and $\bar{\phi}_1 \equiv c_2, \bar{\phi}_2 :=$ his a_2. Actually, Raymond's (H6) has $c_1 > c_2 c_{\Phi,1}^2$, with $c_{\Phi,1} := \sup_{t \geq \tau} |E(t, \tau)|$ (cf. Example 5.1), but let us observe that this can be relaxed into the slightly weaker $c_1 > c_2 c_{\Phi,1}$ (it is weaker since $c_{\Phi,1} \geq 1$, as follows from the semigroup property). When this is substituted, we obtain precisely the desired limit property of Theorem 2.2. Furthermore, in the notation developed above, the orientor field corresponding to [31] is given by

$$Q(t, v) := \{w \in \boldsymbol{R} : w \geq g_0(t, v), (v, v) \in U'(t)\}.$$

By Raymond's (H12), there exist a nonnegative $\Phi(t, |v|)$ (his notation), inf-compact in $|v|$, and an integrable function ϕ such that $g_0(t, v) \geq \Phi(t, |v|) + c_1|v'|^q - \phi(t)$ whenever $v' \in c(t, v)$. Thus, property (K) holds by Remark 2.1, since his conditions

cause $U'(t)$ to be closed, as noted above. All remaining assumptions are easily seen to hold in Raymond's paper (see also Example 5.1). In passing, in the above derivation, an improvement of the growth condition was indicated: $c^2_{\Phi,1}$ can be replaced by $c_{\Phi,1}$ in his (H6). More significant generalizations are that, unlike [31], Corollary 5.2 does not ask for an explicit (inf-)compactness condition for the function g_0 and that the "hitting" condition improves upon the initial condition used in [31]. Also, there is a minor technical improvement in terms of the conditions imposed on the function B (Raymond supposes continuity of B).

Next, we observe that the result of Cellina and Colombo [14] also follows from Corollary 5.2, since their result already follows from Raymond's [31, Thm. 4.2] (they work with a calculus of variations model, which translates into the optimal control model of the corollary by taking $B = 0$ and $b(t, v) := v$).

Corollary 5.2 also generalizes [15, 16.5.i], the principal existence result "without convexity" of Cesari [15]; this result has been formulated in its strongest form by Cesari in his comments on p. 464 of [15] ("alternate hypotheses"), which serve to strengthen the actual statement of 16.5.i in [15]. Of course, the most important difference in Cesari's result lies in fact that he supposes $\bar{g}(t, \cdot)$ to be *linear*, instead of concave (in that case, we can set $\bar{m}_1(|\eta|) := |\eta|$, so the required limit property holds automatically by superlinear growth of χ'). It should be noted that the property (K) condition used in [15], which might seem to be more demanding at first, is equivalent to the one used in this paper. A more minor point is that Cesari assumes—unneccessarily—that B be bounded. His measurability and (semi)continuity conditions are stated in an equivalent "Scorza–Dragoni form."

An even more general version of the above result, which deals with the optimal control of a linear Volterra equation, could now be stated. It considerably generalizes the main existence result of Angell in [1]. However, in [10, Cor. 3.2] such a Volterra equation has already been shown to fit into the model (there this fact was used to derive from Theorem 2.2 a generalization of the *bang-bang* existence result of Angell [1]); with the above corollary in mind, we can easily obtain the intended extension. (Incidentally, as mentioned in the Introduction, applications of the present paper to the subject of optimal bang-bang control can be found in [10].)

Furthermore, the recent existence result of Mariconda [25, Thm. 2], stated first for a parametric problem in the calculus of variations, also follows from Corollary 5.2 above. More precisely, we improve one of Mariconda's conditions in passing. Indeed, as was already observed in [25, p. 296], by a classical argument, his variational problem can be restated in the following optimal control form: Minimize $\int_0^{t_1} g_0(u(t))dt$ over all t_1 in some bounded closed interval $\boldsymbol{R}_+$ (we can suppose this to be [0,1] without any loss of generality) and over all measurable $u : [0, 1] \to U$, where U is the unit sphere in $\boldsymbol{R}^n$ (actually, any compact subset of $\boldsymbol{R}^n$ would do for the present proof). The dynamical system is then $\dot{y}(t) = u(t)$ almost everywhere. Also, the following constraints are imposed in Mariconda's problem: $y(0) \in H'$, $y(t_1) \in D'$, where $H' \subset \boldsymbol{R}^n$ is compact and $D' \subset \boldsymbol{R}^n$ closed. Thus, in terms of Corollary 5.2, we have $H = [0, 1] \times H'$, $B \equiv 0$, $b(t, v) = v$, $U(t) \equiv U$, and the corresponding D is the set of all $(0, t_1, t', p', y_0, y_1)$ with $t_1 = t' \in [0, 1]$, $p' \in H'$, $y_0 \in H'$ and $y_1 \in D'$ (for instance). Note that, in his setup, the conditions of Remark 2.1 are satisified (if we take $g_0 : U \to \boldsymbol{R}$ to be lower semicontinuous on the compact U, it is inf-compact; this is less than what Mariconda needs). As for coercivity, we can, of course, set $\chi'(\xi) = 0$ for $\xi \leq 1$ and $\chi'(\xi) = +\infty$ for $\xi > 1$ (trivially, we set $\bar{g} \equiv 0$, $\phi_i \equiv 0$, and so forth).

Another existence result to which Corollary 5.2 applies can be found in [20],

[21]. Although the main existence result of Ioffe in [20] (in [21], this is Theorem 2 of §9.2.1) addresses existence under convexity conditions, the special problem: Minimize $\int_{t_0}^{t_1}[|u(t)|^{\nu_1} - |y(t)|^{\nu_2}]dt$, $\dot{y}(t) = u(t)$, $y(t_0) = \eta_0$, $y(t_1) = \eta_2$, which is considered in §3 of [20] (and in §9.2.1 of [21]) can be approached by means of Corollary 5.2. For $\nu_1 > \nu_2 \geq 1$, this is straightforward (set $g_0(t,v) := |v|^{\nu_1}$, $\bar{g}(t,\eta) := |\eta|^{\nu_2}$). For $\nu_1 = \nu_2$, the limit property of the corollary (i.e., the one of Theorem 2.2) certainly holds if $t_1 - t_0 < 1$. Ioffe goes further by showing that, in this case, $t_1 - t_0 < 2$ is actually enough. (Incidentally, for $t_1 - t_0 \geq 2$, the infimum of the control problem is $-\infty$, and no optimal solutions exist.)

Let us also recognize that the simple problem (B_0) is out of reach of all the results mentioned above. Yet it follows directly from Corollary 5.2: Set $\Gamma(t_0,t_1,t',p') \equiv [0,1]$ for all (t_0,t_1,t',p') in Π, $g_0(t,u) := (v^2-1)^2$, $\bar{g}(t,\eta) := -\eta^2$, $e'(t_0,t_1,t',p',y) := -\exp(-\|y\|)$, and $D := [0,1]^2 \times H \times \boldsymbol{R}^2$. We also set $b(t,v) := v$ and $B \equiv 0$, $c \equiv 0$, and $\bar{c}(t_0,t_1,y) := -\exp(-\|y\|)$. Then it is easy to see that the assumptions and conditions of Corollary 5.2 hold, with $\chi'(\xi) := [\max(\xi^2-1,0)]^2$ and $\bar{m}_1(\eta) := \eta^2$.

These applications of Theorem 2.2 are finished by stating another existence result that is out of the reach of the current literature; it concerns the optimal control of a functional-integral equation over a fixed compact metric time domain T.

COROLLARY 5.3. *In the notation of Theorem 2.2, but with a trivial control parameter space* $\Pi := \{0\}$, *consider the following optimal control problem* (B_2):

$$\text{minimize } \int_T g_0(t,u(t))\mu(dt) + \int_T \bar{g}(t,y_u(t))\mu(dt)$$

over all continuous $y : T \to \boldsymbol{R}^n$ *and all measurable* $u : T \to U$, *satisfying*

$$y(t) = Sy(t) + \mathcal{A}(0,u)(t) \quad \text{on } T,$$
$$u(t) \in U(t) \quad \text{a.e. in } T.$$

Here S *is the functional-integral operator given by*

$$Sy(t) := \int_T \kappa'(t,\tau)\theta(y)(\tau)\mu(d\tau),$$

where θ *is a bounded linear operator from* $\mathcal{C}(T;\boldsymbol{R}^n)$ *into* $L_\infty(T,\mathcal{T},\mu;\boldsymbol{R}^\nu)$. *It is supposed that* $I - S$ *is an injection (under suitable hereditarity conditions, this is known to be the case* [37, II.5.6]). *Also,* $\kappa' : T \times T \to \boldsymbol{R}^{\nu\times n}$ *is such that* $t \mapsto \kappa'(t,\cdot)$ *can be considered as a continuous mapping from* T *into* $L_1(T,\mathcal{T},\mu;\boldsymbol{R}^{\nu\times n})$. *Let* $g_0 : [0,1]\times U \to (-\infty,+\infty]$ *and* $\bar{g} : [0,1]\times\boldsymbol{R}^n \to (-\infty,+\infty]$ *satisfy the measurability, lower semicontinuity, concavity, growth, and orientor field conditions of Theorem* 2.2, *taking* $c_{\Phi,1} := \|(I-S)^{-1}1\|$ *and* $c_{\Phi,2} := 0$. *Assume that* $\inf(B_2) < +\infty$. *Then there exists an optimal control function for* (B_2).

Proof. Since S is a compact linear operator [37, II.5.5], it follows by [37, I.3.13] that the inverse $(I-S)^{-1}$ is well defined and continuous (Fredholm alternative); cf. [37, II.5.5]. So if we set $\Pi := \{0\}$ and $\Phi(y,0) := (I-S)^{-1}y$, the representation structure for the trajectories, as required in Theorem 2.2, holds. All other conditions are easily seen to hold, so the result follows from Theorem 2.2. □

Let us finish with an application of Theorem 2.3. It coincides with the existence result of Artstein [2], [5].

COROLLARY 5.4. *Consider the problem* (B_3)

$$\text{minimize } \int_T g_0(t,u(t))\mu(dt)$$

over all measurable functions $u : T \to U$ *satisfying*

$$u(t) \in U(t) \quad \text{a.e. in } T$$

and

$$\int_T g_j(t, u(t))\mu(dt) \leq \beta_j, \qquad j = 1, \ldots, m,$$

where $(T, \mathcal{T}, \mu)$ *is a* σ*-finite nonatomic measure space, and* U, M *are as in* §2. *Suppose that* $g_0, \ldots, g_m : M \to (-\infty, +\infty]$ *are all measurable and lower semicontinuous in their second variable and suppose that*

$$g_m(t, \cdot) \text{ is inf-compact for every } t \in T$$

and that $U(t)$ *is closed for all* t. *Finally, suppose that*

$$g_m \text{ is integrably bounded from below}$$

and that, for every $\epsilon > 0$, *there exists an integrable* $\phi_\epsilon \in L_1^+(T)$ *such that*

$$g_i(t, v) + \epsilon g_m(t, v) \geq -\phi_\epsilon(t) \quad \text{on } M$$

for $i = 0, 1, \ldots, m-1$. *Assume that* $\inf(B_3) < +\infty$. *Then there exists an optimal solution for* (B_3).

Proof. By Remark 2.1, we can simply apply Theorem 2.3. □

Acknowledgments. I thank the referees for their helpful comments and criticism. The evolution of this paper was greatly influenced by invited lectures in Italy, France, Canada, and the Netherlands. To all hosts and genii loci involved, I extend my special thanks.

REFERENCES

[1] T. ANGELL, *Existence of optimal control without convexity and a bang-bang theorem for linear Volterra equations*, J. Optim. Theory Appl., 19 (1976), pp. 63–79.

[2] Z. ARTSTEIN, *On a variational problem*, J. Math. Anal. Appl., 45 (1974), pp. 404–415.

[3] R. J. AUMANN AND M. PERLES, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965), pp. 488–503.

[4] E. J. BALDER, *Existence results by extremal properties of the original controls*, unpublished mimeo, Mathematical Institute, University of Utrecht, Utrecht, the Netherlands, 1978.

[5] ———, *On a useful compactification for optimal control problems*, J. Math. Anal. Appl., 72 (1979), pp. 391–398.

[6] ———, *A general approach to lower semicontinuity and lower closure in optimal control theory*, SIAM J. Control Optim., 22 (1984), pp. 570–598.

[7] ———, *An extension of Prohorov's theorem for transition probabilities with applications to infinite-dimensional lower closure problems*, Rend. Circ. Mat. Palermo, (II) 34 (1985), pp. 427–447.

[8] ———, *On seminormality of integral functionals and their integrands*, SIAM J. Control Optim., 24 (1986), pp. 95–121.

[9] ———, *Generalized equilibrium results for games with incomplete information*, Math. Oper. Res., 13 (1988), pp. 265–276.

[10] ———, *Exact bang-bang optimal control for problems with nonlinear costs*, in Advances in Optimization, W. Oettli and D. Pallaschke, eds., Lecture Notes in Economics and Mathematical Systems, Vol. 382, Springer-Verlag, Berlin, 1992, pp. 371–383.

[11] J. M. BALL AND G. KNOWLES, *Young measures and minimization problems of mechanics*, in Elasticity, Mathematical Methods and Applications, G. Eason and R. W. Ogden, eds., Ellis Horwood Ltd., Chichester, UK, 1990, pp. 1–20.

[12] H. Berliocchi and J.-M. Lasry, *Intégrandes normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1973), pp. 129–184.
[13] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics, Vol. 580, Springer-Verlag, Berlin, 1977.
[14] A. Cellina and G. Colombo, *On a classical problem of the calculus of variations without convexity assumptions*, Ann. Inst. H. Poincaré, Analyse non linéaire, 7 (1989), pp. 97–106.
[15] L. Cesari, *Optimization Theory and Applications: Problems with Ordinary Differential Equations*, Springer-Verlag, Berlin, 1983.
[16] G. Choquet, *Lectures on Analysis*, Benjamin, Reading, MA, 1969.
[17] C. Dellacherie and P.-A. Meyer, *Probabilités et Potentiel*, Hermann, Paris, 1975.
[18] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
[19] R. B. Holmes, *Geometric Functional Analysis*, Springer-Verlag, Berlin, 1975.
[20] A. D. Ioffe, *An existence theorem for problems of the calculus of variations*, Soviet Math. Dokl., 13 (1972), pp. 919–923.
[21] A. D. Ioffe and V. M. Tichomirov, *Theorie der Extremalaufgaben*, Nauka, Moscow, 1974; German transl., Deutscher Verlag der Wissenschaften, Berlin, 1979.
[22] A. and C. Ionescu-Tulcea, *Topics in the Theory of Lifting*, Springer-Verlag, Berlin, 1969.
[23] H.-P. Kirschner, *On the risk-equivalence of two methods of randomization in statistics*, J. Multivariate Anal., 6 (1976), pp. 159–166.
[24] P. Marcellini, *Alcune osservazioni sull'esistenza del minimo di integrali del calcolo delle variazioni senza ipotesi di convessità*, Rend. Mat., 13 (1980), pp. 271–281.
[25] C. Mariconda, *On a parametric problem of the calculus of variations without convexity conditions*, J. Math. Anal. Appl., 170 (1992), pp. 291–297.
[26] E. J. McShane, *Relaxed controls and variational problems*, SIAM J. Control, 5 (1967), pp. 438–485.
[27] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
[28] C. Olech, *Integrals of set-valued functions and linear optimal control problems*, Colloque sur la Théorie Mathématique du Contrôle Optimal, C.B.R.M., Vander Louvain, 1970, pp. 109–125.
[29] J.-P. Raymond, *Conditions nécessaires et suffisantes d'existence de solutions en calcul des variations*, Ann. Inst. H. Poincaré, Analyse non linéaire, 4 (1987), pp. 169–202.
[30] ———, *Problèmes de calcul des variations et de contrôle optimal: existence et régularité des solutions*, Habilitation en Sciences Mathématiques, Université Paul Sabatier, Toulouse, France, 1990.
[31] ———, *Existence theorems in optimal control theory without convexity assumptions*, J. Optim. Theory Appl., 67 (1990), pp. 109–132.
[32] R.T. Rockafellar, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, J. P. Gossez et al., eds., Lecture Notes in Mathematics, Vol. 543, Springer-Verlag, Berlin, 1976, pp. 157–205.
[33] P. C. Rosenbloom, *Quelques classes de problèmes extrémaux*, Bull. Soc. Math. France, 80 (1952), pp. 183–216.
[34] J. E. Rubio, *Extremal points in the calculus of variations*, Bull. London Math. Soc., 7 (1975), pp. 159–165.
[35] ———, *Control and Optimisation: The Linear Treatment of Nonlinear Problems*, Manchester University Press, Manchester, UK, 1986.
[36] M. Valadier, *Young measures*, in Methods of Nonconvex Analysis, A. Cellina, ed., Lecture Notes in Mathematics, Vol. 1446, Springer-Verlag, Berlin, 1990, pp. 152–188.
[37] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[38] G. Winkler, *Extreme points of moment sets*, Math. Oper. Res., 13 (1988), pp. 581–587.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 917–936, July 1994

AN ADAPTIVE SERVOMECHANISM FOR A CLASS OF INFINITE-DIMENSIONAL SYSTEMS*

HARTMUT LOGEMANN† AND ACHIM ILCHMANN‡

Abstract. A universal adaptive controller is constructed that achieves asymptotic tracking of a given class of reference signals and asymptotic rejection of a prescribed set of disturbance signals for a class of multivariable infinite-dimensional systems that are stabilizable by high-gain output feedback. The controller does not require an explicit identification of the system parameters or the injection of a probing signal. In contrast to most of the work in universal adaptive control, this paper is based on an input-output approach and the results do not require a state-space representation of the plant. The abstract input-output results are applied to retarded systems and integrodifferential systems.

Key words. servomechanisms, adaptive control, high-gain feedback, infinite-dimensional systems, functional differential equations, input-output methods

AMS subject classifications. 34K20, 93B52, 93C25, 93C30, 93C35, 93C40, 93D15, 93D21, 93D25

1. Introduction.

One of the most important applications of feedback is to achieve servoaction, that is, to obtain a closed-loop system that tracks a given class of reference signals and rejects a given class of external disturbances with zero asymptotic error. This problem has been well understood for many years provided that the plant is linear and time-invariant and the plant uncertainty is sufficiently small (see Wonham [30, p. 203] and Vidyasagar [27, p. 294] for finite-dimensional systems and Francis [4], Callier and Desoer [2], and Curtain [3] for infinite-dimensional systems). The basic design principle in the theory of linear servomechanisms, which is also referred to as the *internal model principle*, says (roughly speaking) that a controller that achieves robust servoaction necessarily contains a duplicate of the dynamics of the reference and disturbance signals.

If the plant uncertainty is large, which is the case if only certain structural information on the plant is available to the designer, it is desirable to construct a universal adaptive servomechanism, that is, a fixed nonlinear controller that achieves servoaction for a whole prescribed class of linear time-invariant systems and all possible initial conditions without explicit identification of the system parameters. Although the problem of universal adaptive stabilization of finite and infinite-dimensional systems has received considerable attention in recent years (cf. e.g., Mårtensson [17], [18], Logemann and Owens [14], Logemann and Mårtensson [13], and the references therein), there are only few papers on universal adaptive servomechanisms, which in addition deal exclusively with finite-dimensional systems. Mårtensson [19] pointed out that adaptive tracking of constant reference signals can be easily achieved for a given class of multivariable systems if a universal adaptive stabilizer is known and the class is invariant under precompensation by an integrator. Helmke, Prätzel-Wolters, and Schmid [8] proved a similar result for single-input single-output systems allowing for a more general class of reference signals including ramps, linear combinations of sinusoidal signals, etc. If the plant is known to lie in a given finite set of (multivariable) systems, if the reference and disturbance signals belong to the solution space of a given linear autonomous differential equation, and if an L^∞-bound on the disturbances is known, Miller and Davison [21] constructed a switching controller that solves the servoproblem for any plant in this finite set. In [20] Miller and Davi-

* Received by the editors July 8, 1991; accepted for publication (in revised form) December 23, 1992.

† Institut für Dynamische Systeme, Fachbereich Mathematik/Informatik, Universität Bremen, Postfach 330440, 2800 Bremen 33, Germany. Current address, School of Mathematical Sciences, University of Bath, Bath BA2 2BZ, United Kingdom.

‡ Centre for Systems and Control Engineering, School of Engineering, University of Exeter, Exeter, Devon EX4 4QF, United Kingdom. Current address, Institut für Angewandte Mathematik der Universität Hamburg, Bundesstr. 55, 2000 Hamburg 13, Germany.

son presented a low-gain controller that carries out asymptotic error regulation for constant reference and disturbance signals for any multivariable plant, provided it is asymptotically stable and has no transmission zeros at zero. For the class of all single-input single-output, relative degree one or two, minimum-phase systems of McMillan degree less than or equal to n, Morse [22] constructed a $4(n+1)$-dimensional model reference adaptive controller that achieves asymptotic tracking of any signal generated by a two-dimensional reference system. On the basis of high-gain concepts, Mareels [16] introduced a control law that solves the tracking problem for any single-input single-output minimum-phase system of known relative degree, provided the sign of the high-frequency gain and an upper bound on its magnitude is known. Finally, for the class of all single-input single-output minimum-phase systems having relative degree one, Helmke, Prätzel-Wolters, and Schmid [9] constructed a high-gain controller that has the property that the resulting closed-loop system tracks any reference signal annihilated by a given linear ordinary differential operator with constant coefficients.

The purpose of this paper is to construct a universal adaptive servomechanism for the class of multivariable infinite-dimensional systems that are minimum-phase and have an invertible high-frequency-gain. We show that the series interconnection of the controller presented in Byrnes and Willems [1] and a suitable precompensator solves the adaptive servoproblem for the class of systems under consideration. This result is also new for the finite-dimensional case. It generalizes the result in [9], where an adaptive tracking problem was solved for finite-dimensional single-input single-output systems. The disturbance rejection problem is not addressed in [9]. Moreover, the proof in [9] does not extend to multivariable systems; neither does it carry over to infinite-dimensional plants, and so the generalization is far from being trivial. We mention that in [9] a state-space approach is used, while our treatment is based on the input-output set-up for high-gain adaptive stabilization as developed by Logemann and Owens [14]. So, in contrast to almost all papers in the area, our approach does not require a state-space model of the plant. Non-zero initial conditions are taken into account by using "initial-condition terms." The input-output results are applied to retarded systems and integrodifferential convolution systems.

The paper is organized as follows. In §2 we introduce a class of infinite-dimensional systems that are stabilizable by high-gain feedback and will be dealt with in the rest of the paper. Moreover, we collect a number of results on a functional differential equation of Volterra type that will be useful in what follows. Section 3 shows that the high-gain based switching algorithm, introduced by Byrnes and Willems [1] in a finite-dimensional state-space set-up, stabilizes any infinite-dimensional plant belonging to the class of systems introduced in §2. Section 4 contains the main result of the paper. We prove that the series connection of the adaptive stabilizer presented in §3, followed by a suitable precompensator containing an internal model of the dynamics of the reference and disturbance signals, achieves servoaction for the class of systems under consideration. Section 5 is devoted to the application of the input-output results of §4 to retarded systems and integrodifferential convolution systems. In particular it is shown that the adaptive servomechanism presented in §4 achieves "internal stability" in the sense that the internal variables of the plant and the precompensator remain bounded provided that the reference signal is bounded. The proof of a technical result is relegated to the Appendix.

Nomenclature.

$\mathbb{C}_+ :=$ open right-half plane.

$\mathbb{C}_- :=$ open left-half plane.

$LL^p(\mathbb{R}_+, \mathbb{R}^n) :=$ vector space of locally p-integrable functions defined on $\mathbb{R}_+$ with values in $\mathbb{R}^n$.

$H^\infty(\mathbb{C}^{n\times n}) :=$ algebra of bounded holomorphic functions defined on $\mathbb{C}_+$ with values in $\mathbb{C}^{n\times n}$.

$H^2(\mathbb{C}^n) :=$ the usual Hardy–Lebesgue space of order 2 of holomorphic functions defined on $\mathbb{C}_+$ with values in $\mathbb{C}^n$.

$BV([a,b],\mathbb{R}^{n\times n}) :=$ vector space of $\mathbb{R}^{n\times n}$-valued functions of bounded variation defined on $[a,b]$.

$M(\mathbb{R}_+,\mathbb{R}^{n\times n}) :=$ vector space of bounded Borel measures on $\mathbb{R}_+$ with values in $\mathbb{R}^{n\times n}$.

Let f be a function defined on $[0,a)$, where $0 < a \le \infty$. Then for all $\tau \in [0,a)$,

$$(\pi_\tau f)(t) := \begin{cases} f(t), & 0 \le t \le \tau, \\ 0, & t > \tau. \end{cases}$$

$\mathcal{L}$ denotes the Laplace transform.

The superscript "$\hat{}$" is used to denote Laplace transformed or Laplace–Stieltjes transformed functions.

2. Preliminaries and system description. We shall assume that externally our plant is described by a transfer-function matrix G of size $m \times m$ which is meromorphic on $\mathbb{C}_+$ and satisfies

$$\text{(2.1)} \qquad \begin{cases} G^{-1}(s) = sD^{-1} + H(s), \\ \text{where} \quad D \in \mathbb{R}^{m\times m},\ \det(D) \neq 0 \quad \text{and} \quad H \in H^\infty(\mathbb{C}^{m\times m}). \end{cases}$$

Of course (2.1) is equivalent to

$$\text{(2.2)} \qquad G(s) = \left(I + \frac{1}{s}DH(s)\right)^{-1} \frac{1}{s}D,$$

i.e., G is the feedback interconnection of the integrator $(1/s)D$ and the transfer-function matrix H.

In order to characterize condition (2.1) in terms of the zeros and the high-frequency behavior of G, we have to make precise what we mean by a zero of a meromorphic transfer-function matrix.

DEFINITION 2.1. *Suppose that R is a matrix of size $m \times m$ whose entries are meromorphic functions defined on a region $\Omega \subset \mathbb{C}$. Let (U,V) be a holomorphic right-coprime factorization of R over Ω, i.e., U and V are holomorphic matrices of size $m \times m$ defined on Ω such that $\det(V(s)) \not\equiv 0$, $R(s) = U(s)V^{-1}(s)$, and there exist holomorphic matrices X and Y of size $m \times m$ defined on Ω satisfying $X(s)U(s)+Y(s)V(s) \equiv I_m$.*[†] *The zeros of $R(s)$ are defined to be the zeros of* $\det(U(s))$.

PROPOSITION 2.2. *Let $G(s)$ be a meromorphic transfer-function matrix of size $m \times m$ defined on a region $\Omega \supset \overline{\mathbb{C}}_+$. Then $G^{-1}(s)$ admits a decomposition of the form* (2.1) *if and only if*

(i) $sG(s) - D = O(1/s)$ *as* $|s| \to \infty$ *in* $\mathbb{C}_+$,

[†] Since the ring of holomorphic functions defined on a region has the property that finitely generated ideals are principal (see Rudin [23, p. 328]) and since the field of meromorphic functions defined on a region is the quotient field of the ring of holomorphic functions defined on that region (see Rudin [23, p. 327]), it follows from Vidyasagar, Schneider, and Francis [28] that such a factorization exists and is unique up to multiplication from the right by unimodular holomorphic matrices.

and

(ii) $G(s)$ *has no zeros in* $\overline{\mathbb{C}}_+$.

Proof. See Logemann and Zwart [15]. □

Note that condition (i) in Proposition 2.2 is a generalization of the relative-degree one condition for finite-dimensional single-input single-output systems.

Remark 2.3. The transfer-function matrix G of a stabilizable and detectable finite-dimensional system $\dot{x} = Ax + Bu$, $y = Cx$ satisfies (2.1) if and only if the system is minimum-phase, i.e.,

$$\det \begin{pmatrix} sI - A & -B \\ C & 0 \end{pmatrix} \neq 0 \quad \text{for all } s \in \overline{\mathbb{C}}_+,$$

and has invertible high-frequency gain, i.e., $\det(CB) \neq 0$. Moreover the matrix D in (2.1) is given by CB.

In the following we shall assign an operator $\mathcal{H} : L^2(\mathbb{R}_+, \mathbb{C}^m) \to L^2(\mathbb{R}_+, \mathbb{C}^m)$ to the transfer-function matrix H by defining $\mathcal{H} := \mathcal{L}^{-1}\mathcal{M}_H\mathcal{L}$, where $\mathcal{L}$ denotes the Laplace transform and $\mathcal{M}_H$ denotes the multiplication by H on the Hardy space $H^2(\mathbb{C}^m)$. The operator $\mathcal{H}$ is linear, bounded, and shift-invariant (in the sense of Vidyasagar [26]). As a consequence $\mathcal{H}$ is causal (see [26]) and therefore has a unique causal extension to $LL^2(\mathbb{R}_+, \mathbb{C}^m)$. This extension will also be denoted by $\mathcal{H}$. The converse is also true, i.e., given a linear, bounded, shift-invariant operator $\mathcal{H} : L^2(\mathbb{R}_+, \mathbb{C}^m) \to L^2(\mathbb{R}_+, \mathbb{C}^m)$, there exists $H \in H^\infty(\mathbb{C}^{m\times m})$ such that $\mathcal{H} = \mathcal{L}^{-1}\mathcal{M}_H\mathcal{L}$ (see Harris and Valenca [7], Logemann [12], and Weiss [29]). Finally we mention that $LL^2(\mathbb{R}_+, \mathbb{R}^m)$ is an $\mathcal{H}$-invariant subspace of $LL^2(\mathbb{R}_+, \mathbb{C}^m)$[‡] if and only if $H(s) = \bar{H}(\bar{s})$ for all $s \in \mathbb{C}_+$. In control applications the latter condition will always be satisfied and it is assumed to hold in the following.

The function G satisfying (2.1) can be thought of as being the transfer-function matrix of

$$\dot{y} = D(u - (\mathcal{H}y + w)), \qquad y(0) = y_0 \in \mathbb{R}^m, \tag{2.3}$$

where $u \in LL^1(\mathbb{R}_+, \mathbb{R}^m)$ and $w \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ takes account of non-zero initial conditions in the system with transfer-function matrix H. The initial value problem (2.3) is a special case of the following initial value problem, which will play an important role in this paper. Consider

$$\begin{aligned} \dot{x}(t) &= (Sx)(t) + f(t, x(t)) + g(t), \qquad t \geq \alpha, \\ x_{|[0,\alpha]} &= x_0 \in C([0, \alpha], \mathbb{R}^n), \qquad \alpha \geq 0, \end{aligned} \tag{2.4}$$

where the following hold.

(i) $S : LL^2(\mathbb{R}_+, \mathbb{R}^n) \to LL^2(\mathbb{R}_+, \mathbb{R}^n)$. We assume that $S(0) = 0$ and that there exists $\kappa > 0$ such that $\|\pi_t(Sx - Sx')\| \leq \kappa\|\pi_t(x - x')\|$ for all $x, x' \in LL^2(\mathbb{R}_+, \mathbb{R}^n)$ and for all $t \geq 0$, i.e., S is unbiased, causal, and of finite incremental gain.

(ii) $f : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ is a function. We assume that $f(t, x)$ is continuous in t and locally Lipschitz continuous in x, uniformly in t on bounded intervals.

(iii) g is in $LL^1(\mathbb{R}_+, \mathbb{R}^n)$.

Of course, if $\alpha = 0$ in (2.4), then $C([0, \alpha], \mathbb{R}^n) = \mathbb{R}^n$. In order to define what we mean by a solution of the initial value problem (2.4) on $[0, \beta)(\alpha < \beta \leq \infty)$, we have to give a meaning to Sx if $x \in C([0, \beta), \mathbb{R}^n)$ (remember that S operates on functions whose domain of definition is $\mathbb{R}_+$). We set $(Sx)(t) = (S\pi_\tau x)(t)$ for $0 \leq t \leq \tau < \beta$. Since S is causal, this definition does not depend on the choice of τ.

[‡] Notice that here $LL^2(\mathbb{R}_+, \mathbb{C}^m)$ is considered as a *real* vector space.

DEFINITION 2.4. *A solution of* (2.4) *on* $[0,\beta)(\alpha < \beta \leq \infty)$ *is an absolutely continuous function* x *on* $[0,\beta)$ *such that* $x_{|[0,\alpha]} = x_0$ *and the differential equation in* (2.4) *is satisfied by* x *almost everywhere on* $[0,\beta)$.

THEOREM 2.5. *The initial-value problem* (2.4) *has a unique solution on some interval* $[0,\beta)$, *where* $\alpha < \beta \leq \infty$. *If* $\beta < \infty$ *and* β *cannot be increased, then there exists a strictly increasing sequence* $t_i \in (0,\beta)$, *satisfying* $\lim_{i\to\infty} t_i = \beta$, *such that* $\lim_{i\to\infty} \|x(t_i)\| = \infty$.

The above theorem has been proved in Logemann and Owens [14]. Similar results can be found in Gripenberg, Londen, and Staffans [5, p. 359], and Hinrichsen and Pritchard [10]. Theorem 2.5 implies in particular that the initial value problem (2.3) has a unique solution for all $w \in L^2(\mathbb{R}_+,\mathbb{R}^m)$, $u \in LL^1(\mathbb{R}_+,\mathbb{R}^m)$, and $y_0 \in \mathbb{R}^m$.

3. Adaptive stabilization. The aim of this section is to construct a universal adaptive control law that stabilizes any system of the form (2.3), i.e., the control law does not depend on D and $\mathcal{H}$, and the closed-loop system satisfies $\lim_{t\to\infty} y(t) = 0$ for all $y_0 \in \mathbb{R}^m$ and $w \in L^2(\mathbb{R}_+,\mathbb{R}^m)$.

In the following, we need a result from linear algebra which has been proved by Mårtensson [17], [18]. For $m \geq 1$ we call a set $\mathcal{U} \subset GL(m,\mathbb{R})$ *unmixing*, if for any $A \in GL(m,\mathbb{R})$ there is a $U \in \mathcal{U}$ such that $\operatorname{spec}(AU) \subset \mathbb{C}_-$.

PROPOSITION 3.1 ([17], [18]). *For all* $m \geq 1$, *there exist unmixing sets of finite cardinality.*

Unfortunately the cardinality of the unmixing sets constructed in [17], [18] is far too large than would be convenient for applications. Hardly anything is known on the minimum cardinality of unmixing sets. However, for $m = 1$ the set $\{1,-1\}$ is obviously unmixing, while for $m = 2$ there exists an unmixing set of cardinality 6 (see [17], [18]). It has been shown by Zhu [31] that $GL(3,\mathbb{R})$ can be unmixed by a set having cardinality 32.

In the following, let $\{K_1,\dots,K_N\}$ be an unmixing set for $GL(m,\mathbb{R})$. Since (2.3) can be stabilized by high-gain feedback of the form $u(t) = ky(t)$, provided that $\operatorname{spec}(D) \subset \mathbb{C}_-$ and k is a sufficiently large positive number, it seems reasonable to consider the following adaptive control law:

$$\begin{aligned} u(t) &= k(t)K_{\sigma(k(t))}y(t), \\ \dot{k}(t) &= \|y(t)\|^2, \qquad k(0) = k_0 \in \mathbb{R}. \end{aligned} \tag{3.1}$$

In (3.1) the function $\sigma : \mathbb{R} \to \{1,\dots,N\}$ is given by

$$\sigma(k) = \begin{cases} 1, & k \in [-\tau_1,\tau_1), \\ i, & k \in [\tau_{lN+i},\tau_{lN+i+1}) \cup [-\tau_{lN+i+1},-\tau_{lN+i}) \quad \text{for some } l \in \mathbb{N}_0, \end{cases} \tag{3.2}$$

where the sequence $(\tau_j)_{j\in\mathbb{N}_0}$ is defined as

$$\tau_{j+1} = \tau_j^2, \qquad \tau_1 > 1. \tag{3.3}$$

Note that the gain $k(t)$ is monotonically increasing and thus the function σ ensures that $K_{\sigma(k(t))}$ will hit some stabilizing gain matrix K_i if $k(t)$ diverges. The growth condition (3.3) captures the intuitive idea that the length of the intervals $[\tau_j,\tau_{j+1})$ should increase rapidly, in order to enable the closed-loop system to settle down. Although the closed-loop system given by (2.3) and (3.1) is of the form (2.4), we cannot apply Theorem 2.5 straight away in order to establish well posedness of the closed loop, since the map $\mathbb{R} \to \{K_1,\dots,K_N\}, k \mapsto K_{\sigma(k)}$ is not continuous. However, Theorem 2.5 can be used to prove the following.

LEMMA 3.2. *For each pair of initial conditions* $(y_0,k_0) \in \mathbb{R}^m \times \mathbb{R}$ *and for each* $w \in L^2(\mathbb{R}_+,\mathbb{R}^m)$, *the closed-loop system given by* (2.3) *and* (3.1) *has a unique absolutely continuous solution* (y,k) *that can be extended to the right as long as it remains bounded.*

Proof. See Appendix.

Now we are in the position to prove the main result of this section. It says that the control law (3.1) stabilizes any system of the form (2.3), or in other words (3.1) is a universal adaptive control law for this class.

THEOREM 3.3. *The solution* (y, k) *of the closed-loop system given by* (2.3) *and* (3.1) *exists on* $\mathbb{R}_+$ *and has the following properties*:

(i) $\lim_{t\to\infty} k(t)$ *exists and is finite*;

(ii) $y \in L^2(\mathbb{R}_+, \mathbb{R}^m) \cap L^\infty(\mathbb{R}_+, \mathbb{R}^m)$;

(iii) $\lim_{t\to\infty} y(t) = 0$.

We shall prove Theorem 3.3 by combining ideas of Byrnes and Willems [1] with the following lemma, which can be found in Ilchmann and Logemann [11].

LEMMA 3.4. *Suppose that* σ *and* τ_j *are given by* (3.2) *and* (3.3), *respectively, and for* $\alpha > 0$ *and* $i \in \{1, \dots, N\}$ *define* $F_i^\alpha : \mathbb{R} \to \{1, -\alpha\}$ *by*

$$F_i^\alpha(x) = \begin{cases} 1, & \text{if } \sigma(x) = i, \\ -\alpha, & \text{if } \sigma(x) \neq i. \end{cases} \tag{3.4}$$

Then we have

$$\sup_{k>k_0} \frac{1}{k - k_0} \int_{k_0}^{k} x F_i^\alpha(x) dx = +\infty \tag{3.5}$$

for all $k_0 \in \mathbb{R}, \alpha > 0, i \in \{1, \dots, N\}$.

Proof of Theorem 3.3. By assumption there exists $i \in \{1, \dots, N\}$ such that $\operatorname{spec}(DK_i) \subset \mathbb{C}_-$. Hence there is a positive definite matrix $Q = Q^T \in GL(m, \mathbb{R})$ satisfying

$$K_i^T D^T Q + QDK_i = -I. \tag{3.6}$$

Furthermore, choose $\alpha > 0$ such that

$$K_j^T D^T Q + QDK_j \leq \alpha I \quad \text{for all } j \in \{1, \dots, N\}. \tag{3.7}$$

By Lemma 3.2, the closed-loop system given by (2.3) and (3.1) has a unique solution (y, k). Let $[0, t^*)$ denote its maximal interval of existence. Setting $\|x\|_Q := (\langle x, Qx\rangle)^{1/2}$ for $x \in \mathbb{R}^m$ and using (2.3), (3.4), (3.6), and (3.7) we obtain

$$\begin{aligned} \frac{d}{dt}\|y(t)\|_Q^2 &= \dot{y}(t)^T Q y(t) + y(t)^T Q \dot{y}(t) \\ &= k(t) y(t)^T (K_{\sigma(k(t))}^T D^T Q + QDK_{\sigma(k(t))}) y(t) - (\mathcal{H}y)(t)^T D^T Q y(t) \\ &\quad - w(t)^T D^T Q y(t) - y(t)^T Q D(\mathcal{H}y)(t) - y(t)^T Q D w(t) \\ &\leq -F_i^\alpha(k(t)) k(t) \dot{k}(t) - 2y(t)^T Q D(\mathcal{H}y)(t) - 2y(t)^T Q D w(t). \end{aligned} \tag{3.8}$$

Using Hölder's inequality and the causality and boundedness of $\mathcal{H}$, it is easy to show that for all $f \in LL^2(\mathbb{R}_+, \mathbb{C}^m)$ and $t \geq 0$,

$$\left| \int_0^t f(\tau)^T Q D(\mathcal{H}f)(\tau) d\tau \right| \leq \|Q\| \|D\| \|\mathcal{H}\| \int_0^t \|f(\tau)\|^2 d\tau. \tag{3.9}$$

Integrating (3.8) from 0 to t, $t < t^*$, changing variables, and applying (3.9) yields

$$\begin{aligned}\|y(t)\|_Q^2 - \|y_0\|_Q^2 &\leq -\int_{k_0}^{k(t)} xF_i^\alpha(x)dx + \lambda_1 \int_0^t \|y(\tau)\|^2 d\tau \\ &\quad + \lambda_2 \|w\|_2 \left(\int_0^t \|y(\tau)\|^2 d\tau\right)^{1/2} \\ &= (k(t) - k_0)\left\{\lambda_1 + \frac{\lambda_2\|w\|_2}{\sqrt{k(t)-k_0}} - \frac{1}{k(t)-k_0}\int_{k_0}^{k(t)} xF_i^\alpha(x)dx\right\},\end{aligned} \tag{3.10}$$

where λ_1 and λ_2 are suitable positive constants depending on $\mathcal{H}$, D, and Q.

In order to show global existence of the solution (y, k) on $\mathbb{R}_+$ it is (by Lemma 3.2) sufficient to show that (y, k) is bounded on $[0, t^*)$. In order to prove that $k(t)$ is bounded on $[0, t^*)$, assume the contrary. It then follows from Lemma 3.4 that the limes inferior of the right-hand side of (3.10) is $-\infty$, contradicting the fact that the left-hand side of (3.10) is bounded from below by $-\|y_0\|_Q^2$. Hence $k(t)$ is bounded on $[0, t^*)$ and from (3.1) and (3.10) we obtain that $y \in L^2(0, t^*; \mathbb{R}^m) \cap L^\infty(0, t^*; \mathbb{R}^m)$. In particular we have $t^* = \infty$, which implies (i) and (ii). In order to prove (iii), notice that by (2.3), (i), and (ii) $\dot{y} \in L^2(\mathbb{R}_+; \mathbb{R}^m)$. As a consequence (iii) holds true. □

Remark 3.5. (i) It is not difficult to see that the sequence given by (3.3) can be replaced by any strictly increasing sequence $(\tau_j)_{j\in\mathbb{N}}$ satisfying $\lim_{j\to\infty} \tau_j/\tau_{j-1} = +\infty$ (cf. Ilchmann and Logemann [11] and Ryan [24]).

(ii) Let $u \in LL^2(\mathbb{R}_+, \mathbb{R}^m), w \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ and suppose that y satisfies (2.3). If $Q \in \mathbb{R}^{m\times m}$ is positive definite, then the inequality

$$\|y(t)\|_Q^2 \leq \|y_0\|_Q^2 + \mu\left(1 + \int_0^t \|y(\tau)\|^2 d\tau\right) + 2\int_0^t y(\tau)^T QDu(\tau)d\tau \tag{3.11}$$

holds for all $t \geq 0$, where μ is a suitable positive constant depending on $\mathcal{H}, D, Q$, and w. Inequality (3.11) has been derived implicitly in the proof of Theorem 3.3 and may be of some independent interest.

Remark 3.6. The controller (3.1) was introduced by Byrnes and Willems [1] in a finite-dimensional state-space set-up. The main result in [1] says that any finite-dimensional state-space system with m inputs and m outputs can be stabilized by the control law (3.1), provided it is minimum-phase and has invertible high-frequency gain. However, the proof is not convincing, since the inequality (3.4) in [1] is in general wrong. A result similar to that in [1] can be found in Mårtensson [17], [18]. The proof in [17], [18] is not convincing either, since it is based on the claim that for the adaptive control system

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t), \qquad x(0) = x_0 \in \mathbb{R}^n, \\ y(t) &= Cx(t), \\ u(t) &= k(t)Qy(t), \\ \dot{k}(t) &= \|u(t)\|^2 + \|y(t)\|^2, \qquad k(0) = k_0 \in \mathbb{R},\end{aligned}$$

there exist constants $c > 0$ and $T \geq 0$ such that

$$\int_t^\infty (\|u(\tau)\|^2 + \|y(\tau)\|^2)d\tau \leq c\|x(t)\|^2$$

for all $x_0 \in \mathbb{R}^n$, $k_0 \in \mathbb{R}$, $t \geq T$, provided that (A, B, C) is minimum-phase and $\sigma(CBQ) \subset \mathbb{C}_-$.

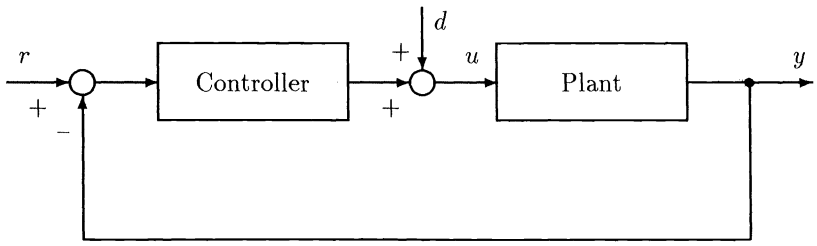


FIG. 1. *Closed-loop system.*

This is not proved in [17], [18] and it seems to the authors that it is unlikely to hold true.

We close this section with a conjecture on the limiting closed-loop system. If the assumptions of Theorem 3.3 are satisfied, then $\lim_{t\to\infty} k(t) =: k_\infty(w, y_0, k_0)$ exists and is finite. The linear system

$$\begin{cases} \dot{\tilde{y}} = D[k_\infty(w, y_0, k_0)K_{\sigma(k_\infty(w,y_0,k_0))}\tilde{y} - \mathcal{H}\tilde{y} - \tilde{w}], \\ \tilde{y}(0) = \tilde{y}_0 \in \mathbb{R}^m, \tilde{w} \in L^2(\mathbb{R}_+, \mathbb{R}^m) \end{cases} \tag{3.12}$$

is called the terminal system of the nonlinear closed-loop system given by (2.3) and (3.1). It is easy to see that (3.12) does not satisfy $\lim_{t\to\infty} \tilde{y}(t) = 0$ for arbitrary $(\tilde{w}, \tilde{y}_0) \in L^2(\mathbb{R}_+, \mathbb{R}^m) \times \mathbb{R}^m$. Indeed, consider the special case that $\mathcal{H} = 0$ and choose $w = 0, y_0 = 0$, and $k_0 = 0$ in (2.3) and (3.1). Since $k_\infty(0,0,0) = 0$, it follows that the solution $\tilde{y}$ of (3.12) is given by $\tilde{y}(t) = \tilde{y}_0 - D\int_0^t \tilde{w}(\tau)d\tau$, and hence $\tilde{y}(t)$ in general does not converge to 0 as $t \to \infty$. However, recent work of Townley [25] on adaptive stabilization of finite-dimensional systems leads us to the following conjecture.

Conjecture. For given $k_0 \in \mathbb{R}$ there exists an open and dense set $\mathcal{I}(k_0) \subset L^2(\mathbb{R}_+, \mathbb{R}^m) \times \mathbb{R}^m$ such that the terminal system (3.12) is stable in the sense that

$$\tilde{y} \in L^2 \cap L^\infty(\mathbb{R}_+, \mathbb{R}^m) \quad \text{and} \quad \lim_{t\to\infty} \tilde{y}(t) = 0 \quad \text{for all } (\tilde{w}, \tilde{y}_0) \in L^2(\mathbb{R}_+, \mathbb{R}^m) \times \mathbb{R}^m,$$

provided that $(w, y_0) \in \mathcal{I}(k_0)$.

4. Adaptive tracking and disturbance rejection. Consider the control scheme in Fig. 1, where the plant is described by (2.1) or, equivalently, by (2.3). The aim of this section is to construct a single controller, such that the closed-loop system asymptotically tracks a given reference trajectory r and asymptotically rejects a given disturbance signal d for all plants of the form (2.1). The signals r and d belong to prespecified vector spaces of functions that are defined as follows. Let $\rho_i, \delta_i \in \mathbb{R}[s]$ be monic polynomials, $1 \le i \le m$, and set $\rho := (\rho_1, \ldots, \rho_m)^T$ and $\delta := (\delta_1, \ldots, \delta_m)^T$. The admissible reference signals are given by

$$\mathcal{S}_\rho := \left\{ r : \mathbb{R}_+ \to \mathbb{R}^m \mid \rho_i\left(\frac{d}{dt}\right) r_i \equiv 0, i = 1, \ldots, m \right\},$$

while the disturbances d are supposed to belong to $\mathcal{S}_\delta + L^2(\mathbb{R}_+, \mathbb{R}^m)$, where $\mathcal{S}_\delta$ is defined as $\mathcal{S}_\rho$ with ρ_i replaced by δ_i. The well-known internal model principle from linear control theory (see e.g., Wonham [30, p. 203], and Vidyasagar [27, p. 294], for the finite-dimensional case and Francis [4], Callier and Desoer [2], and Curtain [3] for the infinite-dimensional case) suggests that the dynamics of the reference and disturbance signals should be replicated in the

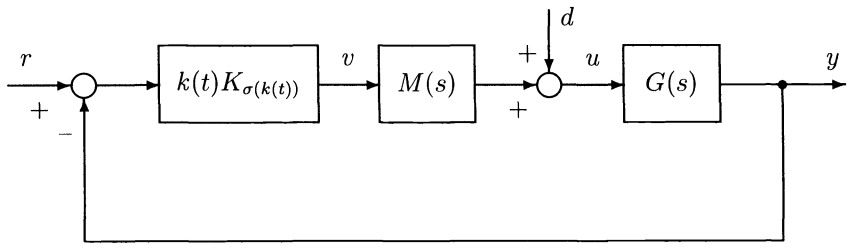


FIG. 2. *High-gain adaptive servomechanism.*

loop via a precompensator. To this end set

$$p(s) = \operatorname{lcm}(s, \rho_1(s), \dots, \rho_m(s), \delta_1(s), \dots, \delta_m(s)),$$

where we choose p to be monic. Moreover, let q be a monic polynomial that is Hurwitz and satisfies $\deg(q) = \deg(p)$. We define the precompensator $M(s)$ containing the internal model to be

$$M(s) = \frac{q(s)}{p(s)} I_m.$$

Note that by construction $M(s)$ contains an integrator. This is required for a purely technical reason: Without $M(s)$ having a pole in 0 we were not able to prove Theorem 4.1 below. Let G_M denote the precompensated plant, i.e., $G_M(s) = G(s)M(s)$. Now realize that, by (2.1),

$$G_M^{-1}(s) = \frac{p(s)}{q(s)}(sD^{-1} + H(s)) = sD^{-1} + H_M(s),$$

where

$$H_M(s) := s\left(\frac{p(s)}{q(s)} - 1\right) D^{-1} + \frac{p(s)}{q(s)} H(s) \tag{4.1}$$

belongs to $H^\infty(\mathbb{C}^{m\times m})$. The important point here is that the structural property (2.1) of the plant G remains invariant under precompensation by $M(s)$. The overall adaptive controller we shall investigate in the following is given by

$$\begin{cases} \hat{u}(s) = M(s)\hat{v}(s) + \hat{d}(s), \\ v(t) = k(t)K_{\sigma(k(t))}(r(t) - y(t)), \\ \dot{k}(t) = \|r(t) - y(t)\|^2, \qquad k(0) = k_0, \end{cases} \tag{4.2}$$

where σ and $K_1, \dots, K_N$ are defined as in §3 (cf. Fig. 2). Using the fact that the first equation of (4.2) can be written as

$$\hat{u}(s) = M(s)(\hat{v}(s) + M^{-1}(s)\hat{d}(s)),$$

setting $d_M(t) := \mathcal{L}^{-1}(M^{-1}\hat{d})(t)$ and $\mathcal{H}_M := \mathcal{L}^{-1}\mathcal{M}_{H_M}\mathcal{L}$,† we obtain the following time-domain description of the closed-loop system given by (2.3) and (4.2):

† As before, the unique causal extension of $\mathcal{H}_M$ to $LL^2(\mathbb{R}_+, \mathbb{R}^m)$ will be denoted by the same symbol $\mathcal{H}_M$.

(4.3)

$$\begin{cases} \dot{y}(t) &= D(v(t) + d_M(t) - (\mathcal{H}_M y)(t) - w_M(t)), \quad y(0) = y_0, \quad w_M \in L^2(\mathbb{R}_+, \mathbb{R}^m) \\ v(t) &= k(t) K_{\sigma(k(t))}(r(t) - y(t)) \\ \dot{k}(t) &= \|r(t) - y(t)\|^2, \qquad k(0) = k_0, \end{cases}$$

where as in §2 the term w_M takes account of non-zero initial conditions (cf. also §5).

We are now in the position to prove the main result of this paper, which shows that the controller (4.2) solves the servoproblem for all systems of the form (2.3).

THEOREM 4.1. *The solution (y, k) of the closed-loop system* (4.3) *exists on $\mathbb{R}_+$ and has the following properties*:

(i) $\lim_{t\to\infty} k(t)$ *exists and is finite*,
(ii) $y - r \in L^2(\mathbb{R}_+, \mathbb{R}^m) \cap L^\infty(\mathbb{R}_+, \mathbb{R}^m)$,
(iii) $\lim_{t\to\infty}(y(t) - r(t)) = 0$.

Proof. Rewriting the first equation of (4.3) as

$$\frac{d}{dt}(r - y) = -D(v + \mathcal{H}_M(r - y) + d_M - \mathcal{H}_M r - D^{-1}\dot{r} - w_M),$$

we see that (i)–(iii) will follow from Theorem 3.3, provided that the term $d_M - \mathcal{H}_M r - D^{-1}\dot{r}$ belongs to $L^2(\mathbb{R}_+, \mathbb{R}^m)$. It is easy to show that $d_M \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. Indeed, by definition we have

$$\hat{d}_M(s) = M^{-1}(s)\hat{d}(s) = M^{-1}(s)\hat{d}_1(s) + M^{-1}(s)\hat{d}_2(s), \tag{4.4}$$

where $d_1 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ and $d_2 \in \mathcal{S}_\delta$. Now, clearly we have

$$M^{-1}(s)\hat{d}_1(s) \in H^2(\mathbb{C}^m). \tag{4.5}$$

Moreover, since $p(d/dt)d_{2i} \equiv 0, 1 \le i \le m$ (where d_{2i} denotes the ith component of d_2), it follows that there exist polynomials $\beta_i \in \mathbb{R}[s]$ such that

$$\hat{d}_{2i}(s) = \frac{\beta_i(s)}{p(s)} \quad \text{and} \quad \deg(\beta_i) \le \deg(p) - 1 = \deg(q) - 1.$$

Therefore

$$M^{-1}(s)\hat{d}_2(s) = \left(\frac{\beta_1(s)}{q(s)}, \ldots, \frac{\beta_m(s)}{q(s)}\right)^T \in H^2(\mathbb{C}^m). \tag{4.6}$$

Combining (4.4)–(4.6) shows that $\hat{d}_M \in H^2(\mathbb{C}^m)$ and hence $d_M \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. It remains to show that $\mathcal{H}_M r + D^{-1}\dot{r} \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. This will be done in two steps.

Step 1. Suppose that $r(0) = 0$. Then we have

$$\mathcal{L}(\mathcal{H}_M r + D^{-1}\dot{r})(s) = H_M(s)\hat{r}(s) + sD^{-1}\hat{r}(s), \tag{4.7}$$

and moreover $\hat{r}(s) = [1/p(s)]\gamma(s)$, where $\gamma(s) := (\gamma_1(s), \ldots, \gamma_m(s))^T, \gamma_i \in \mathbb{R}[s]$, and $\deg(\gamma_i) \le \deg(p) - 2 = \deg(q) - 2$. Using (4.1) it follows from (4.7) that

$$\begin{aligned} \mathcal{L}(\mathcal{H}_M r + D^{-1}\dot{r})(s) &= s\left(\frac{p(s)}{q(s)} - 1\right) D^{-1}\frac{1}{p(s)}\gamma(s) \\ &\quad + \frac{p(s)}{q(s)} H(s)\frac{1}{p(s)}\gamma(s) + sD^{-1}\frac{1}{p(s)}\gamma(s) \\ &= D^{-1}\frac{s}{q(s)}\gamma(s) + H(s)\frac{1}{q(s)}\gamma(s) \in H^2(\mathbb{C}^m), \end{aligned}$$

since $H \in H^\infty(\mathbb{C}^{m\times m})$, $\deg(\gamma_i) \le \deg(q) - 2$ and q is Hurwitz. Hence we have shown that

$$\mathcal{H}_M r + D^{-1}\dot{r} \in L^2(\mathbb{R}_+, \mathbb{R}^m),$$

provided that $p(d/dt)r \equiv 0$ and $r(0) = 0$.

Step 2. Now suppose that $r(0) = r_0 \neq 0$. Define $z(t) := r(t) - \Theta(t)r_0$, where

$$\Theta(t) := \begin{cases} 0, & t < 0, \\ 1, & t \ge 0; \end{cases}$$

notice that

$$\mathcal{H}_M r + D^{-1}\dot{r} = \mathcal{H}_M z + D^{-1}\dot{z} + \mathcal{H}_M(\Theta r_0). \tag{4.8}$$

Since $p(0) = 0$, it follows that $p(d/dt)z \equiv 0$. Moreover $z(0) = 0$ and hence we obtain from Step 1 that $\mathcal{H}_M z + D^{-1}\dot{z} \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. Therefore, by (4.8) it remains to show that $\mathcal{H}_M(\Theta r_0) \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. To this end write $p(s) = s\tilde{p}(s), \tilde{p} \in \mathbb{R}[s]$, which is possible by assumption. Using (4.1) it follows that

$$\mathcal{L}(\mathcal{H}_M(\Theta r_0))(s) = H_M(s)\frac{1}{s}r_0 = \left(\frac{p(s)}{q(s)} - 1\right) D^{-1}r_0 + \frac{\tilde{p}(s)}{q(s)}H(s)r_0 \in H^2(\mathbb{C}^m)$$

and hence $\mathcal{H}_M(\Theta r_0) \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. □

5. Applications to retarded systems and integrodifferential convolution systems. In this section we show how retarded and integrodifferential convolution systems fit into the input-output set-up developed in §§3 and 4. We solve the adaptive servoproblem for these classes of systems under the assumption that the plant is minimum-phase and has invertible high-frequency gain. Moreover, it turns out that the internal variables of the plant and the precompensator remain bounded, provided that the reference signal is bounded.

5.1. Retarded systems. In the following we extend any function $F \in BV([a, b], \mathbb{R}^{n\times n})$ to the whole real axis by setting $F(t) = F(a)$ for $t < a$ and $F(t) = F(b)$ for $t > b$. Any measurable function $f : \Omega \to \mathbb{R}^n, \Omega \subset \mathbb{R}$, will be extended to the whole real axis by defining $f(t) = 0$ for $t \notin \Omega$. For $F = (F_{ij}) \in BV([0, h], \mathbb{R}^{n\times n})$ and $f = (f_1, \dots, f_n)^T, f_i \in LL^1(\mathbb{R}, \mathbb{R}), 1 \le i \le n$, we define

$$dF * f := \begin{pmatrix} \sum_{j=1}^n dF_{1j} * f_j \\ \vdots \\ \sum_{j=1}^n dF_{nj} * f_j \end{pmatrix},$$

where dF_{ij} denote the measure on $\mathbb{R}$ induced by F_{ij} and $dF_{ij} * f_j$ denotes the convolution of the measure dF_{ij} and the function f_j. If f is continuous on $[-h, \infty)$, then of course

$$(dF * f)(t) = \int_0^h dF(\tau)f(t - \tau) \quad \text{for } t \ge 0.$$

Consider the retarded system

$$\begin{aligned} \dot{x} &= dA * x + Bu, \\ y &= Cx, \\ x_{|_{[-h,0]}} &= x_0 \in C([-h, 0], \mathbb{R}^n), \end{aligned} \tag{5.1}$$

where $A \in BV([0,h],\mathbb{R}^{n\times n})$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{m\times n}$. We assume that

$$\det(CB) \neq 0 \tag{5.2}$$

and

$$\det \begin{pmatrix} sI - \hat{A}(s) & -B \\ C & 0 \end{pmatrix} \neq 0 \text{ for all } s \in \overline{\mathbb{C}}_+, \tag{5.3}$$

where $\hat{A}(s) := \int_0^h \exp(-s\tau)dA(\tau)$ denotes the Laplace–Stieltjes transform of A. The transfer function matrix $G(s)$ of (5.1) is given by

$$G(s) = C(sI - \hat{A}(s))^{-1}B.$$

Remark 5.1. As in the finite-dimensional case, we shall call (5.3) the minimum-phase condition. It can be shown that (5.3) holds if and only if the following three conditions hold:

(i) The transfer function matrix $G(s)$ has no zeros in $\overline{\mathbb{C}}_+$;

(ii) $rk(sI - \hat{A}(s), B) = n \quad$ for all $s \in \overline{\mathbb{C}}_+$;

(iii) $rk \begin{pmatrix} sI - \hat{A}(s) \\ C \end{pmatrix} = n \quad$ for all $s \in \overline{\mathbb{C}}_+$.

Let $\rho_i, \delta_i (i = 1,\dots,m), \rho, \delta, p$, and q be as in §4 and let

$$\begin{aligned} \dot{\xi} &= A_M\xi + B_M v, \quad \xi(0) = \xi_0 \in \mathbb{R}^l \\ z &= C_M\xi + I_m v \end{aligned} \tag{5.4}$$

be a stabilizable and detectable realization of $M(s) = [q(s)/p(s)]I_m$. We shall consider the closed-loop system given by (5.1), (5.4),

$$\begin{aligned} v(t) &= k(t)K_{\sigma(k(t))}(y(t) - r(t)), \\ \dot{k}(t) &= \|y(t) - r(t)\|^2, \quad k(0) = k_0 \in \mathbb{R}, \end{aligned} \tag{5.5}$$

and

$$u(t) = z(t) + d(t), \tag{5.6}$$

where $r \in \mathcal{S}_\rho, d \in \mathcal{S}_\delta + L^2(\mathbb{R}_+,\mathbb{R}^m)$ and $K_1,\dots,K_N$ and $\sigma : \mathbb{R} \to \{1,\dots,N\}$ are defined as in §3.

The following result shows that the universal adaptive controller presented in §4 achieves asymptotic tracking and disturbance rejection for the class of retarded systems satisfying (5.2) and (5.3).

THEOREM 5.2. *If* (5.2) *and* (5.3) *are satisfied, then for any* $x_0 \in C([-h,0],\mathbb{R}^n), \xi_0 \in \mathbb{R}^l, k_0 \in \mathbb{R}, r \in \mathcal{S}_\rho$, *and* $d \in \mathcal{S}_\delta + L^2(\mathbb{R}_+,\mathbb{R}^m)$, *the closed-loop system given by* (5.1) *and* (5.4)–(5.6) *has the following properties*:

(i) $\lim_{t\to\infty} k(t)$ *exists and is finite*;

(ii) $y - r \in L^2(\mathbb{R}_+,\mathbb{R}^m) \cap L^\infty(\mathbb{R}_+,\mathbb{R}^m)$;

(iii) $\lim_{t\to\infty}(y(t) - r(t)) = 0$;

(iv) $(x,\xi)^T \in L^\infty(\mathbb{R}_+,\mathbb{R}^{n+l})$ *provided* r *is bounded.*

Proof. First of all it follows from (5.2) and (5.3) that

$$G^{-1}(s) = s(CB)^{-1} + H(s), \tag{5.7}$$

where $H \in H^\infty(\mathbb{C}^{m\times m})$ (see Logemann and Mårtensson [13]), i.e., $G^{-1}(s)$ admits a decomposition of the form (2.1). We proceed in four steps.

Step 1. Recall from the proof of Theorem 4.1 that

$$d_M(t) = \mathcal{L}^{-1}(M^{-1}\hat{d})(t) \in L^2(\mathbb{R}_+, \mathbb{R}^m). \tag{5.8}$$

Defining

$$\chi(t) := \begin{cases} 0, & t = 0, \\ 1, & t \in (0, h], \end{cases}$$

and setting

$$A_{se}(\cdot) := \begin{pmatrix} A(\cdot) & \chi(\cdot)BC_M \\ 0 & \chi(\cdot)A_M \end{pmatrix}, \quad B_{se} := \begin{pmatrix} B \\ B_M \end{pmatrix},$$
$$C_{se} := (C, 0) \quad \text{and} \quad x_{se} := \begin{pmatrix} x \\ \xi \end{pmatrix},$$

the series connection of (5.4) followed by (5.1) in the presence of the disturbance d can be reformulated as follows:

$$\begin{aligned} \dot{x}_{se} &= dA_{se} * x_{se} + B_{se}(v + d_M), \\ y &= C_{se}x_{se} \\ x_{se}(t) &= \begin{pmatrix} x_0(t) \\ \xi_0 \end{pmatrix} \quad \text{for all } t \in [-h, 0]. \end{aligned} \tag{5.9}$$

It follows trivially from (5.2) that

$$\det(C_{se}B_{se}) \neq 0. \tag{5.10}$$

Moreover, since $q(s)$ is Hurwitz, it follows from the stabilizability and detectability of (5.4) that

$$\det \begin{pmatrix} sI - A_M & -B_M \\ C_M & I_m \end{pmatrix} \neq 0 \quad \text{for all } s \in \overline{\mathbb{C}}_+. \tag{5.11}$$

Realizing that

$$\begin{aligned} &\det \begin{pmatrix} sI - \hat{A}_{se}(s) & -B_{se} \\ C_{se} & 0 \end{pmatrix} \\ &\quad = \det(sI - \hat{A}(s)) \det(G(s)) \det(sI - A_M) \det(M(s)) \\ &\quad = \det \begin{pmatrix} sI - \hat{A}(s) & -B \\ C & 0 \end{pmatrix} \det \begin{pmatrix} sI - A_M & -B_M \\ C_M & I_m \end{pmatrix}, \end{aligned}$$

we obtain from (5.3) and (5.11)

$$\det \begin{pmatrix} sI - \hat{A}_{se}(s) & -B_{se} \\ C_{se} & 0 \end{pmatrix} \neq 0 \quad \text{for all } s \in \overline{\mathbb{C}}_+, \tag{5.12}$$

i.e., the series connection of (5.4) followed by (5.1) is minimum-phase.

Step 2. It follows from (5.10) that $\mathbb{R}^{n+l} = \ker C_{se} \oplus \operatorname{im} B_{se}$. Hence there exists a non-singular real transformation $P \in \mathbb{R}^{(n+l)\times(n+l)}$ such that

$$P^{-1}B_{se} = \begin{pmatrix} CB \\ 0 \end{pmatrix}, \quad C_{se}P = (I_m \quad 0).$$

It is useful to partition the matrix $P^{-1}A_{se}(\cdot)P$ as follows:

$$P^{-1}A_{se}(\cdot)P = \begin{pmatrix} A_{11}(\cdot) & A_{12}(\cdot) \\ A_{21}(\cdot) & A_{22}(\cdot) \end{pmatrix},$$

where $A_{11}(\cdot), A_{12}(\cdot), A_{21}(\cdot)$, and $A_{22}(\cdot)$ are matrices with entries in $BV([0,h],\mathbb{R})$ of size $m \times m, m \times (n+l-m), (n+l-m) \times m$ and $(n+l-m) \times (n+l-m)$, respectively. Setting $\eta_{se}(t) = P^{-1}x_{se}(t)$, it follows from (5.9) that

$$\begin{aligned} \dot{\eta}_{se} &= d(P^{-1}A_{se}P) * \eta_{se} + P^{-1}B_{se}(v + d_M), \\ y &= C_{se}P\eta_{se}, \\ \eta_{se}|_{[-h,0]} &= P^{-1}x_{se}|_{[-h,0]}. \end{aligned} \tag{5.13}$$

Since η_{se} can be written as $\eta_{se} = (y,\eta)^T$, it is clear that (5.13) can be decomposed as

$$\dot{y} = CBv_1, \tag{5.14}$$

$$\begin{aligned} \dot{\eta} &= dA_{22} * \eta + dA_{21} * v_2, \\ \gamma &= -(CB)^{-1}(dA_{12} * \eta + dA_{11} * v_2), \end{aligned} \tag{5.15}$$

$$v_1 = v + d_M - \gamma, \qquad v_2 = y \tag{5.16}$$

$$y|_{[-h,0]} = \eta_1, \qquad \eta|_{[-h,0]} = \eta_2, \tag{5.17}$$

where $(\eta_1,\eta_2) = \eta_{se}|_{[-h,0]}$ and in particular $\eta_1 = Cx_0$. Let $\eta(\eta_2,\eta_1,\omega)$ denote the solution of the retarded system (5.15) driven by the initial conditions $\eta|_{[-h,0]} = \eta_2, v_2|_{[-h,0]} = \eta_1$ and the input $v_2|_{[0,\infty)} = \omega \in LL^2(\mathbb{R}_+,\mathbb{R}^m)$. The corresponding output $\gamma(\eta_2,\eta_1,\omega)$ can be written in the form

$$\gamma(\eta_2,\eta_1,\omega) = \mathcal{K}\omega + \tilde{w}, \tag{5.18}$$

where

$$\mathcal{K}w = -(CB)^{-1}(dA_{12} * \eta(0,0,\omega) + dA_{11} * \omega) \tag{5.19}$$

and

$$\tilde{w} = -(CB)^{-1}(dA_{12} * \eta(\eta_2,\eta_1,0) + dA_{11} * \eta_1). \tag{5.20}$$

Step 3. We claim that the retarded system (5.15) is exponentially stable, which is equivalent to saying that $\det(sI - \hat{A}_{22}(s)) \neq 0$ for all $s \in \overline{\mathbb{C}}_+$, where $\hat{A}_{22}(s) = \int_0^h \exp(-s\tau)dA_{22}(\tau)$ (cf. Hale [6, p. 165]). It follows from the properties of P that

$$\det\begin{pmatrix} sI - \hat{A}_{se}(s) & -B_{se} \\ C_{se} & 0 \end{pmatrix} = \det\begin{pmatrix} sI - \hat{A}_{11}(s) & -\hat{A}_{12}(s) & -CB \\ -\hat{A}_{21}(s) & sI - \hat{A}_{22}(s) & 0 \\ I & 0 & 0 \end{pmatrix}.$$

Defining

$$T_1(s) := \begin{pmatrix} I & 0 & -(sI - \hat{A}_{11}(s)) \\ 0 & I & \hat{A}_{21}(s) \\ 0 & 0 & I \end{pmatrix}, \qquad T_2(s) := \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -(CB)^{-1}\hat{A}_{12}(s) & I \end{pmatrix},$$

we obtain

$$\begin{aligned}
&\det\begin{pmatrix} sI-\hat{A}_{se}(s) & -B_{se} \\ C_{se} & 0 \end{pmatrix} \\
&= \det\left\{ T_1(s)\begin{pmatrix} sI-\hat{A}_{11}(s) & -\hat{A}_{12}(s) & -CB \\ -\hat{A}_{21}(s) & sI-\hat{A}_{22}(s) & 0 \\ I & 0 & 0 \end{pmatrix} T_2(s) \right\} \\
&= \det\begin{pmatrix} 0 & 0 & -CB \\ 0 & sI-\hat{A}_{22}(s) & 0 \\ I & 0 & 0 \end{pmatrix} \\
&= (-1)^m \det(CB)\det(sI-\hat{A}_{22}(s)).
\end{aligned}$$

Hence $\det(sI-\hat{A}_{22}(s)) \neq 0$ for all $s \in \overline{\mathbb{C}}_+$ by the minimum-phase property (5.12).

Step 4. As a consequence of the exponential stability of the retarded system (5.15), the linear mapping $\mathcal{K}$ defined by (5.19) is bounded from $L^2(\mathbb{R}_+,\mathbb{R}^m)$ into itself and the function $\tilde{w}$ is in $L^2(\mathbb{R}_+,\mathbb{R}^m)$. Moreover, it is clear that the operator $\mathcal{K}$ is shift-invariant. The system given by (5.14)–(5.17) can be written as

$$\dot{y} = CB(v-\mathcal{K}y-w), \qquad y(0) = Cx_0(0), \tag{5.21}$$

where $w := \tilde{w} - d_M \in L^2(\mathbb{R}_+,\mathbb{R}^m)$. Let $\mathcal{K}$ be the unique element in $H^\infty(\mathbb{C}^{m\times m})$ such that $\mathcal{K} = \mathcal{L}^{-1}\mathcal{M}_K\mathcal{L}$. It is easy to see that K is of the form required for the application of Theorem 4.1, i.e.,

$$K(s) = s\left(\frac{p(s)}{q(s)} - 1\right)(CB)^{-1} + \frac{p(s)}{q(s)}H(s),$$

where $H(s)$ is given by (5.7). Statements (i)–(iii) follow now from Theorem 4.1. Finally, suppose that r is bounded. By statement (ii) this implies that y is bounded, and hence using the exponential stability of (5.15), we see that η is bounded. As a consequence $\eta_{se} = (\eta, y)^T$ is bounded, which in turn implies the boundedness of $x_{se} = (x,\xi)^T$. □

5.2. Integrodifferential convolution systems. Another interesting class of systems covered by Theorem 4.1 is the class of integrodifferential convolution systems. Consider the system

$$\begin{aligned}
\dot{x} &= A * x + Bu, \\
y &= Cx, \\
x(0) &= x_0 \in \mathbb{R}^n,
\end{aligned} \tag{5.22}$$

where $A \in M(\mathbb{R}_+,\mathbb{R}^{n\times n}), B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{m\times n}$. The Volterra integrodifferential system

$$\begin{aligned}
\dot{x}(t) &= A_0x(t) + \int_0^t A_1(t-\tau)x(\tau)d\tau + Bu(t), \\
y(t) &= Cx(t), \\
x(0) &= x_0 \in \mathbb{R}^n,
\end{aligned}$$

where $A_0 \in \mathbb{R}^{n\times n}$ and $A_1 \in L^1(\mathbb{R}_+,\mathbb{R}^{n\times n})$ is obviously a special case of (5.22). We assume that

$$\det(CB) \neq 0 \tag{5.23}$$

and

$$\det \begin{pmatrix} sI - \hat{A}(s) & -B \\ C & 0 \end{pmatrix} \neq 0 \quad \text{for all } s \in \overline{\mathbb{C}}_+, \tag{5.24}$$

where $\hat{A}(s) := \int_0^\infty \exp(-s\tau) dA(\tau)$.

THEOREM 5.3. *If* (5.23) *and* (5.24) *are satisfied, then for any* $x_0 \in \mathbb{R}^n, \xi_0 \in \mathbb{R}^l, k_0 \in \mathbb{R}$, $r \in \mathcal{S}_\rho$, *and* $d \in \mathcal{S}_\delta + L^2(\mathbb{R}_+, \mathbb{R}^m)$, *the closed-loop system given by* (5.22) *and* (5.4)–(5.6) *has the following properties*:

(i) $\lim_{t\to\infty} k(t)$ *exists and is finite*;
(ii) $y - r \in L^2(\mathbb{R}_+, \mathbb{R}^m) \cap L^\infty(\mathbb{R}_+, \mathbb{R}^m)$;
(iii) $\lim_{t\to\infty}(y(t)) - r(t)) = 0$;
(iv) $(x, \xi)^T \in L^\infty(\mathbb{R}_+, \mathbb{R}^{n+l})$, *provided* r *is bounded.*

Proof. Defining

$$A_{se} := \begin{pmatrix} A & \delta_0 B C_M \\ 0 & \delta_0 A_M \end{pmatrix}, \qquad B_{se} := \begin{pmatrix} B \\ B_M \end{pmatrix},$$
$$C_{se} := (C, 0) \quad \text{and} \quad x_{se} := \begin{pmatrix} x \\ \xi \end{pmatrix},$$

where δ_0 denotes the unit point mass at 0, the series connection of (5.4) followed by (5.22) in the presence of the disturbance d can be formulated as follows:

$$\begin{aligned} \dot{x}_{se} &= A_{se} * x_{se} + B_{se}(v + d_M), \\ y &= C_{se} x_{se}, \\ x_{se}(0) &= \begin{pmatrix} x_0 \\ \xi_0 \end{pmatrix}, \end{aligned} \tag{5.25}$$

where d_M is given by (5.8).

Using the same coordinate transformation P as in §5.1, it is clear that (5.25) can be written in the form

$$\dot{y} = CBv_1, \tag{5.26}$$

$$\begin{aligned} \dot{\eta} &= A_{22} * \eta + A_{21} * v_2, \\ \gamma &= -(CB)^{-1}(A_{12} * \eta + A_{11} * v_2); \end{aligned} \tag{5.27}$$

$$v_1 = v + d_M - \gamma, \qquad v_2 = y, \tag{5.28}$$

$$(y(0), \eta(0))^T = P^{-1} x_{se}(0), \tag{5.29}$$

where $(y, \eta)^T = P^{-1} x_{se}$ and the A_{ij} are bounded matrix-valued measures on $\mathbb{R}_+$. Let R denote the differential resolvent of the integrodifferential system (5.27), i.e., R is the unique solution of

$$\dot{R} = A_{22} * R, \qquad R(0) = I.$$

The solution η is then given by

$$\eta(t) = R(t)\eta(0) + (R * A_{21} * v_2)(t)$$

(see Gripenberg, Londen, and Staffans [5, p. 76]) and the output γ can be written in the form

$$\gamma = \mathcal{K}v_2 + \tilde{w},$$

where

$$\mathcal{K}v_2 = -(CB)^{-1}(A_{12} * R * A_{21} + A_{11}) * v_2 \tag{5.30}$$

and

$$\tilde{w} = -(CB)^{-1}(A_{12} * R)\eta(0). \tag{5.31}$$

Now we can show, as in §5.1, that

$$\det(sI - \hat{A}_{22}(s)) \neq 0 \quad \text{for all } s \in \overline{\mathbb{C}}_+, \tag{5.32}$$

where $\hat{A}_{22}(s) = \int_0^\infty \exp(-s\tau)dA_{22}(\tau)$, and hence R is integrable (see Gripenberg, Londen, and Staffans [5, p. 83]). It follows that the linear operator $\mathcal{K}$ defined by (5.30) is bounded from $L^2(\mathbb{R}_+, \mathbb{R}^m)$ into itself. Moreover it is trivial to show that $\mathcal{K}$ is shift-invariant. Since R is integrable we obtain from Gripenberg, Londen, and Staffans [5, p. 83], that the entries of R are square-integrable as well. Therefore the function $\tilde{w}$ defined by (5.31) is in $L^2(\mathbb{R}_+, \mathbb{R}^m)$. Finally it follows that the system (5.25) can be written as

$$\dot{y} = CB(v - \mathcal{K}y - w), \qquad y(0) = Cx_0,$$

where $w := \tilde{w} - d_M$ is in $L^2(\mathbb{R}_+, \mathbb{R}^m)$ (by (5.8)) and $\mathcal{K} : L^2(\mathbb{R}_+, \mathbb{R}^m) \to L^2(\mathbb{R}_+, \mathbb{R}^m)$ is linear bounded and shift-invariant. The claim now follows in exactly the same way as in the proof of Theorem 5.2. □

6. Conclusions. In this paper we have presented an input-output approach to the adaptive servoproblem for multivariable infinite-dimensional minimum-phase systems with invertible high-frequency gains. In particular, we have shown the following:

- The switching algorithm, introduced by Byrnes and Willems [1] in a finite-dimensional state-space set-up, stabilizes any infinite-dimensional plant belonging to the class of systems given by (2.1).
- The series interconnection of the Byrnes–Willems controller with a suitable precompensator solves the adaptive servoproblem for the class of systems satisfying (2.1).
- The input-output results obtained in §§3 and 4 apply to retarded systems and integro-differential convolution systems.

The adaptive control laws presented in §§3 and 4 give positive answers to feasibility and existence questions. They do not provide satisfying adaptive controllers from an engineer's point of view. However, the following comments show that the results of this paper might also be of some practical importance.

- It seems plausible that the technique in §4 (or variations thereof) can be used in order to obtain adaptive servomechanisms from various adaptive stabilization algorithms available in the literature.
- If the conjecture formulated in §3 turns out to be true, the high-gain switching algorithm can be used in order to identify a stabilizing linear controller or a linear servocompensator for the class of infinite-dimensional systems under consideration by a single simulation.

7. Appendix.

Proof of Lemma 3.2. The closed-loop system is given by

$$\begin{aligned} \dot{y}(t) &= D(k(t)K_{\sigma(k(t))}y(t) - (\mathcal{H}y)(t) - w(t)), \\ \dot{k}(t) &= \|y(t)\|^2, \qquad t \geq 0, \\ y(0) &= y_0, \qquad k(0) = k_0. \end{aligned} \tag{7.1}$$

Without loss of generality we may assume that $k_0 \geq 0$. The proof is divided into three steps.

Step 1. Existence and uniqueness on a "small" interval.

Consider equation (7.1) with $\sigma(k(t))$ replaced by $\sigma(k_0)$, i.e.,

$$\begin{aligned} \dot{y}(t) &= D(k(t)K_{\sigma(k_0)}y(t) - (\mathcal{H}y)(t) - w(t)), \\ \dot{k}(t) &= \|y(t)\|^2, \qquad t \geq 0, \\ y(0) &= y_0, \qquad k(0) = k_0 \end{aligned} \tag{7.2}$$

By Theorem 2.5, (7.2) has a unique absolutely continuous solution $(\tilde{y}, \tilde{k})$ on some interval $[0, T)$. Set $\tau(k_0) = \min\{\tau_i \mid \tau_i > k_0\}$ and let $T' \in (0, T)$ be such that $\tilde{k}(T') < \tau(k_0)$. Since $\sigma(\tilde{k}(t)) = \sigma(k_0)$ for all $t \in [0, T']$, it follows that $(\tilde{y}, \tilde{k})$ is the unique solution of (7.1) on $[0, T')$.

Step 2. Extended uniqueness.

Let (y_i, k_i) be solutions of (7.1) on $[0, T_j), i = 1, 2$. We claim that $(y_1, k_1) \equiv (y_2, k_2)$ on $[0, T)$, where $T := \min(T_1, T_2)$. Let us assume the contrary, i.e. there exists $t \in (0, T)$ for which $(y_1(t), k_1(t)) \neq (y_2(t), k_2(t))$. Defining

$$t^* := \inf\{t \in (0, T) \mid (y_1(t), k_1(t)) \neq (y_2(t), k_2(t))\},$$

it follows that $t^* > 0$ (by Step 1) and $(y_1(t^*), k_1(t^*)) = (y_2(t^*), k_2(t^*))$ (by continuity). Now set $k^* := k_1(t^*) = k_2(t^*)$ and realize that the initial-value problem

$$\begin{aligned} \dot{y}(t) &= D(k(t)K_{\sigma(k^*)}y(t) - (\mathcal{H}y)(t) - w(t)), \\ \dot{k}(t) &= \|y(t)\|^2, \qquad t \geq t^*, \\ y|_{[0,t^*]} &= y_1|_{[0,t^*]}, \qquad k|_{[0,t^*]} = k_1|_{[0,t^*]} \end{aligned} \tag{7.3}$$

is solved by (y_1, k_1) and (y_2, k_2) on $[0, t^* + \epsilon)$ for some sufficiently small $\epsilon > 0$. It follows from Theorem 2.5 that $(y_1(t), k_1(t)) = (y_2(t), k_2(t))$ for all $t \in [0, t^* + \epsilon)$, which contradicts the definition of t^*.

Step 3. Continuation of solutions.

Let $(\tilde{y}, \tilde{k})$ be a solution of (7.1) on $[0, T), 0 < T < \infty$. Assume that $(\tilde{y}, \tilde{k})$ is bounded. We claim that under these conditions the solution $(\tilde{y}, \tilde{k})$ can be continued to the right (beyond T). Since $\tilde{k}$ is bounded, continuous, and nondecreasing, it is clear that $\lim_{t \to T} \tilde{k}(t) =: \tilde{k}_T$ exists and is finite. As a consequence we have $\tilde{y} \in L^2(0, T; \mathbb{R}^m)$ and hence, by (7.1), $\dot{\tilde{y}} \in L^2(0, T; \mathbb{R}^m) \subset L^1(0, T; \mathbb{R}^m)$. Using the fact that

$$\tilde{y}(t) = y_0 + \int_0^t \dot{\tilde{y}}(\tau) d\tau,$$

it follows that $\lim_{t \to T} \tilde{y}(t) =: \tilde{y}_T$ exists and is finite. By Theorem 2.5 the initial-value problem

$$\begin{aligned} \dot{y}(t) &= D(k(t)K_{\sigma(\tilde{k}_T)}y(t) - (\mathcal{H}y(t) - w(t)), \\ \dot{k}(t) &= \|y(t)\|^2, \qquad t \geq T, \\ y(t) &= \begin{cases} \tilde{y}(t), & t \in [0, T), \\ \tilde{y}_T, & t = T, \end{cases} \qquad k(t) = \begin{cases} \tilde{k}(t), & t \in [0, T), \\ \tilde{k}_T, & t = T, \end{cases} \end{aligned}$$

has a unique absolutely continuous solution $(\bar{y}, \bar{k})$ on $[0, T+\epsilon)$ for some $\epsilon > 0$. Finally let $\delta \in (0, \epsilon)$ be such that

$$\bar{k}(T+\delta) < \min\{\tau_i \mid \tau_i > \tilde{k}_T\}.$$

Then $(\bar{y}, \bar{k})$ is a solution of (7.1) on $[0, T+\delta)$ extending the solution $(\tilde{y}, \tilde{k})$. □

Acknowledgments. A substantial part of the research for this paper was done while H. Logemann was visiting the Department of Mathematics and the School of Engineering at the University of Exeter. A. Ilchmann was supported by the Deutsche Forschungsgemeinschaft and the University of Exeter.

REFERENCES

[1] C. I. Byrnes and J. C. Willems, *Adaptive stabilization of multivariable linear systems*, Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, NV, 1984, pp. 1574–1577.

[2] F. M. Callier and C. A. Desoer, *Stabilization, tracking and disturbance rejection in multivariable convolution systems*, Ann. Soc. Sci. Bruxelles, 94 (1980), pp. 7–51.

[3] R. F. Curtain, *Tracking and regulation for distributed parameter systems*, Matematica Aplicada E Computacional, 2 (1983), pp. 199–218.

[4] B. A. Francis, *The multivariable servomechanism problem from the input-output viewpoint*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 322–328.

[5] G. Gripenberg, S.-O. Londen, and O. Staffans, *Volterra Integral and Functional Equations*, Cambridge University Press, Cambridge, UK, 1990.

[6] J. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[7] C. J. Harris and J. M. E. Valenca, *The Stability of Input-Output Dynamical Systems*, Academic Press, London, 1983.

[8] U. Helmke, D. Prätzel-Wolters, and S. Schmid, *Sufficient conditions for adaptive stabilization and tracking*, Bericht 38 der Arbeitsgruppe Technomathematik, Fachbereich Mathematik, Universität Kaiserslautern, Kaiserslautern, Germany, 1989.

[9] ———, *Adaptive tracking for scalar minimum-phase systems*, in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Boston, 1990, pp. 101–117.

[10] D. Hinrichsen and A. J. Pritchard, *Destabilization by output feedback*, Differential Integral Equations, 5 (1992), pp. 357–386.

[11] A. Ilchmann and H. Logemann, *High-gain adaptive stabilization of multivariable systems—revisited*, Systems Control Lett., 18 (1992), pp. 355–364.

[12] H. Logemann, *Funktionentheoretische Methoden in der Regelungstheorie undendlichdimensionaler Systeme*, Ph.D. thesis, Institut für Dynamische Systeme, Universität Bremen, Bremen, Germany, 1986 (Report 156).

[13] H. Logemann and B. Mårtensson, *Adaptive stabilization of infinite-dimensional systems*, IEEE Trans. Automat. Control, AC-37 (1992), pp. 1869–1883.

[14] H. Logemann and D. H. Owens, *Input-output theory of high-gain adaptive stabilization of infinite-dimensional systems with non-linearities*, Internat. J. Adaptive Control Signal Processing, 2 (1988), pp. 193–216.

[15] H. Logemann and H. Zwart, *On robust PI-control of infinite-dimensional systems*, SIAM J. Control Optim., 30 (1992), pp. 573–593.

[16] I. Mareels, *A simple selftuning controller for stably invertible systems*, Systems Control Lett., 4 (1984), pp. 5–16.

[17] B. Mårtensson, *Adaptive stabilization*, Ph.D. thesis, Lund Institute of Technology, Dept. of Automatic Control, Lund, Sweden, 1986.

[18] ———, *Adaptive stabilization of multivariable systems*, Contemp. Math., 68 (1987), pp. 191–225.

[19] ———, *Adaptive stabilization without high-gain*, in Identification and Adaptive Control, C. I. Byrnes and A. Kurzmansky, eds., Springer-Verlag, Berlin, 1988, pp. 226–238.

[20] D. E. Miller and E. J. Davison, *The self-tuning robust servomechanism problem*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 511–523.

[21] ———, *Adaptive control of a family of plants*, in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Boston, 1990, pp. 197–219.

[22] A. S. Morse, *A $4(n+1)$-dimensional model reference adaptive stabilizer for any relative degree one or two, minimum-phase system of dimension n or less*, Automatica, 23 (1987), pp. 123–125.

[23] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1974.

[24] E. P. Ryan, *Adaptive stabilization of multi-input nonlinear systems*, Int. J. Robust Nonlinear Control, 3 (1993), pp. 169–181.

[25] S. B. TOWNLEY, *Topological aspects of universal adaptive stabilization*, preprint, 1992.
[26] M. VIDYASAGAR, *A note on time-invariance and causality*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 929–931.
[27] ———, *Control Systems Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[28] M. VIDYASAGAR, H. SCHNEIDER, AND B. A. FRANCIS, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880–894.
[29] G. WEISS, *Representation of shift invariant operators on L^2 by H^∞ transfer functions: an elementary proof, a generalization to L^p and a counterexample for L^∞*, Math. Control Signals Systems, 4 (1991), pp. 193–203.
[30] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 2nd, ed., Springer-Verlag, New York, 1979.
[31] X.-J. ZHU, *A finite spectrum unmixing set for $GL(3, \mathbb{R})$*, in Computation and Control, K. Bowers and J. Lund, eds., Birkhäuser, Boston, 1989, pp. 403–410.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 937–951, July 1994

002

MINIMAX-OPTIMAL STRATEGIES FOR THE BEST-CHOICE PROBLEM WHEN A BOUND IS KNOWN FOR THE EXPECTED NUMBER OF OBJECTS*

T. P. HILL† AND D. P. KENNEDY‡

Abstract. For the best-choice (or secretary) problem with an unknown number N of objects, minimax-optimal strategies for the observer and minimax distributions for N are derived under the assumption that N is a random variable with expected value at most M, where M is known. The solution is derived as a special case of the situation where N is constrained by $Ef(N) \leq M$, where f is increasing with $f(i) - f(i-1)$ convex.

Key words. best-choice problem; secretary problem; minimax strategies; optimal stopping; convexity; Lagrangian; Lagrange multiplier

AMS subject classification. 60G40

1. Introduction.

In the classical best-choice (or secretary) problem a known fixed number, n, of rankable objects is presented one by one in random order (all $n!$ possible orderings being equally likely). As each object is presented, the observer must either select it and stop observing or reject it and continue observing. He may never return to a previously rejected object, and his decision to stop must be based solely on the relative ranks of the objects he has observed so far. The goal is to maximize the probability that the best object is selected. For a history and review of the literature of this problem and its numerous variants the reader is referred to Freeman (1983) and Ferguson (1989). In the best-choice setting, the optimal strategy for the observer is to view k_n objects without selecting and subsequently to take the first object, if any, better than all its predecessors, where $k_1 = 0$ and for $n > 1$, k_n is the unique positive integer satisfying

$$\sum_{i=k_n}^{n-1} \frac{1}{i} \geq 1 > \sum_{i=k_n+1}^{n-1} \frac{1}{i}.$$

If the number of objects is not known, but is a random variable taking values in the positive integers, then minimax-optimal strategies for selecting the best object are known for several situations (cf. Freeman (1983), Ferguson (1989)). For example, Presman and Sonin (1972) derived the optimal stop rules when the distribution of N is known, and Hill and Krengel (1991) found minimax-optimal stop rules (and distributions) when N has unknown distribution but known upper bound n. It is the purpose of this paper to derive the analogous minimax-optimal strategies when N again has unknown distribution, but has expectation at most M, where M is known. Since the arguments in this case generalize easily to the constraint $Ef(N) \leq M$, where f is a known positive function for which $f(i) - f(i-1)$ is nondecreasing and convex, the proofs will be given in the more general setting. The reader may want to keep in mind the natural case $f(i) \equiv i$, which corresponds to the expected-value constraint.

In the (zero-sum, two-person) game-theoretic interpretation of this problem there are two players, a controller P and an observer Q. Given $M > 0$ and a constraint function f, player P first picks a distribution for the number of objects, subject to the constraint $E(f(N)) \leq M$, and then the actual number N of objects to be presented to Q is chosen randomly according to this distribution. Then, knowing only the constraint (and not N itself), player Q begins

* Received by the editors July 23, 1992; accepted for publication (in revised form) January 7, 1993.

† Department of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332. The work of this author was partially supported by National Science Foundation grant DMS 89-01267.

‡ Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, United Kingdom.

his observation-selection of the objects and receives one dollar from player P if the object he selects is the best of the N objects and pays player Q one dollar if it is not the best.

Formally, the strategies available to the two players are given as follows. For $M \geq f(1)$, the set of allowable strategies for player P is

$$\mathcal{P}_M = \left\{ \mathbf{p} = (p_1, p_2, \cdots) : p_j \geq 0, \sum_{j=1}^{\infty} p_j = 1, \sum_{j=1}^{\infty} f(j) p_j \leq M \right\}$$

(the set of distributions for N for which $Ef(N) \leq M$), and

$$\mathcal{Q} = \{ \mathbf{q} = (q_1, q_2, \cdots) : 0 \leq q_j \leq 1 \}$$

is the set of allowable strategies for player Q, where if strategy $\mathbf{q}$ is used player Q stops at object j with probability q_j (independently of the rest of the process) if object j is the best so far. If the strategies $\mathbf{p}$, $\mathbf{q}$ are used by the respective players, the pay-off function $V(\mathbf{p}, \mathbf{q})$, which is the probability that player Q selects the best object, is given (cf. Hill and Krengel (1991)) by

$$V(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{\infty} \frac{p_j}{j} \sum_{i=1}^{j} q_i \prod_{m=1}^{i-1} \left(1 - \frac{q_m}{m}\right). \tag{1}$$

For each value of $M \geq f(1)$, the aim is to derive minimax-optimal strategies $\mathbf{p}_M \in \mathcal{P}_M$, $\mathbf{q}_M \in \mathcal{Q}$ satisfying

$$V(\mathbf{p}_M, \mathbf{q}) \leq V(\mathbf{p}_M, \mathbf{q}_M) \leq V(\mathbf{p}, \mathbf{q}_M) \quad \text{for all } \mathbf{p} \in \mathcal{P}_M, \mathbf{q} \in \mathcal{Q}. \tag{2}$$

Example 1.1. As will follow from the main results below, for the optimal strategies when N is a random variable with expected value at most 3, the optimal strategy for the observer is $(\frac{7}{13}, 1, 1, \cdots)$, i.e., stop with the first object with probability $\frac{7}{13}$, and otherwise stop with the first object thereafter, if any, that is better than any previously seen. Using this strategy, the best object will be selected with probability at least $\frac{16}{39}$ no matter what the distribution of N is, provided its expectation is at most 3. Conversely, the optimal P-strategy (worst-case distribution) for this case is $(\frac{2}{13}, 0, \frac{7}{13}, \frac{4}{13}, 0, 0, \cdots)$, i.e., there is only the one object with probability $\frac{2}{13}$, exactly three objects with probability $\frac{7}{13}$, and exactly four objects with probability $\frac{4}{13}$; and against this distribution no stop rule will select the best object with probability exceeding $\frac{16}{39}$.

That the optimal Q-strategy is monotonic is intuitive (since if it is good to stop at time j with the best object seen so far, it is even better to stop at later times if that object is the best yet seen), but that the optimal P-strategy typically (as in Example 1.1) places mass on *two* large numbers seems surprising, and it is never the case for the uniformly bounded problem. In general, the optimal value is a complicated piecewise linear function of the form $\lambda + \mu M$ for appropriate λ and μ. It will be seen that there are real numbers $a_1 < a_2 < \cdots$, such that the minimax-optimal $\mathbf{q}_M$ is constant over each interval $a_i \leq M \leq a_{i+1}$, while the minimax-optimal $\mathbf{p}_M$ is linear in M in the interval and is thus a convex combination of the distributions that are optimal at the end points $M = a_i$ and $M = a_{i+1}$. The next example, which identifies values and optimal strategies for an interval of values of M (including the special case $M = 3$ of Example 1.1), shows typical behavior of the optimal strategies and value as M varies.

Example 1.2. As will follow from the main results below, for the expected-value constraint $f(i) \equiv i$, the optimal strategies and value for $19/7 \leq M \leq 317/75$ are as follows.

If $19/7 \le M < 101/29$:

$$\mathbf{q}_M = \left(\frac{7}{13}, 1, 1, \cdots\right);$$
$$\mathbf{p}_M = a_M \left(\frac{1}{7}, 0, \frac{6}{7}, 0, 0, \cdots\right) + (1 - a_M) \left(\frac{5}{29}, 0, 0, \frac{24}{29}, 0, 0, \cdots\right),$$

where $a_M = 7(101 - 29M)/156$;

$$V(\mathbf{p}_M, \mathbf{q}_M) = (47 - 5M)/78.$$

If $101/29 \le M < 69/17$:

$$\mathbf{q}_M = \left(\frac{105}{213}, 1, 1, \cdots\right);$$
$$\mathbf{p}_M = b_M \left(\frac{5}{29}, 0, 0, \frac{24}{29}, 0, \cdots\right) + (1 - b_M) \left(\frac{3}{17}, 0, 0, \frac{4}{17}, \frac{10}{17}, 0, \cdots\right),$$

where $b_M = 29(69 - 17M)/284$;

$$V(\mathbf{p}_M, \mathbf{q}_M) = (459 - 39M)/852.$$

If $69/17 \le M < 317/75$:

$$\mathbf{q}_M = \left(\frac{50}{107}, \frac{14}{19}, 1, 1, \cdots\right);$$
$$\mathbf{p}_M = c_M \left(\frac{3}{17}, 0, 0, \frac{4}{17}, \frac{10}{17}, 0, 0, \cdots\right) + (1 - c_M) \left(\frac{13}{75}, \frac{2}{75}, 0, 0, \frac{60}{75}, 0, \cdots\right),$$

where $c_M = 17(317 - 75M)/214$;

$$V(\mathbf{p}_M, \mathbf{q}_M) = (54 - 4M)/107.$$

Although the statements of the main results in this article are probabilistic in nature, the proofs are primarily optimization-theoretic. Since general optimization theory saddle-point theorems do not seem to yield a direct solution to the problem formulated here, optimization arguments using a Lagrangian, but heavily based on ad hoc convexity tools, have been developed. In principle, one could use the same techniques to handle a larger class of constraint functions such as those reflecting known bounds on means *and* variances (or several other moments), but this would involve examination of the many cases corresponding to criticality of the various constraints and is not done here.

The organization of the paper is as follows. Section 2 introduces notation, important parameters, and the value of the game and establishes a number of useful identities and inequalities, the proofs of which may be skipped at first reading. Section 3 identifies the minimax-optimal strategy for the observer player Q, which is obtained by solving for the coefficient of p_j/j in a Lagrangian, and §4 builds on these results to establish the optimal (worst-case) distribution for N (i.e., the minimax-optimal strategy for the controller player P), and summarizes all the results in the main theorem, Theorem 4.3.

2. Notation, preliminaries and the value of the game. The first lemma records some easy convexity results; the proof is left to the reader.

LEMMA 2.1. *Suppose* $g : \mathbb{N} \to (0, \infty)$. *Then*

(i) g *is convex (respectively, strictly convex) if and only if* $g(i) - g(i-1)$ *is nondecreasing (increasing) in* i;

(ii) *if* g *is convex (respectively, strictly convex), then* $\sum_{j=k+1}^{i} g(j)/(i-k)$ *is convex (strictly convex) in* $i > k$ *for each* $k \geq 0$.

Basic assumption. Throughout this paper, $f(0) = 0$ and $f : \mathbb{N} \to (0, \infty)$ and

$$f(i) - f(i-1) \quad \text{is nondecreasing and convex,} \tag{3}$$

the canonical example being $f(i) \equiv i$.

Let $s_0 = 0$, $s_k = \sum_{j=1}^{k} 1/j$ for $k \geq 1$ and for $1 \leq k < n$, set $s_k^n = s_n - s_k$ and $F_k^n = nf(n) - kf(k)$.

LEMMA 2.2. *For all such* f,

(i) $f(i)$ *is increasing and convex on* $i \in \mathbb{N}$;

(ii) $if(i)$ *is increasing and strictly convex on* $i \in \mathbb{N}$;

(iii) $if(i) - (i-1)f(i-1)$ *is increasing and convex on* $i \in \mathbb{N}$;

(iv) $s_k^n/(n-k)$ *is decreasing and strictly convex in* $n > k$;

(v) $F_k^n/(n-k)$ *is increasing and convex in* $n > k$;

(vi) F_k^n/s_{k-1}^{n-1} *is increasing in* $n > k$,

(vii) $\frac{F_k^{n+1}}{n-k+1} > \frac{F_k^n}{n-k} \geq F_k^{k+1} > f(k) > (\sum_{j=1}^{k-1} \frac{f(j)}{j+1} + f(k))/s_k$.

Proof. The proof is routine, using (3), Lemma 2.1 and the definitions of s_k^n and F_k^n. □

The next objective will be to define some basic parameters that play a central role in the main results of this paper and to establish some useful inequalities and equalities interrelating these parameters.

For $1 \leq k < n$, define $\alpha_{n,k} > 0$, $\overline{\alpha}_{n,k} > 0$, $m_{n,k} > 0$, $\overline{m}_{n,k} > 0$, $\lambda_{n,k} > 0$, and $\mu_{n,k} < 0$, as follows:

$$\frac{1}{\alpha_{n,k}} = s_k + \frac{n-k}{ks_{k-1}^{n-1}}, \qquad \frac{1}{\overline{\alpha}_{n,k}} = s_k + \frac{n-k+n(1-s_k^{n-1})}{k+1}, \tag{4}$$

$$\begin{aligned} \frac{m_{n,k}}{\alpha_{n,k}} &= \sum_{j=1}^{k-1} \frac{f(j)}{j+1} + f(k) + \frac{F_k^n}{ks_{k-1}^{n-1}}, \\ \frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}} &= \sum_{j=1}^{k-1} \frac{f(j)}{j+1} + f(k) + \frac{F_k^n + nF_n^{n+1}(1-s_k^{n-1})}{k+1}, \end{aligned} \tag{5}$$

$$\lambda_{n,k} = \frac{\alpha_{n,k}m_{n+1,k} - \alpha_{n+1,k}m_{n,k}}{m_{n+1,k} - m_{n,k}}, \qquad \mu_{n,k} = \frac{\alpha_{n+1,k} - \alpha_{n,k}}{m_{n+1,k} - m_{n,k}}. \tag{6}$$

Note that $\alpha_{n,k}$ and $\overline{\alpha}_{n,k}$ do not depend on f. Using these parameters, the value of the game $\mathbf{v}_M = V(\mathbf{p}_M, \mathbf{q}_M)$ appearing in (2) can now be stated (although proof that it is indeed the value is the subject of the subsequent sections).

Recall that k_n is the optimal cutoff value for the classical secretary problem with n objects and so $s_{k_n-1}^{n-1} \geq 1 > s_{k_n}^{n-1}$ and $k_n \leq k_{n+1} \leq k_n + 1$. Set $m_n = m_{n,k_n}$, $\overline{m}_n = \overline{m}_{n,k_n}$, $\alpha_n = \alpha_{n,k_n}$, and $\overline{\alpha}_n = \overline{\alpha}_{n,k_n}$.

DEFINITION. *For all* $M > f(1)$,

$$v_M = \begin{cases} \lambda_{n,k_n} + \mu_{n,k_n} M & \text{if } m_n \le M < m_{n+1} \text{ and } k_{n+1} = k_n, \text{ or} \\ & \text{if } m_n \le M < \overline{m}_n \text{ and } k_{n+1} = k_n + 1, \\ \lambda_{n,k_{n+1}} + \mu_{n,k_{n+1}} M & \text{if } \overline{m}_n \le M < m_{n+1} \text{ and } k_{n+1} = k_n + 1. \end{cases} \tag{7}$$

Notice that $v_{m_n} = \alpha_n$, *and* $v_{\overline{m}_n} = \overline{\alpha}_n$ *when* $k_{n+1} = k_n + 1$.

Example 2.3. For the canonical expected-value case $f(i) \equiv i$,

$$(m_3, m_4, \overline{m}_4, m_5) = \left(\frac{19}{7}, \frac{101}{29}, \frac{69}{17}, \frac{317}{75}\right),$$

$$(\alpha_{3,1}, \alpha_{4,1}, \alpha_{5,1}, \alpha_{4,2}, \alpha_{5,2}) = \left(\frac{3}{7}, \frac{11}{29}, \frac{25}{73}, \frac{10}{27}, \frac{26}{75}\right),$$

$$(m_{3,1}, m_{4,1}, m_{5,1}, m_{4,2}, m_{5,2}) = \left(\frac{19}{7}, \frac{101}{29}, \frac{313}{73}, \frac{97}{27}, \frac{317}{75}\right),$$

$$(\lambda_{3,1}, \lambda_{4,1}, \lambda_{4,2}) = \left(\frac{47}{78}, \frac{459}{852}, \frac{54}{107}\right), \quad \text{and}$$

$$(\mu_{3,1}, \mu_{4,1}, \mu_{4,2}) = \left(\frac{-5}{78}, \frac{-39}{852}, \frac{-4}{107}\right),$$

and together with (7) these yield the value $v_M = V(\mathbf{p}_M, \mathbf{q}_M)$ in Example 1.2. □

The next lemma establishes some useful inequalities.

LEMMA 2.4. *For* $n > 1$,

(i) *for fixed* k, $\alpha_{n,k}$ *is decreasing in* $n > k$;

(ii) *for fixed* k, $m_{n,k}/\alpha_{n,k}$ *and* $m_{n,k}$ *are increasing in* $n > k$;

(iii) *for fixed* n, $\alpha_{n,k}$ *is maximized at* $k = k_n$; $\alpha_{n,k}$ *is increasing in* k, $1 \le k \le k_n$, *and decreasing in* k, $k_n \le k \le n$;

(iv) *for fixed* n, $m_{n,k}/\alpha_{n,k}$ *and* $m_{n,k}$ *are minimized at* $k = k_n$; *they are decreasing in* k, $1 \le k \le k_n$ *and increasing in* k, $k_n \le k \le n$;

(v) $\overline{\alpha}_n < \alpha_n < \alpha_{n-1}$;

(vi) $\overline{m}_n > m_n > m_{n-1}$;

(vii) $\overline{\alpha}_n \ge \alpha_{n+1}$ *when* $k_{n+1} = k_n + 1$;

(viii) $m_{n,k_{n+1}} \le \overline{m}_n \le m_{n+1} < m_{n+1,k_n}$ *when* $k_{n+1} = k_n + 1$; *and*

(ix) $\lambda_{n,k} > 0$ *and* $\mu_{n,k} < 0$.

Proof. (i) By Lemma 2.2 (iv),

$$1/\alpha_{n,k} = s_k + (n-k)/(ks_{k-1}^{n-1}) < s_k + (n-k+1)/(ks_{k-1}^{n}) = 1/\alpha_{n+1,k}.$$

(ii) By Lemma 2.2 (vi),

$$\frac{m_{n+1,k}}{\alpha_{n+1,k}} - \frac{m_{n,k}}{\alpha_{n,k}} = \frac{F_k^{n+1}}{ks_{k-1}^{n}} - \frac{F_k^{n}}{ks_{k-1}^{n-1}} > 0.$$

Also, (i), (4), and Lemma 2.2 (vii) give

$$m_{n+1,k} - m_{n,k} = (\alpha_{n,k} - \alpha_{n+1,k}) \left(\frac{s_k F_k^n}{n-k} - \sum_{j=1}^{k-1} \frac{f(j)}{j+1} - f(k) \right) + \left(\frac{F_k^{n+1}}{n-k+1} - \frac{F_k^n}{n-k} \right) (1 - s_k \alpha_{n+1,k}) > 0.$$

(iii) Observing that $n-k-1 > ks_k^{n-1}$, it may be seen that the difference

$$\frac{1}{\alpha_{n,k+1}} - \frac{1}{\alpha_{n,k}} = \frac{(s_k^{n-1}-1)(1-(n-k-1)/(ks_k^{n-1}))}{(k+1)s_{k-1}^{n-1}}$$

is ≤ 0 or > 0 according as $k < k_n$ or $k \geq k_n$.

(iv) By Lemma 2.2 (ii), $F_{k+1}^n > (n-k-1)F_k^{k+1}$ and since $n-k-1 > ks_k^{n-1}$,

$$\frac{F_{k+1}^n}{ks_k^{n-1}} = \left(\frac{F_{k+1}^n}{n-k-1}\right)\left(\frac{n-k-1}{ks_k^{n-1}}\right) > F_k^{k+1}.$$

It follows that

$$\frac{m_{n,k+1}}{\alpha_{n,k+1}} - \frac{m_{n,k}}{\alpha_{n,k}} = \frac{(s_k^{n-1}-1)}{(k+1)s_{k-1}^{n-1}}\left(F_k^{k+1} - \frac{F_{k+1}^n}{ks_k^{n-1}}\right)$$

is ≤ 0 or > 0 according as $k < k_n$ or $k \geq k_n$. This last equality implies

$$\begin{aligned} m_{n,k+1} - m_{n,k} &= \alpha_{n,k+1}\left[\frac{m_{n,k+1}}{\alpha_{n,k+1}} - \frac{m_{n,k}}{\alpha_{n,k}} + m_{n,k}\left(\frac{1}{\alpha_{n,k}} - \frac{1}{\alpha_{n,k+1}}\right)\right] \\ &= \frac{\alpha_{n,k+1}(s_k^{n-1}-1)}{(k+1)s_{k-1}^{n-1}}\left[F_k^{k+1} - m_{n,k} + \left(\frac{(n-k-1)m_{n,k} - F_{k+1}^n}{ks_k^{n-1}}\right)\right]. \end{aligned}$$

From (4), (5), and Lemma 2.2 (vii) note that

$$\frac{m_{n,k}}{\alpha_{n,k}} < s_k f(k) + \frac{F_k^n}{ks_{k-1}^{n-1}} = s_k\left(f(k) - \frac{F_k^n}{n-k}\right) + \frac{F_k^n}{(n-k)\alpha_{n,k}} < \frac{F_k^n}{(n-k)\alpha_{n,k}};$$

hence

$$m_{n,k} < \frac{F_k^n}{n-k} < \frac{F_{k+1}^n}{n-k-1} < F_{n-1}^n. \tag{8}$$

Since $n-k-1 > ks_k^{n-1}$, it follows that

$$F_k^{k+1} - m_{n,k} + \left(\frac{(n-k-1)m_{n,k} - F_{k+1}^n}{ks_k^{n-1}}\right) < F_k^{k+1} - \frac{F_{k+1}^n}{n-k-1} < 0,$$

which implies that $m_{n,k+1} \leq m_{n,k}$ or $m_{n,k+1} > m_{n,k}$ according as $k < k_n$ or $k \geq k_n$.

(v) Note that $ns_{k-1}^{n-1} > n-k$. Since from (4),

$$\frac{1}{\overline{\alpha}_{n,k}} - \frac{1}{\alpha_{n,k}} = \frac{(1-s_k^{n-1})(ns_{k-1}^{n-1} - (n-k))}{(k+1)s_{k-1}^{n-1}},$$

putting $k = k_n$ and recalling the fact that $1 > s_{k_n}^{n-1}$ gives the result.

(vi) From (5),

$$\frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}} - \frac{m_{n,k}}{\alpha_{n,k}} = \frac{(1-s_k^{n-1})}{k+1}\left(nF_n^{n+1} - \frac{F_k^n}{s_{k-1}^{n-1}}\right).$$

Hence

$$\overline{m}_{n,k} - m_{n,k} = \overline{\alpha}_{n,k}\left[\frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}} - \frac{m_{n,k}}{\alpha_{n,k}} + m_{n,k}\left(\frac{1}{\alpha_{n,k}} - \frac{1}{\overline{\alpha}_{n,k}}\right)\right]$$
$$= \frac{\overline{\alpha}_{n,k}(1-s_k^{n-1})}{k+1}\left[n(F_n^{n+1} - m_{n,k}) - \frac{F_k^n - (n-k)m_{n,k}}{s_{k-1}^{n-1}}\right].$$

Again using the inequality $s_{k-1}^{n-1} > (n-k)/n$ and (8),

$$n(F_n^{n+1} - m_{n,k}) - \frac{F_k^n - (n-k)m_{n,k}}{s_{k-1}^{n-1}} > n\left(F_n^{n+1} - \frac{F_k^n}{n-k}\right) > 0,$$

whence $\overline{m}_{n,k} - m_{n,k}$ is >0 or ≤ 0 according as $1 > s_k^{n-1}$ or $1 \leq s_k^{n-1}$. From (ii) and (iv) above, $m_n = m_{n,k_n} \leq m_{n,k_{n+1}} < m_{n+1,k_{n+1}} = m_{n+1}$.

(vii) Similarly,

$$\frac{1}{\overline{\alpha}_{n,k}} - \frac{1}{\alpha_{n+1,k+1}} = \frac{(1-s_k^n)(ns_k^n - (n-k))}{(k+1)s_k^n},$$

which is ≤ 0 or >0 according as $k < k_{n+1}$ or $k \geq k_{n+1}$; taking $k = k_n < k_{n+1}$ gives the result.

(viii) Note that

$$\frac{m_{n+1,k+1}}{\alpha_{n+1,k+1}} - \frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}} = \frac{(1-s_k^n)(F_{k+1}^{n+1} - nF_n^{n+1}s_k^n)}{(k+1)s_k^n},$$

and it follows using the expression in the proof of (vii) above that

$$m_{n+1,k+1} - \overline{m}_{n,k} = \overline{\alpha}_{n,k}\left[\frac{m_{n+1,k+1}}{\alpha_{n+1,k+1}} - \frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}} + m_{n+1,k+1}\left(\frac{1}{\overline{\alpha}_{n,k}} - \frac{1}{\alpha_{n+1,k+1}}\right)\right]$$
$$= \frac{\overline{\alpha}_{n,k}(1-s_k^n)}{(k+1)s_k^n}[F_{k+1}^{n+1} - (n-k)m_{n+1,k+1} - ns_k^n(F_n^{n+1} - m_{n+1,k+1})].$$

Observe from (8) that $m_{n+1,k+1} < F_n^{n+1}$ and using $ns_k^n > n-k$ gives

$$F_{k+1}^{n+1} - (n-k)m_{n+1,k+1} - ns_k^n(F_n^{n+1} - m_{n+1,k+1}) < 0,$$

from which it may be seen that $\overline{m}_n \leq m_{n+1} < m_{n+1,k_n}$ by setting $k = k_n < k_{n+1}$ and using (iv) above. That $m_{n,k_{n+1}} \leq \overline{m}_n$ in this case follows in a similar fashion.

(ix) The conclusion follows easily from (6) and (i)–(viii) above. □

The next lemma records some useful identities relating the parameters.

LEMMA 2.5. *For* $1 \leq k < n$, *the following equations* (9)–(17) *hold*:

$$\frac{\overline{\alpha}_{n,k} - \alpha_{n,k}}{\overline{m}_{n,k} - m_{n,k}} = \mu_{n,k} = \frac{\alpha_{n+1,k} - \overline{\alpha}_{n,k}}{m_{n+1,k} - \overline{m}_{n,k}}; \tag{9}$$

$$\frac{\alpha_{n,k}\overline{m}_{n,k} - \overline{\alpha}_{n,k}m_{n,k}}{\overline{m}_{n,k} - m_{n,k}} = \lambda_{n,k} = \frac{\overline{\alpha}_{n,k}m_{n+1,k} - \alpha_{n+1,k}\overline{m}_{n,k}}{m_{n+1,k} - \overline{m}_{n,k}}; \tag{10}$$

$$\frac{\overline{\alpha}_{n,k} - \alpha_{n,k+1}}{\overline{m}_{n,k} - m_{n,k+1}} = \mu_{n,k+1} = \frac{\alpha_{n+1,k+1} - \overline{\alpha}_{n,k}}{m_{n+1,k+1} - \overline{m}_{n,k}}; \tag{11}$$

$$\frac{\alpha_{n,k+1}\overline{m}_{n,k} - \overline{\alpha}_{n,k}m_{n,k+1}}{\overline{m}_{n,k} - m_{n,k+1}} = \lambda_{n,k+1} = \frac{\overline{\alpha}_{n,k}m_{n+1,k+1} - \alpha_{n+1,k+1}\overline{m}_{n,k}}{m_{n+1,k+1} - \overline{m}_{n,k}}. \tag{12}$$

$$\left(\frac{1 - s_k^{n-1}}{s_{k-1}^{n-1}}\right)\frac{1}{\alpha_{n+1,k}} - \left(\frac{1 - s_k^n}{s_{k-1}^n}\right)\frac{1}{\alpha_{n,k}} = \left(\frac{k+1}{nks_{k-1}^{n-1}s_{k-1}^n}\right)\frac{1}{\overline{\alpha}_{n,k}}; \tag{13}$$

$$\left(\frac{1 - s_k^{n-1}}{s_{k-1}^{n-1}}\right)\frac{m_{n+1,k}}{\alpha_{n+1,k}} - \left(\frac{1 - s_k^n}{s_{k-1}^n}\right)\frac{m_{n,k}}{\alpha_{n,k}} = \left(\frac{k+1}{nks_{k-1}^{n-1}s_{k-1}^n}\right)\frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}}; \tag{14}$$

$$\frac{m_{n+1,k} - \overline{m}_{n,k}}{m_{n+1,k} - m_{n,k}} = \frac{\alpha_{n+1,k} - \overline{\alpha}_{n,k}}{\alpha_{n+1,k} - \alpha_{n,k}} = \frac{nk\overline{\alpha}_{n,k}s_{k-1}^{n-1}(s_k^n - 1)}{\alpha_{n,k}(k+1)}; \tag{15}$$

$$\frac{\overline{m}_{n,k} - m_{n,k}}{m_{n+1,k} - m_{n,k}} = 1 - \frac{m_{n+1,k} - \overline{m}_{n,k}}{m_{n+1,k} - m_{n,k}} = \frac{nk\overline{\alpha}_{n,k}s_{k-1}^n(1 - s_k^{n-1})}{\alpha_{n+1,k}(k+1)}; \tag{16}$$

$$-\frac{\lambda_{n,k}}{\mu_{n,k}} = \frac{ns_{k-1}^{n-1}F_n^{n+1} - F_k^n}{ns_{k-1}^{n-1} - (n-k)}. \tag{17}$$

Proof. It is sufficient to prove just one side of the relations (9)–(12) in each case. For example, for (9) think of the slopes of the lines joining the points $(m_{n,k}, \alpha_{n,k})$, $(\overline{m}_{n,k}, \overline{\alpha}_{n,k})$ and $(m_{n+1,k}, \alpha_{n+1,k})$. To prove (9) and (10), first note that

$$-\frac{\lambda_{n,k}}{\mu_{n,k}} = \frac{(m_{n+1,k}/\alpha_{n+1,k}) - (m_{n,k}/\alpha_{n,k})}{(1/\alpha_{n+1,k}) - (1/\alpha_{n,k})} = \frac{(\overline{m}_{n,k}/\overline{\alpha}_{n,k}) - (m_{n,k}/\alpha_{n,k})}{(1/\overline{\alpha}_{n,k}) - (1/\alpha_{n,k})}, \tag{18}$$

which follows by observing that

$$\frac{1}{\overline{\alpha}_{n,k}} - \frac{1}{\alpha_{n,k}} = \frac{(1 - s_k^{n-1})(ns_{k-1}^{n-1} - (n-k))}{(k+1)s_{k-1}^{n-1}} = \frac{nks_{k-1}^n(1 - s_k^{n-1})}{k+1}\left[\frac{1}{\alpha_{n+1,k}} - \frac{1}{\alpha_{n,k}}\right]$$

and

$$\begin{aligned}\frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}} - \frac{m_{n,k}}{\alpha_{n,k}} &= \frac{(1 - s_k^{n-1})}{k+1}\left(nF_n^{n+1} - \frac{F_k^n}{s_{k-1}^{n-1}}\right) \\ &= \frac{nks_{k-1}^n(1 - s_k^{n-1})}{k+1}\left[\frac{m_{n+1,k}}{\alpha_{n+1,k}} - \frac{m_{n,k}}{\alpha_{n,k}}\right].\end{aligned}$$

To see (9) (and hence (10)), notice that (18) implies that

$$\begin{aligned}\frac{1}{\mu_{n,k}} &= \frac{1}{\alpha_{n,k}}\left[\frac{(m_{n+1,k}/\alpha_{n+1,k}) - (m_{n,k}/\alpha_{n,k})}{(1/\alpha_{n,k}) - (1/\alpha_{n+1,k})}\right] + \frac{m_{n,k}}{\alpha_{n,k}} \\ &= \frac{1}{\alpha_{n,k}}\left[\frac{(\overline{m}_{n,k}/\overline{\alpha}_{n,k}) - (m_{n,k}/\alpha_{n,k})}{(1/\alpha_{n,k}) - (1/\overline{\alpha}_{n,k})}\right] + \frac{m_{n,k}}{\alpha_{n,k}} \\ &= \frac{\overline{m}_{n,k} - m_{n,k}}{\overline{\alpha}_{n,k} - \alpha_{n,k}}.\end{aligned}$$

The relations (11) and (12) are derived in an identical manner after proving that

$$-\frac{\lambda_{n,k+1}}{\mu_{n,k+1}} = \frac{(m_{n+1,k+1}/\alpha_{n+1,k+1}) - (m_{n,k+1}/\alpha_{n,k+1})}{(1/\alpha_{n+1,k+1}) - (1/\alpha_{n,k+1})} = \frac{(\overline{m}_{n,k}/\overline{\alpha}_{n,k}) - (m_{n,k+1}/\alpha_{n,k+1})}{(1/\overline{\alpha}_{n,k}) - (1/\alpha_{n,k+1})},$$

which comes from the calculations

$$\frac{1}{\overline{\alpha}_{n,k}} - \frac{1}{\alpha_{n,k+1}} = \frac{(s_k^{n-1}-1)(n-k-ns_k^n)}{(k+1)s_k^{n-1}} = ns_k^n(1-s_k^{n-1})\left[\frac{1}{\alpha_{n+1,k+1}} - \frac{1}{\alpha_{n,k+1}}\right]$$

and

$$\frac{\overline{m}_{n,k}}{\overline{\alpha}_{n,k}} - \frac{m_{n,k+1}}{\alpha_{n,k+1}} = \frac{(1-s_k^{n-1})}{k+1}\left(nF_n^{n+1} - \frac{F_{k+1}^n}{s_k^{n-1}}\right) = ns_k^n(1-s_k^{n-1})\left[\frac{m_{n+1,k+1}}{\alpha_{n+1,k+1}} - \frac{m_{n,k+1}}{\alpha_{n,k+1}}\right].$$

The identities (13)–(16) may be obtained from direct calculation from the definitions (4) and (5). □

3. The optimal *Q*-strategy. For each pair n, k, $1 \le k < n$ and $1 \le j \le k+1$, define

$$q_j^{n,k} = \frac{\lambda_{n,k} + \mu_{n,k}F_{j-1}^j}{1 - s_{j-1}\lambda_{n,k} - \mu_{n,k}\left(\sum_{i=1}^{j-2}(f(i)/(i+1)) + f(j-1)\right)}, \tag{19}$$

where an empty sum is zero, and define the strategy $\mathbf{q}^{n,k} = (q_1^{n,k}, q_2^{n,k}, \ldots, q_k^{n,k}, 1, 1, \cdots)$.

Using these strategies, the minimax-optimal *Q-strategies* $\mathbf{q}_M$ appearing in (2) can now be given.

DEFINITIONS. *For all* $M > f(1)$,

$$\begin{aligned} n(M) &= n \quad \textit{when } m_n \le M < m_{n+1}; \\ k(M) &= \begin{cases} k_n & \textit{if } m_n \le M < m_{n+1} \textit{ and } k_{n+1} = k_n, \textit{ or} \\ & \textit{if } m_n \le M < \overline{m}_n \textit{ and } k_{n+1} = k_n + 1, \\ k_{n+1} & \textit{if } \overline{m}_n \le M < m_{n+1} \textit{ and } k_{n+1} = k_n + 1; \end{cases} \quad \textit{and} \\ \mathbf{q}_M &= \mathbf{q}^{n(M),k(M)}. \end{aligned} \tag{20}$$

Example 3.1. For the expected-value case $f(i) \equiv i$, it follows from the calculations in Example 2.3 that

$$\mathbf{q}^{3,1} = \left(\frac{7}{13}, 1, 1, \cdots\right),$$

$$\mathbf{q}^{4,1} = \left(\frac{105}{213}, 1, 1, \cdots\right), \quad \text{and}$$

$$\mathbf{q}^{4,2} = \left(\frac{50}{107}, \frac{14}{19}, 1, 1, \cdots\right),$$

and together these yield the minimax-optimal $\mathbf{q}_M$ in Example 1.2. □

First it must be shown that $\mathbf{q}^{n,k} \in \mathbf{Q}$ for all $1 \le k < n$; that is, each coordinate must be shown to be a probability.

LEMMA 3.2. *For* $1 \le j \le k+1 \le n$, $0 \le q_j^{n,k} \le 1$.

Proof. First, the numerator of $q_j^{n,k}$ is > 0. This is because $\lambda_{n,k} + \mu_{n,k} F_{j-1}^j =$

$$\begin{aligned}
&\frac{\alpha_{n,k}\alpha_{n+1,k}}{m_{n+1,k} - m_{n,k}} \left[\frac{m_{n+1,k}}{\alpha_{n+1,k}} - \frac{m_{n,k}}{\alpha_{n,k}} + \left(\frac{1}{\alpha_{n,k}} - \frac{1}{\alpha_{n+1,k}} \right) F_{j-1}^j \right] \\
&= \frac{\alpha_{n,k}\alpha_{n+1,k}}{m_{n+1,k} - m_{n,k}} \left[\frac{n-k+1}{k s_{k-1}^n} \left(\frac{F_k^{n+1}}{n-k+1} - F_{j-1}^j \right) - \frac{n-k}{k s_{k-1}^{n-1}} \left(\frac{F_k^n}{n-k} - F_{j-1}^j \right) \right] > 0,
\end{aligned} \tag{21}$$

by Lemma 2.2 (iv) and (v) and the fact that $F_k^n/(n-k)$ exceeds F_{j-1}^j for $j \le k+1$. Denote the denominator of $q_j^{n,k}$ in (19) by β_j. Then to show $\beta_j > 0$, note that

$$\beta_j - \beta_{j+1} = (\lambda_{n,k} + \mu_{n,k} F_{j-1}^j)/j > 0,$$

by (21) for $j \le k+1$, so it is sufficient to show that $\beta_{k+1} > 0$. But now, from (4) and (5) calculate that

$$\begin{aligned}
\beta_{k+1} &= \frac{(1 - s_k \alpha_{n,k})\alpha_{n+1,k}}{m_{n+1,k} - m_{n,k}} \left[\frac{m_{n+1,k}}{\alpha_{n+1,k}} - \frac{m_{n,k}}{\alpha_{n,k}} + \left(\frac{1}{\alpha_{n,k}} - \frac{1}{\alpha_{n+1,k}} \right) \frac{F_k^n}{n-k} \right] \\
&= \frac{(1 - s_k \alpha_{n,k})(1 - s_k \alpha_{n+1,k})}{m_{n+1,k} - m_{n,k}} \left[\frac{F_k^{n+1}}{n-k+1} - \frac{F_k^n}{n-k} \right] > 0.
\end{aligned} \tag{22}$$

To show that $q_j^{n,k} \le 1$ for $j \le k+1$, note that the statement that $q_j^{n,k} \le 1$ is equivalent to

$$\gamma_j = \lambda_{n,k}(1 + s_{j-1}) + \mu_{n,k} \left(F_{j-1}^j + f(j-1) + \sum_{i=1}^{j-2} \frac{f(i)}{i+1} \right) \le 1.$$

Thus it is sufficient to show that $\gamma_{j+1} - \gamma_j \ge 0$ for $j \le k$ and $q_{k+1}^{n,k} \le 1$. First, using (4) and the expressions for the numerator and denominator from above, showing that

$$q_{k+1}^{n,k} = \frac{k s_{k-1}^{n-1}(F_k^{n+1} - (n-k+1)F_k^{k+1}) - k s_{k-1}^n (F_k^n - (n-k)F_k^{k+1})}{(n-k)F_k^{n+1} - (n-k+1)F_k^n} \le 1$$

is equivalent after rearrangement to showing that

$$\frac{1 - [(k s_{k-1}^n)/(n-k+1)]}{1 - [(k s_{k-1}^{n-1})/(n-k)]} \le \frac{(F_k^{n+1}/(n-k+1)) - F_k^{k+1}}{(F_k^n/(n-k)) - F_k^{k+1}}. \tag{23}$$

But by the convexity of $1/i$ in i, the left-hand side of (23) is dominated by $(n-k)/(n-k-1)$, which in turn is dominated by the right-hand side of (23) using Lemma 2.2 (iii). Furthermore,

$$\gamma_{j+1} - \gamma_j = [\lambda_{n,k} + \mu_{n,k}(j F_j^{j+1} - (j-1)F_{j-1}^j)]/j,$$

and since F_{j-1}^j is increasing in j and $\mu_{n,k} < 0$, to prove that $\gamma_{j+1} \ge \gamma_j$ for $j \le k$, it is sufficient to show that

$$-\frac{\lambda_{n,k}}{\mu_{n,k}} \ge k F_k^{k+1} - (k-1)F_{k-1}^k. \tag{24}$$

From (17),

$$-\frac{\lambda_{n,k}}{\mu_{n,k}} = \frac{ns_{k-1}^{n-1}F_n^{n+1} - F_k^n}{ns_{k-1}^{n-1} - (n-k)} \geq \frac{(n+k-1)F_n^{n+1} - 2(k-1)(F_k^n/(n-k))}{n-k+1}, \tag{25}$$

the last inequality because, by the convexity of $1/i$,

$$s_{k-1}^{n-1} \leq \frac{1}{2}(n-k)\left(\frac{1}{n} + \frac{1}{k-1}\right) = \frac{(n-k)(n+k-1)}{2n(k-1)}.$$

Finally, it may be seen that the right-hand side of (25) exceeds that of (24) using Lemma 2.2 (iii) again. □

It is now possible to prove an inequality that will imply that if the Q-player uses strategy $\mathbf{q}_M$, then it forces the P-player to put positive probability mass only on the points $1, 2, \ldots, k(M), n(M), n(M)+1$. Recall the definitions of $\mathbf{q}_M$ and $\mathbf{v}_M$.

PROPOSITION 3.3. *For all* $M > f(1)$,

$$V(\mathbf{p}, \mathbf{q}_M) \geq v_M \quad \textit{for all } \mathbf{p} = (p_1, p_2, \cdots) \in \mathcal{P}_M, \tag{26}$$

with equality in (26) *if and only if* $\sum_{j=1}^{\infty} p_j f(j) = M$ *and* $\mathbf{p}$ *assigns positive mass only to points in* $\{1, \ldots, k(M), n(M), n(M)+1\}$.

Proof. By the definitions of $\mathbf{q}_M$ and v_M, it is enough to show

$$V(\mathbf{p}, \mathbf{q}^{n,k}) \geq \lambda_{n,k} + \mu_{n,k}M, \quad \text{for all } \mathbf{p} = (p_1, p_2, \cdots) \in \mathcal{P}_M, \tag{27}$$

with equality if and only if $\sum_{j=1}^{\infty} p_j f(j) = M$ and $\mathbf{p}$ assigns positive mass only to points $\{1, \ldots, k, n, n+1\}$. First observe that

$$\prod_{m=1}^{j}\left(1 - \frac{q_m^{n,k}}{m}\right) = 1 - s_j\lambda_{n,k} - \mu_{n,k}\left(\sum_{m=1}^{j-1}\frac{f(m)}{m+1} + f(j)\right) \tag{28}$$

and

$$\sum_{i=1}^{j} q_i^{n,k} \prod_{m=1}^{i-1}\left(1 - \frac{q_m^{n,k}}{m}\right) = j\lambda_{n,k} + jf(j)\mu_{n,k}. \tag{29}$$

For strategies $\mathbf{p} = (p_1, p_2, \cdots)$ and $\mathbf{q} = (q_1, q_2, \cdots)$, recall (1) and define the Lagrangian

$$\begin{aligned} L(\mathbf{p}, \mathbf{q}) &= V(\mathbf{p}, \mathbf{q}) + \lambda_{n,k}\left(1 - \sum_{j=1}^{\infty} p_j\right) + \mu_{n,k}\left(M - \sum_{j=1}^{\infty} p_j f(j)\right) \\ &= \lambda_{n,k} + \mu_{n,k}M + \sum_{j=1}^{\infty}\frac{p_j}{j}\left[\sum_{i=1}^{j} q_i \prod_{m=1}^{i-1}\left(1 - \frac{q_m}{m}\right) - \lambda_{n,k}j - \mu_{n,k}jf(j)\right]. \end{aligned} \tag{30}$$

The dependence of L on n and k will be suppressed in the notation. For $\mathbf{p} \in \mathcal{P}_M$, since $\mu_{n,k} < 0$, it is immediate that $V(\mathbf{p}, \mathbf{q}) \geq L(\mathbf{p}, \mathbf{q})$, with equality if and only if $\sum_{j=1}^{\infty} p_j f(j) = M$. When $\mathbf{q} = \mathbf{q}^{n,k}$ it is now sufficient to show that the coefficient of p_j/j in (30) is ≥ 0 for all j and is $= 0$ if and only if $j = 1, 2, \ldots, k, n, n+1$. But for $1 \leq j \leq k$ this is true from (29).

For $j > k$ the coefficient of p_j/j is

$$\sum_{i=1}^{k} q_i^{n,k} \prod_{m=1}^{i-1} \left(1 - \frac{q_m^{n,k}}{m}\right) + \prod_{m=1}^{k} \left(1 - \frac{q_m^{n,k}}{m}\right) \sum_{i=k+1}^{j} \prod_{r=k+1}^{i-1} \left(\frac{r-1}{r}\right)$$

$$- \lambda_{n,k} j - \mu_{n,k} j f(j) = \lambda_{n,k}(k-j) - \mu_{n,k} F_k^j + k s_{k-1}^{j-1} \tag{31}$$

$$\cdot \left[1 - s_k \lambda_{n,k} - \mu_{n,k} \left(\sum_{i=1}^{k-1} \frac{f(i)}{i+1} + f(k)\right)\right],$$

from (29), with the convention that an empty product is 1. Using (4)–(6) and (22) and putting $G_s^r = F_s^r/(r-s)$ for $r > s$, (31) reduces to

$$\frac{(1 - s_k \alpha_{n,k})(1 - s_k \alpha_{n+1,k})(j-k) k \delta_j}{m_{n+1,k} - m_{n,k}},$$

where

$$\delta_j = \frac{s_{k-1}^{j-1}}{j-k}(G_k^{n+1} - G_k^n) - \frac{s_{k-1}^{n-1}}{n-k}(G_k^{n+1} - G_k^j) + \frac{s_{k-1}^n}{n-k+1}(G_k^n - G_k^j).$$

Note that $\delta_n = \delta_{n+1} = 0$, and

$$\delta_{j+1} - \delta_j = \left(\frac{s_{k-1}^j}{j-k+1} - \frac{s_{k-1}^{j-1}}{j-k}\right)(G_k^{n+1} - G_k^n) - \left(\frac{s_{k-1}^n}{n-k+1} - \frac{s_{k-1}^{n-1}}{n-k}\right)(G_k^{j+1} - G_k^j).$$

But, using Lemma 2.1 (ii), $G_k^{j+1} - G_k^j$ is positive and nondecreasing and the expression $(s_{k-1}^j/(j-k+1)) - (s_{k-1}^{j-1}/(j-k))$ is negative and increasing in $j > k$; hence $\delta_{j+1} < \delta_j$ for $j < n$ and $\delta_{j+1} > \delta_j$ for $j > n$, which shows that (31) is > 0 when $j > k$, $j \neq n$, $n+1$, which completes the proof. □

Once the analogue of Proposition 3.3 is proved for the P-strategy (Proposition 4.2 below), this will establish the minimax-optimality of both $\mathbf{q}_M$ and $\mathbf{p}_M$ and that v_M is the value of the game.

4. The optimal P-strategy and main theorem. For $1 \leq k \leq k_n < n$, define distributions $\mathbf{p}^{n,k} = (p_1^{n,k}, p_2^{n,k}, \cdots)$ concentrated on the points $\{1, 2, \ldots, k, n\}$ by

$$p_j^{n,k} = \begin{cases} \alpha_{n,k}/(j+1) & \text{for } 1 \leq j < k, \\ \alpha_{n,k}(s_{k-1}^{n-1} - 1)/s_{k-1}^{n-1} & \text{for } j = k, \\ n\alpha_{n,k}/(k s_{k-1}^{n-1}) & \text{for } j = n, \end{cases}$$

with $p_j^{n,k} \equiv 0$ otherwise. Note that $\mathbf{p}^{n,k}$ is a distribution only for $k \leq k_n$ and that $m_{n,k} = \sum_{j=1}^{\infty} f(j) p_j^{n,k}$. Set $\mathbf{p}^n = \mathbf{p}^{n,k_n}$ for $n \geq 2$, and $\mathbf{p}^1 = (1, 0, 0, \cdots)$. Also, when $k_{n+1} = k_n + 1$, define the distribution $\overline{\mathbf{p}}^n = (\overline{p}_1^n, \overline{p}_2^n, \cdots)$ concentrated on $\{1, 2, \ldots, k_n, n, n+1\}$ by

$$\overline{p}_j^n = \begin{cases} \overline{\alpha}_n/(j+1) & \text{for } 1 \leq j \leq k_n, \\ \overline{\alpha}_n n^2 (s_{k_n}^n - 1)/(k_n + 1) & \text{for } j = n, \\ \overline{\alpha}_n n(n+1)(1 - s_{k_n}^{n-1})/(k_n + 1) & \text{for } j = n+1, \end{cases}$$

with $\overline{p}_j^n \equiv 0$ otherwise. Note that $\overline{m}_n = \sum_{j=1}^{\infty} f(j) \overline{p}_j^n$.

Remark. The distributions $\mathbf{p}^n$, $n = 1, 2, \cdots$ are exactly the minimax-optimal P-player strategies for the $N \leq n$ problem studied in Hill and Krengel (1991).

The minimax-optimal (worst-case distribution) strategy $\mathbf{p}_M$ in (2) for player P can now be shown to be convex combinations of these base strategies $\mathbf{p}^n$ and $\overline{\mathbf{p}}^n$.

DEFINITION. *For all* $M > f(1)$, *define* $\mathbf{p}_M \in \mathcal{P}_M$ *by*

$$(32)\quad \mathbf{p}_M = \begin{cases} \left(\dfrac{m_{n+1}-M}{m_{n+1}-m_n}\right)\mathbf{p}^n + \left(\dfrac{M-m_n}{m_{n+1}-m_n}\right)\mathbf{p}^{n+1} \in \mathcal{P}_M \\ \qquad\qquad \text{if } k_{n+1} = k_n \text{ and } m_n \le M < m_{n+1}; \\ \left(\dfrac{\overline{m}_n-M}{\overline{m}_n-m_n}\right)\mathbf{p}^n + \left(\dfrac{M-m_n}{\overline{m}_n-m_n}\right)\overline{\mathbf{p}}^n \in \mathcal{P}_M \\ \qquad\qquad \text{if } k_{n+1} = k_n + 1 \text{ and } m_n \le M < \overline{m}_n; \\ \left(\dfrac{m_{n+1}-M}{m_{n+1}-\overline{m}_n}\right)\overline{\mathbf{p}}^n + \left(\dfrac{M-\overline{m}_n}{m_{n+1}-\overline{m}_n}\right)\mathbf{p}^{n+1} \in \mathcal{P}_M \\ \qquad\qquad \text{if } k_{n+1} = k_n + 1 \text{ and } \overline{m}_n \le M < m_{n+1}. \end{cases}$$

Notice that the strategy $\mathbf{p}_M$ *places positive probability mass only on the points in the set* $\{1, \ldots, k(M), n(M), n(M)+1\}$ *and that, when* $k_{n+1} = k_n$, $\overline{m}_n$ *is a changeover point in that when* M *increases through the value* $M = \overline{m}_n$, $\mathbf{p}_M$ *increases the number of points in its support by* 1.

Example 4.1. For the expected-value case $f(i) \equiv i$, it follows from the calculations in Example 2.3 that

$$\mathbf{p}^3 = \mathbf{p}^{3,1} = \left(\frac{1}{7}, 0, \frac{6}{7}, 0, 0, \cdots\right),$$

$$\mathbf{p}^4 = \mathbf{p}^{4,1} = \left(\frac{5}{29}, 0, 0, \frac{24}{29}, 0, 0, \cdots\right),$$

$$\overline{\mathbf{p}}^4 = \left(\frac{3}{17}, 0, 0, \frac{4}{17}, \frac{10}{17}, 0, 0, \cdots\right), \quad \text{and}$$

$$\mathbf{p}^5 = \mathbf{p}^{5,2} = \left(\frac{13}{75}, \frac{2}{75}, 0, 0, \frac{60}{75}, 0, 0, \cdots\right),$$

and together these yield the minimax-optimal strategies $\mathbf{p}_M$ in Example 1.2. □

Now the analogue of Proposition 3.3 will be proved, which, together with Proposition 3.3, will establish the minimax-optimality of $\mathbf{q}_M$ and $\mathbf{p}_M$ simultaneously.

PROPOSITION 4.2. *For all* $M > f(1)$,

$$(33)\qquad V(\mathbf{p}_M, \mathbf{q}) \le v_M \quad \text{for all } \mathbf{q} = (q_1, q_2, \cdots) \in Q.$$

Proof. The argument of Theorem C of Hill and Krengel (1991) demonstrates that

$$V(\mathbf{p}^{n,k}, (q_1, q_2, \ldots, q_k, 1, 1, \cdots)) \equiv \alpha_{n,k}, \quad \text{for } 0 \le q_i \le 1,\ i = 1, \ldots, k,$$

and a similar argument shows that, when $k_{n+1} = k_n + 1$,

$$V(\overline{\mathbf{p}}^n, (q_1, q_2, \ldots, q_{k_{n+1}}, 1, 1, \cdots)) \equiv \overline{\alpha}_n, \quad \text{for } 0 \le q_i \le 1,\ i = 1, \ldots, k_{n+1}.$$

Furthermore, Hill and Krengel (1991) established that for any $\mathbf{q} = (q_1, q_2, \cdots) \in \mathcal{Q}$ and $k \ge k_n$,

$$(34)\quad V(\mathbf{p}^n, \mathbf{q}) \le V(\mathbf{p}^n, (q_1, \ldots, q_k, 1, 1, \cdots)) \le V(\mathbf{p}^n, (q_1, \ldots, q_{k_n}, 1, 1, \cdots)) = \alpha_n,$$

and again it is straightforward to prove that when $k_{n+1} = k_n + 1$ and $k \geq k_{n+1}$,

$$(35)\quad V(\overline{\mathbf{p}}^n, \mathbf{q}) \leq V(\overline{\mathbf{p}}^n, (q_1, \ldots, q_k, 1, 1, \cdots)) \leq V(\overline{\mathbf{p}}^n, (q_1, \ldots, q_{k_{n+1}}, 1, 1, \cdots)) = \overline{\alpha}_n.$$

Three cases corresponding to the three possibilities in (32) must be considered.

Case 1. $k_{n+1} = k_n$ and $m_n \leq M < m_{n+1}$.

$$V(\mathbf{p}_M, \mathbf{q}) = \left(\frac{m_{n+1} - M}{m_{n+1} - m_n}\right) V(\mathbf{p}^n, \mathbf{q}) + \left(\frac{M - m_n}{m_{n+1} - m_n}\right) V(\mathbf{p}^{n+1}, \mathbf{q})$$
$$\leq \left(\frac{m_{n+1} - M}{m_{n+1} - m_n}\right) \alpha_n + \left(\frac{M - m_n}{m_{n+1} - m_n}\right) \alpha_{n+1} = \lambda_{n,k_n} + \mu_{n,k_n} M,$$

for all $\mathbf{q} \in \mathcal{Q}$, using (34).

Case 2. $k_{n+1} = k_n + 1$ and $m_n \leq M < \overline{m}_n$. Using the relations in (9)–(16) and setting $M = \overline{m}_n$ it may be seen that $\mathbf{p}_{\overline{m}_n} = \overline{\mathbf{p}}_n$, whence

$$\mathbf{p}_M \equiv \left(\frac{m_{n+1,k_n} - M}{m_{n+1,k_n} - m_n}\right) \mathbf{p}^n + \left(\frac{M - m_n}{m_{n+1,k_n} - m_n}\right) \mathbf{p}^{n+1,k_n}.$$

Note from Lemma 2.4 (viii) that for M in this range, $M \leq m_{n+1,k_n}$. For any $\mathbf{q} \in \mathcal{Q}$,

$$V(\mathbf{p}_M, \mathbf{q}) = \left(\frac{m_{n+1,k_n} - M}{m_{n+1,k_n} - m_n}\right) V(\mathbf{p}^n, \mathbf{q}) + \left(\frac{M - m_n}{m_{n+1,k_n} - m_n}\right) V(\mathbf{p}^{n+1,k_n}, \mathbf{q}),$$

which, by (34), is dominated by

$$(36)\quad \left(\frac{m_{n+1,k_n} - M}{m_{n+1,k_n} - m_n}\right) V(\mathbf{p}^n, (q_1, \ldots, q_{k_n+1}, 1, 1, \cdots)) + \left(\frac{M - m_n}{m_{n+1,k_n} - m_n}\right) V(\mathbf{p}^{n+1,k_n}, (q_1, \ldots, q_{k_n+1}, 1, 1, \cdots)).$$

Now consider the coefficient of $q_{k_n+1} \prod_{m=1}^{k_n} (1 - q_m/m)$ in (36) which equals

$$\frac{p_n^{n,k_n}}{n} \left(\frac{m_{n+1,k_n} - M}{m_{n+1,k_n} - m_n}\right) (1 - s_{k_n}^{n-1}) + \frac{p_{n+1}^{n+1,k_n}}{n+1} \left(\frac{M - m_n}{m_{n+1,k_n} - m_n}\right) (1 - s_{k_n}^n)$$
$$= \frac{\alpha_n}{k_n} \left(\frac{m_{n+1,k_n} - M}{m_{n+1,k_n} - m_n}\right) \left(\frac{1 - s_{k_n}^{n-1}}{s_{k_n-1}^{n-1}}\right) + \frac{\alpha_{n+1,k_n}}{k_n} \left(\frac{M - m_n}{m_{n+1,k_n} - m_n}\right) \left(\frac{1 - s_{k_n}^n}{s_{k_n-1}^n}\right).$$

Rearranging this expression as

$$\frac{\alpha_n \alpha_{n+1,k_n}}{k_n (m_{n+1,k_n} - m_n)} \left[\left(\frac{1 - s_{k_n}^{n-1}}{s_{k_n-1}^{n-1}}\right) \left(\frac{m_{n+1,k_n} - M}{\alpha_{n+1,k_n}}\right) + \left(\frac{1 - s_{k_n}^n}{s_{k_n-1}^n}\right) \left(\frac{M - m_n}{\alpha_n}\right)\right],$$

and using (13) and (14) (with $k = k_n$) this equals

$$\frac{\alpha_n \alpha_{n+1,k_n} (k_n + 1)}{n k_n s_{k_n-1}^{n-1} s_{k_n-1}^n (m_{n+1,k_n} - m_n)} \left(\frac{\overline{m}_n - M}{\overline{\alpha}_n}\right) \geq 0,$$

for $M \leq \overline{m}_n$. It follows that (36) is not decreased by taking $q_{k_n+1} = 1$, so it may be seen that

$$V(\mathbf{p}_M, \mathbf{q}) \leq \left(\frac{m_{n+1,k_n} - M}{m_{n+1,k_n} - m_n}\right) V(\mathbf{p}^n, (q_1, \ldots, q_{k_n}, 1, 1, \cdots)) + \left(\frac{M - m_n}{m_{n+1,k_n} - m_n}\right) V(\mathbf{p}^{n+1,k_n}, (q_1, \ldots, q_{k_n}, 1, 1, \cdots)),$$

which in turn implies that

$$V(\mathbf{p}_M, \mathbf{q}) \le \left(\frac{m_{n+1,k_n} - M}{m_{n+1,k_n} - m_n}\right) \alpha_n + \left(\frac{M - m_n}{m_{n+1,k_n} - m_n}\right) \alpha_{n+1,k_n} = \lambda_{n,k_n} + \mu_{n,k_n} M.$$

Case 3. $k_{n+1} = k_n + 1$ and $\overline{m}_n \le M < m_{n+1}$. It follows from (34) and (35) as in Case 1 that

$$\begin{aligned} V(\mathbf{p}_M, \mathbf{q}) &= \left(\frac{m_{n+1} - M}{m_{n+1} - \overline{m}_n}\right) V(\overline{\mathbf{p}}^n, \mathbf{q}) + \left(\frac{M - \overline{m}_n}{m_{n+1} - \overline{m}_n}\right) V(\mathbf{p}^{n+1}, \mathbf{q}) \\ &\le \left(\frac{m_{n+1} - M}{m_{n+1} - \overline{m}_n}\right) \overline{\alpha}_n + \left(\frac{M - \overline{m}_n}{m_{n+1} - \overline{m}_n}\right) \alpha_{n+1} \\ &= \lambda_{n,k_{n+1}} + \mu_{n,k_{n+1}} M, \end{aligned}$$

using (11) and (12). □

The main results in this paper can now be summarized in the following theorem.

THEOREM 4.3. *With* $\mathbf{p}_M, \mathbf{q}_M$, *and* v_M *as in* (32), (20), *and* (7), *respectively*,

$$V(\mathbf{p}_M, \mathbf{q}) \le V(\mathbf{p}_M, \mathbf{q}_M) = v_M \le V(\mathbf{p}, \mathbf{q}_M), \quad \text{for all } \mathbf{p} \in \mathcal{P}_M, \mathbf{q} \in \mathcal{Q}.$$

Proof. The proof is immediate from Propositions 3.3 and 4.2 and the definitions. □

Acknowledgment. The main question addressed in this paper was raised by Shmuel Gal in the AMS-IMS-SIAM Conference on Sequential Search Strategies in Amherst (1990), and the authors are grateful to the organizers Tom Ferguson and Steve Samuels for invitations to speak at that conference.

REFERENCES

[1] T. S. FERGUSON, *Who solved the secretary problem*? Statist. Sci., 4 (1989), pp. 282–296.

[2] P. R. FREEMAN, *The secretary problem and its extensions: A review*, Internat. Statist. Rev., 51 (1983), pp. 189–206.

[3] T. P. HILL AND U. KRENGEL, *Minimax-optimal stop rules and distributions in secretary problems*, Ann. Probab., 19 (1991), pp. 342–353.

[4] E. PRESMAN AND I. SONIN, *The best choice problem for a random number of objects*, Theory Probab. Appl., 17 (1972), pp. 657–668.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 952–964, July 1994

003

THE H_∞-PROBLEM WITH CONTROL CONSTRAINTS*

VIOREL BARBU†

Abstract. Necessary and sufficient conditions for the existence of a solution to the suboptimal H_∞-problem for input-output linear systems with control constraints are established.

Key words. stabilizing feedback, Hamilton–Jacobi equations, analytic semigroup, convex function, subdifferential

AMS subject classifications. 93B50, 93C35, 49A40

1. Problem formulation.

Consider the input-output system

$$\begin{aligned} &x'(t) = Ax(t) + B_2u(t) + B_1w(t), && t \in R = [0,\infty),\\ &x(0) = x_0,\\ &z(t) = C_1x(t) + D_{12}u(t) && \text{a.e. } t \in R^+,\\ &u(t) \in U_0 && \text{a.e. } t > 0, \end{aligned} \tag{1.1}$$

where A is the infinitesimal generator of a C_0-semigroup e^{At} on $X, B_2 \in L(U,X), B_1 \in L(W,X), C_1 \in L(X,Z), D_{12} \in L(U,Z)$ and $z' = dx/dt$. Here X, Z, U, W are separable real Hilbert spaces and U_0 is a closed convex subset of U such that $0 \in U_0$.

The system (1.1) will be studied under the following standard hypotheses:

$$D_{12}^*D_{12} = I, \qquad D_{12}^*C_1 = 0, \tag{1.2}$$

$$\text{The pair } (A, C_1) \text{ is exponentially detectable.} \tag{1.3}$$

Here and throughout in the sequel we shall use the asterisk symbol to denote the dual operators. Also we shall denote by $|\cdot|, |\cdot|_Z, |\cdot|_U, |\cdot|_W$ the norm in X, Z, U, and W, respectively, and by $(\cdot,\cdot), (\cdot,\cdot)_Z, (\cdot,\cdot)_U, (\cdot,\cdot)_W$ the corresponding scalar products.

In system (1.1) $x \in X, u \in U$, and $z \in Z$ are the state, the control variable, and the disturbance (exogeneous variable), respectively.

By definition an *admissible feedback control* is a multivalued mapping $F : X \to U_0$ having the property that for every $x_0 \in X$ and $f \in L^2_{\text{loc}}(R^+;X)$ the Cauchy problem

$$x' \in Ax + B_2Fx + f \quad \text{in } R^+; \qquad x(0) = x_0 \tag{1.4}$$

has at least one mild solution x, i.e., there is $u(t) \in Fx(t)$ almost everywhere $t > 0$, $u \in L^2_{\text{loc}}(R^+;U)$ such that

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}(B_2u(s) + f(s))\,ds, \qquad t > 0. \tag{1.4$'$}$$

An admissible feedback control $F$is said to be stabilizing if for all $x_0 \in X$ and $f \in L^2(R^+;X)$ the Cauchy problem (1.4) has at least one mild solution $x \in C(R^+;X) \cap L^2(R^+;X)$, i.e., there exists $u \in L^2(R^+;X)$ such that $u(t) \in Fx(t)$ almost everywhere $t > 0$ and (1.4)$'$ holds.

* Received by the editors June 15, 1992; accepted for publication (in revised form) January 20, 1993. This paper represents results obtained at Institut National de Recherche en Informatique et en Automatique, Rocquencourt and was supported in part by National Science Foundation grant NSF-DMS-91-11794.

† Department of Mathematics, University of Iasi, 6600 Iasi, Romania.

We shall denote by $\mathcal{F}$ the set of all stabilizing feedback controls F. For every $f \in \mathcal{F}$ and $w \in L^2(R^+; X), x_0 \in X$ we set

$$S_F(x_0, w) = z = C_1 x + D_{12} u, \tag{1.5}$$

where $(x, u) \in L^2(R^+; X) \times L^2(R^+; U), u(t) \in Fx(t)$ almost everywhere $t > 0$ satisfy system (1.1), i.e.,

$$x(t) = e^{At} x_0 + \int_0^t e^{A(t-s)} (B_2 u(s) + B_1 w(s))\, ds, \qquad t \geq 0. \tag{1.6}$$

The operator $S_F : X \times L^2(R^+; W) \to L^2(R^+; Z)$ is in general multivalued but everywhere defined on $X \times L^2(R^+; W)$.

According to the theory of standard H_∞-problem [2], we shall define the H_∞-suboptimal control problem for system (1.1) as follows: given $\delta > 0$ and $\alpha > 0$ finds $F \in \mathcal{F}$ such that

$$|S_F(x_0, w)|^2 \leq \rho^2 \|w\|^2_{L^2(R^+;W)} + \alpha |x_0|^2, \qquad \forall (x_0, w) \in X \times L^2(R^+; W), \tag{1.7}$$

where $0 < \rho < \delta$.

Here

$$|S_F(x_0, w)| = \sup\{\|\theta\|_{L^2(R^+;Z)}; \theta \in S_F(x_0, w)\}.$$

The main result of this work, Theorem 1 below, solves the above problem in terms of a stationary Hamilton–Jacobi equation and is a generalization of known results [2]–[4], [5], [6] to the unconstrained case. This result seems to be new even in the finite-dimensional framework. Although the approach borrows an idea already used in the study of standard H_∞-problem, namely to reduce the problem to a differential game associated with system (1.1), the proof is quite different and there are significant differences between our treatment and the standard one.

2. The main result. Throughout in the sequel we shall assume that system (1.1) satisfies hypotheses (1.2) and (1.3) and also that (i) e^{At} are compacts for all $t > 0$; and (ii) e^{At} is a C_0-analytic semigroup.

THEOREM 1. *If the H_∞-suboptimal control problem has a solution $F \in \mathcal{F}$, then there is a continuous, convex function $\varphi : X \to R$ such that $\partial\varphi : X \to X$ is compact,*

$$0 \leq 2\varphi(x) \leq \alpha |x|^2, \qquad \forall x \in X, \tag{2.1}$$

$$\begin{aligned} 2(Ax, \eta) &+ |P_{U_0}(-B_2^* \eta)|^2_U + \delta^{-2} |B_1^* \eta|^2_W + 2(B_2^* \eta, P_{U_0}(-B_2^* \eta))_U + |C_1 x|^2 \\ &= 0, \quad \forall x \in D(A), \quad \forall \eta \in \partial\varphi(x). \end{aligned} \tag{2.2}$$

Moreover, the Cauchy problem

$$\begin{aligned} x' &= Ax + B_2 P_{U_0}(-B_2^* \partial\varphi(x)) + \delta^{-2} B_1 B_1^* \partial\varphi(x), \\ x(0) &= x_0 \end{aligned} \tag{2.3}$$

has for every $x_0 \in X$ at least one mild solution

$$x^* \in C(R^+; X) \cap L^2(R^+; X); \qquad \lim_{t \to \infty} x^*(t) = 0. \tag{2.4}$$

Conversely, if (2.2) *has a solution φ with the above properties then the feedback $F = P_{U_0}(-B_2^* \partial\varphi)$ is stabilizing and guarantees inequality* (1.7) *with $\rho = \delta$.*

Here $P_{U_0} : U \to U_0$ is the projection operator on U_0 and $\partial\varphi$ is the subdifferential of φ, i.e.,

$$\partial\varphi(x_0) = \{\eta \in X; \varphi(x_0) \leq \varphi(x) + (\eta, x_0 - x), \forall x \in X\}.$$

The multivalued mapping $\partial\varphi$ is said to be compact if it maps bounded subsets into compact sets.

We note that in the case of unconstrained H_∞-control problem, i.e., $U_0 = U$, (2.2) reduces to the Riccati equation corresponding to the regular H_∞-problem [3], [4], [8], while the closed-loop inequality (1.7) becomes

$$\|S_F(0, w)\|_{L^2(R^+;Z)} \leq \rho\|w\|_{L^2(R^+;W)}, \qquad \forall w \in L^2(R^+; W).$$

However, in our case a gap arises between the necessary and sufficient conditions for existence of solution to H_∞-problem. Perhaps in most significant cases, the existence of a solution φ to (2.2) is necessary and sufficient for existence of a solution to H_∞-suboptimal control problem.

Now, we shall illustrate Theorem 1 with a few examples.

Example 1. Consider the system

$$x' = -x + u + w \quad \text{in } R^+; \qquad z = \{x, u\}, \quad u(t) \geq 0, \quad t \geq 0. \tag{2.5}$$

Here $X = U = W = R$, $Z = R \times R$ and $U_0 = \{u \in R; u \geq 0\}$. Equation (2.2) has therefore in this case the following form:

$$2x\varphi'(x) - |(\varphi'(x))^-|^2 - (\delta^2)^{-1}|\varphi'(x)|^2 - x^2 = 0. \tag{2.6}$$

By a little calculation we see that for $\delta > 1$ this equation has a unique convex solution φ satisfying (2.1) and which is given by

$$\varphi(x) = \begin{cases} 2^{-1}(\delta^2 + \delta(\delta^2 - 1)^{1/2})x^2 & \text{for } x \geq 0, \\ 2^{-1}(\delta^2 - 1)^{-1/2}((2\delta^2 - 1)^{1/2} - \delta)x^2 & \text{for } x < 0, \end{cases}$$

and the suboptimal H_∞ feedback control F is given by

$$u = Fx = \begin{cases} 0 & \text{for } x \geq 0, \\ -(\delta^2 - 1)^{-1/2}((2\delta^2 - 1)^{1/2} - \delta)x & \text{for } x < 0. \end{cases} \tag{2.7}$$

Example 2. Consider the input-output system

$$y_t - \Delta y + ay = u + w \quad \text{in } \Omega \times R^+,$$

$$y = 0 \quad \text{in } \partial\Omega \times R^+,$$

$$z = (y, u) \in L^2(\Omega) \times L^2(\Omega); \qquad u \geq 0 \text{ a.e. in } \Omega \times R^+,$$

where $a \in R, w \in W = L^2(\Omega)$ and Ω is a bounded open subset of R^N.

In this case $A = -\Delta, D(A) = H_0^1(\Omega) \cap H^2(\Omega), U = L^2(\Omega), U_0 = \{u \in L^2(\Omega); u \geq 0\}$ almost everywhere in Ω and equation (2.2) has the following form:

$$2\int_\Omega \nabla y(x) \cdot \nabla_x \varphi_y(y(x))\, dx - \int_\Omega ((\varphi_y(y(x)))^-)^2\, dx + \delta^{-2}\int_\Omega (\varphi_y(y(x)))^2\, dx$$
$$+ \int_\Omega |y(x)|^2\, dx = 0, \qquad \forall y \in H_0^1(\Omega),$$

while the H_∞-suboptimal control is given by

$$u = F(y) = (\varphi_y(y(x)))^- \qquad \text{a.e. } x \in \Omega.$$

3. Proof of Theorem 1. We shall assume first that there is $F \in \mathcal{F}$ such that (1.7) is satisfied. Define on the space $L^2(R^+;U) \times L^2(R^+;W)$ the function

$$\begin{aligned} K(u,w) &= 2^{-1}\int_0^\infty (|z(t)|_Z^2 + h(u(t)) - \delta^2|w(t)|_W^2)\,dt \\ &= 2^{-1}\int_0^\infty (|C_1x(t)|_Z^2 + h(u(t)) - \delta^2|w(t)|_W^2)\,dt, \end{aligned} \tag{3.1}$$

where x is the mild solution to (1.1), $h(u) = |u|_U^2 + I_{U_0}(u)$ and $I_{U_0} : U \to (-\infty, +\infty]$ is the indicator function of U_0, i.e., $I_{U_0}(u) = 0$ for $u \in U_0$, $I_{U_0}(u) = +\infty$ for $u \bar{\in} U_0$.

Denote $\mathcal{U} = L^2(R^+;U), \mathcal{W} = L^2(R^+;W)$ and consider the problem

$$\sup_{w\in\mathcal{W}} \inf_{u\in\mathcal{U}} K(u,w). \tag{3.2}$$

We shall denote by $\varphi(x_0)$ the value of (3.2), i.e.,

$$\varphi(x_0) = \sup_{w\in\mathcal{W}} \inf_{u\in\mathcal{U}} K(u,w), \qquad x_0 \in X. \tag{3.3}$$

Clearly we have

$$2\varphi(x_0) \geq \sup\left\{\int_0^\infty (|C_1x(t)|_Z^2 - \delta^2|w(t)|_W^2)\,dt; w \in \mathcal{W}\right\} \geq 0,$$

while by closed-loop inequality (1.7) we see that

$$0 \leq 2\varphi(x_0) \leq \alpha|x_0|^2, \qquad \forall\, x_0 \in X. \tag{3.4}$$

We will prove that the function is a solution to Hamilton–Jacobi equation (2.1) in the sense precised in Theorem 1. To this aim we shall consider a family of approximately sup inf problems on the finite intervals $[0,n]$. Namely,

$$\sup_{w\in\mathcal{W}_n} \inf_{u\in\mathcal{U}_n} K_n(u,w), \qquad n = 1,2,\ldots, \tag{3.5}$$

where

$$K_n(u,w) = 2^{-1}\int_0^n (|C_1x(t)|_Z^2 + h(u(t)) - \delta^2|w(t)|_W^2)\,dt.$$

x is the corresponding solution to (1.1) on $[0,n]$ and $\mathcal{U}_n = L^2(0,n;U), \mathcal{W}_n = L^2(0,n;W)$. We denote by φ_n the corresponding value of problem (3.5), i.e.,

$$\varphi_n(x_0) = \sup_{w\in\mathcal{W}_n} \inf_{u\in\mathcal{U}_n} K_n(u,w). \tag{3.6}$$

As a supremum of the family of convex lower-semicontinuous functions

$$x_0 \to \inf\left\{2^{-1}\int_0^\infty (|C_1x(t)|_Z^2 + h(u(t)) - \delta^2|w(t)|_W^2)\,dt; u \in \mathcal{U}\right\},$$

the function φ is itself convex and lower semicontinuous. Since it is everywhere defined it is continuous on X. Similarly, the functions φ_n are convex and continuous. Moreover, we have

$$0 \leq 2\varphi_n(x_0) \leq \alpha|x_0|^2, \qquad \forall\, x_0 \in X. \qquad \square$$

LEMMA 1. *Problem* (3.5) *has at least one solution* (u_n, w_n), *which is expressed as*

$$u_n(t) = P_{U_0}(B_2^* p_n(t)); \qquad w_n(t) = -\delta^{-2} B_1^* p_n(t) \quad \text{a.e. } t \in (0,n), \tag{3.7}$$

where

$$p_n' = -A^* p_n + C_1^* C_1 x_n \quad \text{in } [0,n]; \qquad p_n(n) = 0. \tag{3.8}$$

Moreover, we have

$$\lim_{n\to\infty} \varphi_n(x_0) = \varphi(x_0), \qquad \forall\, x_0 \in X. \tag{3.9}$$

Proof. It is readily seen that for every n there exists $F \in \mathcal{F}$ such that

$$\|S_F(x_0,w)\|^2_{L^2(0,n;Z)} \le \rho^2 \|w\|^2_{\mathcal{W}_n} + \alpha |x_0|^2, \qquad \forall\, (x_0,w) \in X \times \mathcal{W}_n, \tag{3.10}$$

where $0 < \rho < \delta$ (it suffices to take in (1.7), $w = w_0$ on $(0,n)$ and $w = 0$ on (n,∞)). Then for every $w \in \mathcal{W}_n$ the minimization problem

$$\inf\{K_n(u,w); u \in \mathcal{U}_n\}$$

has a unique solution $\bar{u} = \Gamma w$, because $K(\cdot\,, w)$ is strictly convex, lower semicontinuous, coercive, and $\not\equiv +\infty$ (by (1.7)). In fact this is an optimal control problem governed by state system (1.1) and in virtue of standard results, u satisfies the Euler–Lagrange system (see e.g., [1], p. 258)

$$p' = -A^* p + C_1^* C_1 \bar{x} \quad \text{on } (0,n); \quad p(n) = 0, \tag{3.11}$$

$$B_2^* p(t) - \bar{u}(t) \in N_{U_0}(\bar{u}(t)) \quad \text{a.e. } t \in (0,n), \tag{3.12}$$

where $N_{U_0}(u)$ is the normal cone to U_0 at u and $\bar{x}$ is the corresponding solution to (1.1) with $u = \bar{u}$. Recall that (3.12) can be rewritten as

$$\bar{u}(t) = P_{U_0}(B_2^* p(t)) \quad \text{a.e. } t \in (0,n). \tag{3.13}$$

For $\bar{u} = u_n, \bar{x} = x_n$, (3.11) and (3.13) reduce to (3.8) and the first equation in (3.7), respectively. Now problem (3.5) reduces to

$$\inf\{-K(\Gamma w, w); w \in \mathcal{W}_n\}. \tag{3.14}$$

We note $\Gamma : \mathcal{W}_n \to \mathcal{U}_n$ is weakly-strongly continuous and $\Phi : \mathcal{W}_n \to R$,

$$\Phi(w) = 2^{-1} \int_0^n (|C_1 x^w(t)|_Z^2 + h(\Gamma w(t)))\, dt$$

is weakly continuous. (Here x^w is the solution to (1.1) where $u = \Gamma w$.) Indeed, if $w_k \to w$ weakly in $\mathcal{W}_n$, then we have

$$2\Phi(w_k) \le \int_0^n (|C_1 y_k(t)|_Z^2 + h(\Gamma w(t)))\, dt,$$

where y_k is the solution to (1.1) with $u = \Gamma w$ and $w = w_k$. Since by the assumption (i) $w \to x^w$ is compact from $L^2(0,n;W) = \mathcal{W}_n$ to $C([0,n];X)$, on a subsequence, again denoted k, we have

$$C_1 y_k \to C_1 x^w \quad \text{in } C([0,n];X), \quad u_k = \Gamma w_k \to \bar{u}$$

weakly in $\mathcal{U}_n$ and $C_1 x^{w_k} \to C_1 x$ in $C([0,n];X)$ where (x,u,w) satisfies system (1.1). We have

$$2^{-1}\int_0^n (|C_1x(t)|_Z^2 + h(\bar{u}(t)))\,dt \le \Phi(w) = \inf\{K_n(u,w); u \in \mathcal{U}_n\} + 2^{-1}\delta^2\|w\|_{\mathcal{W}_n}^2.$$

Hence $\bar{u} = \Gamma w, x = x^w$, and $\Gamma w_k \to \Gamma w$ strongly in $\mathcal{U}_n$ as claimed. Hence the function $w \to K(\Gamma w, w)$ is weakly lower semicontinuous on $\mathcal{W}_n$ while by (3.10) we see that

$$-2K_n(\Gamma w, w) \ge -\alpha|x_0|^2 + (\delta - \rho)\|w\|_{\mathcal{W}_n}^2, \qquad \forall\, w \in \mathcal{W}_n. \tag{3.14}'$$

This implies that problem (3.14) has at least one solution w_n. Then clearly $(u_n = \Gamma w_n, w_n)$ is a solution to problem (3.5). We shall prove now that the second equation in (3.7) holds. To this end we note that the function Φ defined above is convex and its subdifferential $\partial\Phi$ is given by

$$\partial\Phi(w) = -B_1^* p^w,$$

where p^w is the solution to system (3.11) where $\bar{x} = x^w$. Since $\partial\Phi$ is single valued we conclude that Φ is Gâteaux differentiable and so the solution w_n to (3.14) satisfies the equation

$$\partial\Phi(w_n) = -B_1^* p_n = \delta^2 w_n, \qquad \text{a.e. in } (0,n),$$

where p_n is the solution to (3.7).

To prove (3.9) we note first that

$$\varphi_n(x_0) \le \sup_{w\in\mathcal{W}} \inf_{u\in\mathcal{U}} K(u,w) = \varphi(x_0). \tag{3.15}$$

Now let $\varepsilon > 0$ and $\bar{w} \in \mathcal{W}$ be such that $\inf\{K(u,\bar{w}); u \in \mathcal{U}\} \ge \varphi(x_0) - \varepsilon$. We have

$$\begin{aligned} 2\varphi_n(x_0) &= \int_0^n (|C_1x_n(t)|_Z^2 + h(u_n(t)) - \delta^2|w_n(t)|_W^2)\,dt \\ &\ge \int_0^n (|C_1y_n(t)|_Z^2 + h(v_n(t)) - \delta^2|\bar{w}(t)|_W^2)\,dt, \end{aligned}$$

where $v_n = \arg\inf\{K_n(u,\bar{w}); u \in \mathcal{U}_n\}$ and y_n is the corresponding solution to (1.1). We have therefore

$$\int_0^n (|C_1y_n(t)|_Z^2 + h(v_n(t)))\,dt \le \int_0^n (|C_1x^{\bar{w}}(t)|_Z^2 + h(\Gamma_0\bar{w}(t)))\,dt,$$

where $\Gamma_0\bar{w} = \arg\inf\{K(u,\bar{w}); u \in \mathcal{U}\}$. On the other hand, on a subsequence, we have

$$v_n \to v \quad \text{weakly in } L^2(R^+;U),$$

$$C_1y_n \to C_1y \quad \text{weakly in } L^2(R^+;Z),$$

$$y_n \to y \quad \text{uniformly on compacta,}$$

where (y,v,w) satisfy system (1.1). This yields

$$2^{-1}\int_0^\infty (|C_1y|_Z^2 + h(v))\,dt = \inf\{K(u,\bar{w}); u \in \mathcal{U}\} + 2^{-1}\delta^2\|\bar{w}\|_{\mathcal{W}}^2.$$

Hence $v = \Gamma_0 \bar{w}$ and for $n \to \infty$,

$$2^{-1} \int_0^n (|C_1 y_n(t)|_Z^2 + h(v_n(t)))\, dt \to \inf\{K(u, \bar{w}); u \in \mathcal{U}\} + 2^{-1}\delta^2 \|\bar{w}\|_{\mathcal{W}}^2 \geq \varphi(x_0) - \varepsilon.$$

Then by (3.15) we see that

$$\varphi(x_0) - \varepsilon \leq \varphi_n(x_0) \leq \varphi(x_0)$$

for $n \geq N(\varepsilon, x_0)$, thereby completing the proof. □

Incidentally, we have also proved that (see (3.14)′)

$$\int_0^n (|C_1 x_n(t)|_Z^2 + |u_n(t)|_U^2 + |w_n(t)|_W^2)\, dt \leq C, \qquad \forall n,$$

and so on a subsequence we have

$$\begin{array}{ll} u_n \to u^* & \text{weakly in } L^2(R^+; U), \\ w_n \to w^* & \text{weakly in } L^2(R^+; W), \\ x_n(t) \to x^*(t) & \text{strongly in } X \text{ and uniformly on compacta}, \\ C_1 x_n \to C_1 x^* & \text{weakly in } L^2(R^+; Z), \end{array} \tag{3.16}$$

where (x^*, u^*, w^*) satisfy system (1.1).

For every $x_0 \in X$ we shall denote by $P_n x_0$ the set

$$P_n x_0 = \{-p_n(0)\},$$

where p_n is any solution to system (3.8) arising from a solution (x_n, u_n, w_n) to problem (3.5).

LEMMA 2. *We have*

$$P_n = \partial \varphi_n. \tag{3.17}$$

Proof. It is readily seen that $P_n x_0 \subset \partial \varphi_n(x_0)$, for all $x_0 \in X$. Indeed we have

$$\begin{aligned} 2(\varphi_n(x_0) - \varphi_n(y_0)) &\leq \int_0^n (|C_1 x_n(t)|_Z^2 + h(u_n(t)))\, dt - \int_0^n (|C_1 y_n(t)|_Z^2 + h(v_n(t)))\, dt \\ &\leq 2 \int_0^n ((C_1 x_n(t), (C_1(x_n(t) - y_n(t))_Z \\ &\quad + (\partial h(u_n(t), u_n(t) - v_n(t))_U)\, dt, \end{aligned}$$

where $v_n = \arg\inf_{u \in \mathcal{U}_n} K_n(u, w_n) = \Gamma_{w_n}$ and y_n is the solution to (1.1) where $w = w_n$ and $u = v_n$. Now if p_n is a solution to (3.8) we have

$$\begin{aligned} \varphi_n(x_0) - \varphi_n(y_0) &\leq \int_0^n ((p_n' + A^* p_n, x_n - y_n) + (B_2^* p_n, u_n - v_n)_U)\, dt \\ &= -(p_n(0), x_0 - y_0), \qquad \forall x_0, y_0 \in X, \end{aligned}$$

as claimed. To prove that $P_n = \partial \varphi_n$, it suffices to show that P_n is maximal monotone, i.e., the range $R(\lambda I + P_n)$ is all of X for some $\lambda > 0$ (see e.g., [1]). To this end let $y_0 \in X$ be arbitrary but fixed. To solve the equation

$$x_0 + P_n x_0 = y_0 \tag{3.18}$$

consider the sup inf problem

$$\sup_{w\in\mathcal{W}_n}\inf_{u\in U_n, x(0)\in X}\left\{\int_0^n (|C_1x(t)|_Z^2 + h(u(t)) - \delta^2|w(t)|_W^2)\,dt + \lambda|x(0)|^2 - 2(x(0),y_0) \text{ subject to (1.1)}\right\}. \tag{3.19}$$

Clearly the inf problem has for every $w\in\mathcal{W}_n$ a unique solution $\bar{u}=\Gamma_n w$ given by (see [1], p. 258)

$$\begin{aligned} \bar{u}(t) &= P_{U_0}(B_2^*p(t)) \quad \text{a.e. } t\in(0,n),\\ p' &= -A^*p + C_1^*C_1\bar{x} \quad \text{in } (0,n),\\ p(0) &= \lambda\bar{x}(0) - y_0; \qquad p(n) = 0. \end{aligned} \tag{3.20}$$

Now in virtue of inequality (1.7) we have for all λ sufficiently large

$$\begin{aligned} &\int_0^n (|C_1\bar{x}(t)|_Z^2 + h(\bar{x}(t)) - \delta^2|w(t)|_W^2)\,dt + \lambda|\bar{x}(0)|^2 - 2(\bar{x}(0),y_0)\\ &\le -(\delta^2-\rho^2)\int_0^n |w(t)|_W^2\,dt + \lambda|\bar{x}(0)|^2 - 2(\bar{x}(0),y_0) + \alpha|\bar{x}(0)|^2\\ &\le -\alpha_0^2\int_0^n |w(t)|_W^2\,dt, \qquad \forall w\in\mathcal{W}_n, \end{aligned} \tag{3.21}$$

because

$$\begin{aligned} &\int_0^n (|C_1\bar{x}(t)|_Z^2 + h(\bar{u}(t)))\,dt + \lambda|\bar{x}(0)|^2 - 2(\bar{x}(0),y_0)\\ &\quad\le \int_0^n |C_1\tilde{x}(t)|_Z^2\,dt \le M\int_0^n |w(t)|_W^2\,dt, \qquad \forall w\in\mathcal{W}_n, \end{aligned} \tag{3.22}$$

where $\tilde{x}' = A\tilde{x} + B_1w$; $\tilde{x}(0)=0$. Then by (3.20) and (3.22) we see that $\lambda^2|\bar{x}(0)|^2 \le M_1\int_0^n |w(t)|_W^2\,dt$, which implies (3.21) for λ sufficiently large.

This implies as in the proof of Lemma 1 that problem (3.19) has at least one solution $(\bar{u}_n,\bar{w}_n)\in\mathcal{U}_n\times\mathcal{W}_n$ given by

$$\bar{u}_n(t) = P_{U_0}(B_2^*\bar{p}_n(t)), \qquad w_n(t) = -\delta^{-2}B_1^*\bar{p}_n(t), \tag{3.23}$$

where $\bar{p}_n$ is the solution to (3.20) with $\bar{x}=\bar{x}_n$.

On the other hand, it is readily seen that $(\bar{u}_n,\bar{w}_n)$ is also the solution to problem (3.5) where $x_0=\bar{x}_n(0)$, i.e.,

$$(\bar{u}_n,\bar{w}_n) = \arg\sup_w\inf_u\{K_n(u,w);\ x(0)=\bar{x}_n(0)\}.$$

Then by (3.20) we conclude that $x_0=\bar{x}_n(0)$ is the solution to (3.18). Hence $\partial\varphi_n = P_n$ thereby completing the proof of Lemma 2. □

Consider the function $\psi_n : [0,n]\times X \to R$ defined by

$$\psi_n(t,x_0) = \sup_w\inf_u 2^{-1}\int_t^n (|C_1x|_Z^2 + h(u) - \delta^2|w|_W^2)\,dt \tag{3.24}$$

subject to $u\in L^2(t,n;U)$, $w\in L^2(t,n;W)$, and

$$x' = Ax + B_2u + B_1w \quad \text{in } [t,n]; \qquad x(t)=x_0.$$

LEMMA 3. *Let p_n be a solution to system* (3.8). *Then*

$$p_n(t) \in -\partial\psi_n(t, x_n(t)), \qquad \forall\, t \in [0, n], \tag{3.25}$$

and there is C independent of n such that

$$|p_n(t)| \leq C, \qquad \forall\, t \in [0, n]. \tag{3.26}$$

Proof. For every $t \in [0, n]$, the function $\psi_n(t, \cdot)$ is convex, continuous, and as seen in Lemma 3, $-p_n^t(t) \in \partial\psi_n(t, x_0)$, where $\partial\psi_n$ is the subdifferential of $\psi_n(t, \cdot)$ and p_n^t is the solution to

$$(p_n^t)' = -A^* p_n^t + C_1^* C_1 x_n^t \quad \text{in } [t, n]; \qquad p_n^t(n) = 0,$$

and u_n^t, w_n^t, x_n^t are optimal in (3.24). Moreover, it is readily seen (the dynamic programming principle) that if $x_0 = x_n(t)$, then $u_n^t = u_n, w_n^t = w_n, x_n^t = x_n$ on $[t, n]$ where (u_n, w_n) is the solution to problem (3.5). We may therefore infer that in this case $p^n(s) = p_n(s)$, for all $s \in [t, n]$ and

$$p_n(t) = -\partial\psi_n(t, x_n(t)), \qquad \forall\, t \in [0, n],$$

where p_n is the solution to system (3.8).

On the other hand

$$\psi_n(t, x_n(t)) \leq \psi_n(t, x_n(t) + \beta\theta) - \beta(\partial\psi_n(t, x_n(t)), \theta)$$

for all $\theta \in X$ and $\beta > 0$. This yields for $\theta = p_n(t)|p_n(t)|^{-1}$,

$$\begin{aligned} |p_n(t)| &\leq \beta^{-1}\psi_n(t, x_n(t) + \beta\theta) \leq \beta^{-1}\varphi(x_n(t) + \beta\theta) \\ &\leq \alpha\beta^{-1}|x_n(t) + \beta\theta|^2/2. \end{aligned}$$

Let $K \in L(Z, X)$ be such that $e^{(A+KC_1)t}$ is exponentially stable (1.3). Then we have, respectively,

$$x_n' = (A + KC_1)x_n + B_1 w_n + B_2 u_n - KC_1 x_n$$

and

$$(x^*)' = (A + KC_1)x^* + B_1 w^* + B_2 u^* - KC_1 x^*,$$

where (x^*, u^*, w^*) are defined by (3.16).

We have therefore

$$x^* \in L^2(R^+; X) \cap C(R^+; X), \qquad \lim_{t\to\infty} x^*(t) = 0, \tag{3.27}$$

and

$$x_n(t) \to x^*(t) \qquad \text{uniformly on } R^+. \tag{3.28}$$

Hence

$$\limsup_{n\to\infty} |p_n(t)| \leq \alpha\beta^{-1}(|x^*(t)|^2 + 2\beta|x^*(t)| + \beta^2)/2, \tag{3.29}$$

and this completes the proof of Lemma 3. □

Selecting further sequence if necessary we may assume that

$$\begin{aligned} &p_n \to p \quad \text{weak star in } L^\infty(R^+;X), \\ &p_n(t) \to p(t) \quad \text{uniformly on compact intervals,} \end{aligned} \tag{3.30}$$

where p is a mild solution to equation

$$p' = -A^*p + C_1^*C_1x^* \quad \text{in } R^+, \tag{3.31}$$

i.e.,

$$p(t) = e^{A^*(s-t)}p(s) - \int_t^s e^{A^*(\varsigma-t)}C_1^*C_1x^*(\varsigma)\,d\varsigma, \tag{3.31}'$$

for all $0 \le t < s < \infty$.

Now by (3.29) we see that

$$2|p(t)| \le \alpha\beta^{-1}(|x^*(t)|^2 + 2\beta|x^*(t)| + \beta^2), \qquad \forall\, t \ge 0, \beta > 0,$$

and by (3.27) we infer that $\lim_{t\to\infty} |p(t)| \le \alpha\beta/2$ for all $\beta > 0$. Hence

$$\lim_{t\to\infty} p(t) = p(\infty) = 0. \tag{3.32}$$

Note also that by Lemma 1 and (3.16), (3.30) we have

$$u^*(t) = P_{U_0}(-B_2^*p(t)), \qquad w^*(t) = -\delta^{-2}B_1^*p(t), \qquad \text{a.e. } t > 0. \tag{3.33}$$

We shall denote by $P : X \to X$ the mapping defined by

$$Px_0 = \{-p(0)\}, \qquad \forall\, x_0 \in X,$$

where $p \in C(R^+;X)$ is any solution to (3.31), (3.32).

LEMMA 4. *We have* $\partial\varphi = P$.

Proof. Letting n tend to $+\infty$ in $p_n(0) \in -\partial\varphi_n(x_0)$, i.e.,

$$-(p_n(0), x_0 - y_0) \ge \varphi_n(x_0) - \varphi_n(y_0), \qquad \forall\, y_0 \in X,$$

it follows by Lemmas 1 and 2 that $-p(0) \in \partial\varphi(x_0)$, i.e., $P \subset \partial\varphi$. To prove that $P = \partial\varphi$, it suffices to show that $R(I + P) = X$, i.e., for every $y_0 \in X$ the equation

$$x_0 + Px_0 \ni y_0 \tag{3.34}$$

has at least one solution. Since by Lemma 2, $R(I + P_n) = X$ for all n, the equation $x + P_nx \ni y_0$ has a unique solution: x_0^n, i.e.,

$$x_0^n - p_n(0) = y_0,$$

where p_n is a solution to system (3.8) with $x(0) = x_0^n$. Since $(I + P_n)^{-1}$ is nonexpansive on X, $\{x_0^n\}$ is bounded and so on a subsequence, $x_0^n \to x_0$ weakly in X. Then clearly the corresponding sequence of solutions (u_n, w_n) to problem (3.5) is bounded in $L^2(R^+;U) \times L^2(R^+;W)$, and so we may assume (3.16) holds and this implies as above that

$$p_n(t) \to p(t) \quad \text{uniformly on every } [0,T],$$

where p is a solution to (3.31), (3.32). Hence $x_0 - p(0) = y_0$ and so x_0 is the solution to (3.34) as desired.

Now by (3.31)$'$ and by assumption (i) we see that $\partial\varphi = P$ is compact, i.e., maps bounded sets in compacta.

Let $x_0 \in D(A)$ be arbitrary but fixed and let η be any element of $\partial\varphi(x_0)$. According to Lemma 4 there are $(x^*, u^*, w^*) \in L^2(R^+; X) \times L^2(R^+; U) \times L^2(R^+; W)$ and $p \in C(R^+; X)$ such that

$$(x^*)' = Ax^* + B_2 P_{U_0}(-B_2^* p) - \delta^{-2} B_1 B_1^* p \quad \text{in } R^+,$$

$$p' = -A^* p + C_1^* C_1 x^* \quad \text{in } R^+;$$

$$x^*(0) = x_0, \qquad p(\infty) = 0,$$

$$p(0) = \eta.$$

By assumption (ii) we see that x^* and p are strong solutions to these equations. Then multiplying the first equation by p' the second by $(x^*)'$ and subtracting the results we get

$$2(Ax^*(t), p(t)) - \delta^{-2}|B_1^* p(t)|^2 - h^*(-B_2^* p(t)) - |C_1 x^*(t)|^2 = 0, \qquad \forall\, t \geq 0,$$

where

$$h^*(v) = \sup\{(v, x) - h(x); x \in X\} = 2^{-1}(|v|_U^2 - |v - P_{U_0}(v)|_U^2).$$

Now letting t tend to zero, we get (2.2). On the other hand, by (3.25) we have

$$(p(t), x^*(t) - y_0) \geq \varphi(x^*(t) - \varphi(y_0)), \qquad \forall\, t \geq 0, y_0 \in X,$$

because by Lemma 1 we know that

$$\lim_{n\to\infty} \psi_n(t, x) = \sup_w \inf_u \left\{ \frac{1}{2} \int_t^\infty (|C_1 x|_Z^2 + h(u) - \delta^2 |w|_W^2)\, dt, u \in L^2(t, \infty; U), \right.$$

$$\left. w \in L^2(t, \infty; W); x' = Ax + B_2 u + B_1 w \text{ in } (t, \infty), x(t) = x_0 \right\}$$

$$= \varphi(x_0).$$

Hence $-p(t) \in \partial\varphi(x^*(t))$, for all $t \geq 0$ and so by (3.27) we conclude that x^* is a mild solution to multivalued equation (2.3) satisfying (2.4).

Let us assume now that (2.2) has a solution φ which is convex, continuous with $\partial\varphi$, compact, and which satisfies (2.2). We must prove that the feedback control $F = P_{U_0}(-B_2^* \partial\varphi)$ belongs to $\mathcal{F}$ and

$$|S_F(x_0, w)|^2 \leq \delta^2 \|w\|_{L^2(R^+;W)}^2 + \alpha |x_0|^2 \tag{3.35}$$

for all $(x_0, w) \in X \times L^2(R^+; W)$.

Consider the Cauchy problem

$$x' = Ax + B_2 P_{U_0}(-B_2^* \partial\varphi(x)) + B_1 w; \qquad x(0) = x_0 \tag{3.36}$$

for $x_0 \in X$ and $w \in L^2(R^+; W)$. We shall prove first that (3.36) has a mild solution x on R^+.

Since the mapping $x \to B_2 P_{U_0}(-B_2^* \partial\varphi)$ is not convex-valued, the standard existence theory does not apply in this case. However we consider the approximating equation

$$x'_\lambda = Ax_\lambda + B_2 P_{U_0}(-B_2^* \nabla\varphi_\lambda(x_\lambda)) + B_1 w; \qquad x_\lambda(0) = x_0,$$

where $\nabla\varphi_\lambda = \lambda^{-1}(I - (I + \lambda\partial\varphi)^{-1})$ (see [1]). Since $\nabla\varphi_\lambda$ is Lipschitzian, this equation has a unique mild solution x_λ. Moreover, by (2.2) it is readily seen that

$$\sup\{|\eta|; \eta \in \partial\varphi(x)\} \leq C(|x| + 1) \qquad \forall x \in X$$

and therefore

$$|\nabla\varphi_\lambda(x_\lambda)| \leq C_1(x_\lambda| + 1), \qquad \forall \lambda > 0.$$

This implies that $\{x_\lambda\}$ is uniformly bounded on every interval $[0, T]$, and by standard compacity arguments (see e.g., [7]) we infer that $\{x_\lambda\}$ is uniformly convergent on every interval to a continuous function $x = x(t)$. Since $\nabla\varphi_\lambda(x_\lambda) \in \partial\varphi((I + \lambda\partial\varphi)^{-1}x_\lambda)$ and $\partial\varphi$ is compact, we have on a subsequence

$$P_{U_0}(-B_2^* \nabla\varphi_\lambda(x_\lambda)) \to P_{U_0}(-B_2^*\eta(t)), \qquad \text{a.e. } t > 0,$$

where $\eta(t) \in \partial\varphi(x(t)), t \geq 0$. This implies that x is a mild solution to (3.36). Since e^{At} is analytic, the function x is a strong solution to (3.36) (i.e., is absolutely continuous and satisfies almost everywhere this equation).

Now multiplying (3.36) by $\eta(t) \in \partial\varphi(x(t))$ and using (2.2), we get

$$\begin{aligned} \frac{d}{dt}\varphi(x(t)) &= (Ax(t) + B_2 P_{U_0}(-B_2^*\eta(t)), \eta(t)) + (w(t), B_1^*\eta(t))_W \\ &= -2^{-1}|C_1 x(t)|_Z^2 + \delta^{-2}|B_1^*\eta(t)|_W^2 - |P_{U_0}(-B_2^*\eta(t))|_U^2 \\ &\quad + (w(t), B_1^*\eta(t))_W, \qquad \forall t \geq 0. \end{aligned} \tag{3.37}$$

Integrating (3.37) on $(0, t)$ we get

$$\begin{aligned} &\varphi(x(t)) + 2^{-1}\int_0^t (|C_1 x(s)|_Z^2 + |u(s)|_U^2 - \delta^{-2}|B_1^*\eta(s)|_W^2)\, ds \\ &\quad = \varphi(x_0) + \int_0^t (w(s), B_1^*\eta(s))W\, ds, \qquad \forall t \geq 0, \end{aligned} \tag{3.38}$$

and therefore

$$\int_0^t (|C_1 x(s)|_Z^2 + |u(s)|_U^2)\, ds \leq C, \qquad \forall t \geq 0. \tag{3.39}$$

On the other hand, we may write (3.36) as

$$x' = (A + KC_1)x - KC_1 x + B_2 u + B_1 w, \qquad x(0) = x_0,$$

where $e^{(A+KC_1)t}$ is exponentially stable. Then by estimate (3.39) we see that

$$x \in L^2(R^+; W) \quad \text{and} \quad \lim_{t\to\infty} x(t) = 0.$$

Since in the previous argument $B_1 w$ can be replaced by any L^2 function f, we conclude that $F \in \mathcal{F}$. We note that (3.38) holds for any mild solution x to system (1.1) where $u(t) \in P_{U_0}(-B_2^*\partial\varphi(x(t)))$ almost everywhere $t > 0$. This yields

$$\frac{1}{2}\int_0^\infty (|C_1 x(t)|_Z^2 + h(u(t)))\, dt \leq \varphi(x_0) + \frac{\delta^2}{2}\int_0^\infty |w(t)|_W^2\, dt, \qquad \forall w \in L^2(R^+; W)$$

as claimed. The proof of Theorem 1 is complete. □

Acknowledgments. The author is indebted to Professors A. Halanay and R. Curtain for useful discussions concerning the subject of H_∞-control theory. We also thank the anonymous referee for some suggestions concerning the content of this paper.

REFERENCES

[1] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, the Netherlands, 1986.

[2] J. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_2 and H_∞-control problems*, IEEE Trans. Automat. Control, AC34 (1989), pp. 831–847.

[3] A. ICHIKAWA, *H_∞-control and min-max problems in Hilbert spaces*, to appear.

[4] B. VAN KEULEN, M. PETERS, AND R. CURTAIN, *H_∞-control with state feedback: The infinite dimensional case*, J. Math. Systems, Estimation Control, 3 (1993), pp. 1–39.

[5] G. TADMOR, *Worst case design in the time domain: The maximum principle and the standard H_∞ problem*, Math. Control Signals Systems, 3 (1990), pp. 301–324.

[6] ———, *The standard H_∞ problem and the maximum principle: The general linear case*, SIAM J. Control Optim., 31 (1993), pp. 813–846.

[7] I. VRABIE, *Compactness Methods for Nonlinear Evolutions*, Longman, London, 1987.

[8] K. ZHOU AND P. P. KHARGONEKAR, *An algebraic equation approach to H_∞ optimization*, Systems Control Lett., 11 (1988), pp. 85–91.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 965–974, July 1994

004

POSITIVE DEPENDENCE OF A CLASS OF MULTIVARIATE EXPONENTIAL DISTRIBUTIONS*

INGRAM OLKIN† AND Y. L. TONG‡

Abstract. The positive dependence of a subclass of multivariate exponential distributions is examined. This class is characterized by an index vector $\mathbf{k}$ and a parameter vector $\boldsymbol{\lambda}$, which are used as an ordering to yield degrees of positive dependence. The results presented have a direct implication on the reliability function of a system and the survival probability function of a shock model, and consequently on the optimal assembly of systems.

Key words. system reliability, component systems, shock models, majorization, Schur-convex functions

AMS subject classifications. 62N05, 60K10, 90B25

1. Introduction.

There is a large literature dealing with the role of multivariate exponential distributions in reliability theory. (Here we use the term multivariate exponential distribution to mean any joint distribution with univariate exponential marginal distributions.) Early attention was focused on probabilistic models for generating bivariate exponential distributions. These arise from shock models, physical applications, statistical considerations, and so on (see, e.g., [Freund (1961)], [Marshall and Olkin (1967a), (1967b)], [Downton (1970)], [Block and Basu (1974)], [Friday (1976)], [Friday and Patil (1977)]). Some of these bivariate exponential distributions extend naturally to the multivariate case [Marshall and Olkin (1967a)], [Arnold (1968)], [Proschan and Sullo (1974)], [Basu and Block (1975)]. A key feature of these distributions is that the number of parameters is large, thereby creating difficulties in statistical inference. For small dimensionality some headway has been made. Statistical estimation of the parameters was studied in [Proschan and Sullo (1976)], and a test for independence was proposed in [Al-Saadi and Young (1982)]. For a comprehensive general review of the literature on multivariate exponential distributions, see [Johnson and Kotz (1972), Chap. 41], [Friday (1976)], [Friday and Patil (1977)], [Block (1985)], and [Marshall and Olkin (1985)].

For two nonnegative random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{X}' = (X_1', \ldots, X_n')$ with the same marginal distributions, we say that $\mathbf{X}$ is more positively dependent than $\mathbf{X}'$ if

$$P\left(\bigcap_{i=1}^{n} \{X_i \in \mathcal{B}\}\right) \geq P\left(\bigcap_{i=1}^{n} \{X_i' \in \mathcal{B}\}\right) \quad \text{for all Borel sets } \mathcal{B} \tag{1.1}$$

(see [Tong (1989)]). In effect, by taking $\mathcal{B}$ to be an interval in $\mathcal{R}$ whose indicator function is monotonic, more positive dependence concentrates probabilities more heavily in the permutation symmetric lower and upper orthants.

Positive dependence is an important characteristic of random variables, and multivariate exponential distributions have been a source of considerable study. In particular, we determine configurations of the parameters for which comparisons of positive dependence can be made.

In this paper we study a subclass of multivariate exponential distributions of [Marshall and Olkin (1967a)], characterized by an index vector $\mathbf{k}$ and a parameter vector $\boldsymbol{\lambda}$, for which we obtain results on positive dependence. Examples of parallel-series systems and shock models are shown to fit in this subclass when $\mathcal{B}$ is the set $\mathcal{B} = \{x : x \leq t\}$ or $\mathcal{B} = \{x : x > t\}$ for

* Received by the editors August 12, 1991; accepted for publication (in revised form) January 28, 1993.

† Department of Statistics, Stanford University, Stanford, California 94305. The research of this author was supported in part by National Science Foundation grant DMS-9002502.

‡ School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332. The research of this author was supported in part by National Science Foundation grants DMS-8801327, A01, and DMS-9001721.

arbitrary but fixed t (§2). In §3 we obtain results concerning positive dependence via a partial ordering of the index vectors and a partial ordering of the parameter vectors. These results have a direct implication on the reliability function of a system and the survival probability function of a shock model.

To fix notation we say that a random variable X is exponentially distributed with parameter λ if $\overline{G}(x) = P\{X > x\} = e^{-\lambda x}$. For a multivariate distribution of $(X_1, \ldots, X_n)$ we write $F(x_1, \ldots, x_n) = P(\cap_{i=1}^n \{X_i \le x_i\})$ and for simplicity $F(t, \ldots, t) \equiv P(\cap_{i=1}^n \{X_i \le t\})$; the survival function is denoted $\overline{F}(x_1, \ldots, x_n) = P(\cap_{i=1}^n \{X_i > x_i\})$ or $\overline{F}(t, \ldots, t) = P(\cap_{i=1}^n \{X_i > t\})$.

2. A class of multivariate exponential distributions with a common univariate marginal distribution. To motivate the subclass considered, we briefly review the Marshall–Olkin (M–O) multivariate exponential for the trivariate case [Marshall and Olkin (1967a), §4]. Let U_1, U_2, U_3, V_{12}, V_{13}, V_{23}, W_{123} denote independent exponential random variables and let

$$\begin{aligned} X_1 &= \min(U_1, V_{12}, V_{13}, W_{123}), \\ X_2 &= \min(U_2, V_{12}, V_{23}, W_{123}), \\ X_3 &= \min(U_3, V_{13}, V_{23}, W_{123}). \end{aligned} \tag{2.1}$$

Then (X_1, X_2, X_3) has the M–O trivariate exponential distribution. The key point to note is that this formulation requires $2^n - 1$ independent random variables to generate an n-variate exponential distribution. However, not all component systems require the full range of variables. For example,

$$\begin{aligned} X_1' &= \min(U_1, V_{12}, W_{123}), \\ X_2' &= \min(U_2, V_{12}, W_{123}), \\ X_3' &= \min(U_3, W_{123}) \end{aligned} \tag{2.2}$$

is a subclass of (2.1) and requires only five instead of seven variables to generate the trivariate model.

Thus, the motivation for the subclass being considered is exactly to define a subclass of models that might be useful in some applications. One subclass is the following: Let $U_1, \ldots, U_n$, $V_1, \ldots, V_n$, and W be independent univariate exponential random variables with $EU_i = \lambda_1^{-1}$, $EV_i = \lambda_2^{-1}$ $(i = 1, \ldots, n)$, and $EW = \lambda_0^{-1}$. Let $\mathbf{k} = (k_1, \ldots, k_n)$ be a vector of nonnegative integers with

$$\sum_{s=1}^{n} k_s = n, \qquad k_1 \ge \cdots \ge k_r \ge 1, \qquad k_{r+1} = \cdots = k_n = 0, \tag{2.3}$$

for some $r \le n$. In the results given below, the monotonicity of the k_i's in (2.3) is not essential, as we shall see; this condition is made primarily for notational convenience. For given $\mathbf{k}$ let $\mathbf{X} \equiv \mathbf{X}(\mathbf{k}) = (X_1, \ldots, X_n)$ be an n-dimensional multivariate exponential random vector defined by

$$X_j = \begin{cases} \min(U_j, V_1, W), & j = 1, \ldots, K_1, \\ \min(U_j, V_2, W), & j = k_1 + 1, \ldots, K_2, \\ \quad \vdots \qquad \vdots & \\ \min(U_j, V_r, W), & j = K_{r-1} + 1, \ldots, n, \end{cases} \tag{2.4}$$

where $K_0 = 0$, $K_1 = k_1$, $K_2 = k_1 + k_2, \ldots, K_{r-1} = k_1 + \cdots + k_{r-1}$. That is, each of the X_j's depends on a different U_j and a common W, the first k_1 depend on V_1, the next k_2

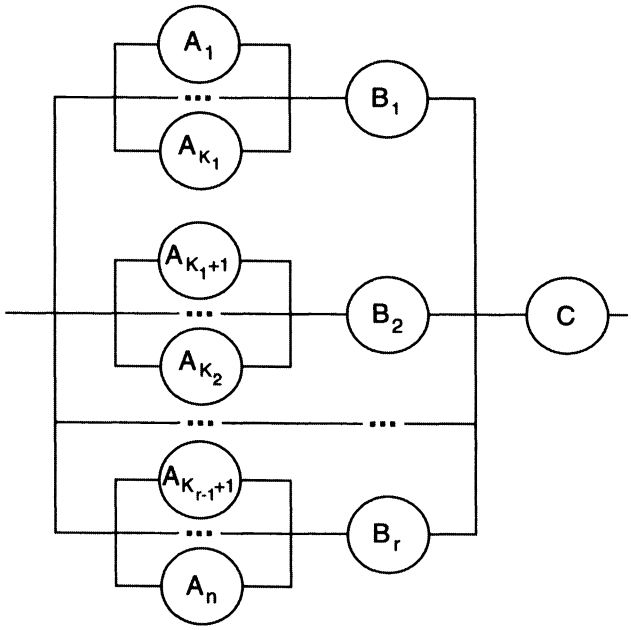


FIG. 1. *A parallel-series system.*

depend on V_2, and so on. From the construction the univariate marginal distributions of the X_j's are exponential with a mean $\lambda^{-1} \equiv (\lambda_1 + \lambda_2 + \lambda_0)^{-1}$.

The joint distribution of the X_i's is exchangeable only when $\mathbf{k} = (n, 0, \ldots, 0)$ or $\mathbf{k} = (1, 1, \ldots, 1)$. Furthermore, the construction suggests that the components $X_j = \min(U_j, V_1, W)$, $j = 1, \ldots, n$ of $\mathbf{X}(n, 0, \ldots, 0)$, are more positively dependent (in the sense of (1.1)) than the components $X_j = \min(U_j, V_j, W)$, $j = 1, \ldots, n$ of $\mathbf{X}(1, 1, \ldots, 1)$ because the former depends on the same variable V_1, whereas the latter permits the V_j's to differ. (An analytical proof is provided in [Shaked and Tong (1985)].)

We now compare the positive dependence for intermediate cases of the components of $\mathbf{X}(\mathbf{k})$. Let

$$F_{\mathbf{k},\boldsymbol{\lambda}}(t, \ldots, t) \equiv P_{\mathbf{k},\boldsymbol{\lambda}}\left[\bigcap_{i=1}^{n} \{X_i \leq t\}\right], \qquad \overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(t, \ldots, t) = P_{\mathbf{k},\boldsymbol{\lambda}}\left[\bigcap_{i=1}^{n} \{X_i > t\}\right], \tag{2.5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_0)$. We first examine how these probability functions depend on λ_1, λ_2, and λ_0 when both $\mathbf{k}$ and the sum $\lambda = \lambda_0 + \lambda_1 + \lambda_2$ are kept fixed.

Example 2.1. *System reliability.* Suppose that three types of components labeled A, B, and C are connected in a parallel-series fashion such that subsets of size k_s of the type A components comprise a subsystem connected in parallel, the s-th subsystem connected in series to a component $B_s (s = 1, \ldots, r)$; all of the type B components are connected to a type C component in parallel as shown in Fig. 1. If the lifelength distributions of the components A, B, and C are exponential with means λ_1^{-1}, λ_2^{-1}, and λ_0^{-1}, respectively, then the system reliability is simply

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) \equiv P_{\mathbf{k},\boldsymbol{\lambda}}\{\text{lifelength of system } > t\} = 1 - F_{\mathbf{k},\boldsymbol{\lambda}}(t, \ldots, t). \tag{2.6}$$

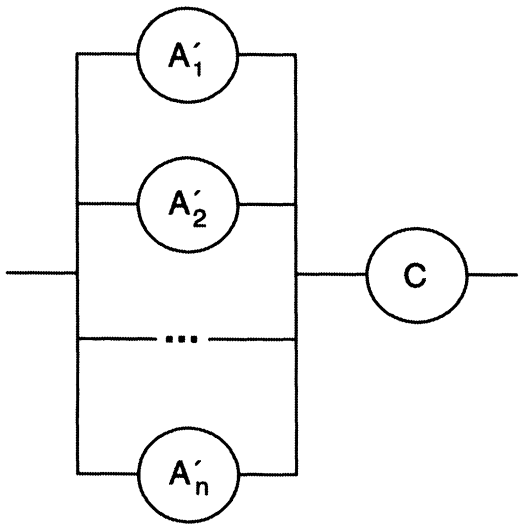


FIG. 2. *The case* $k_1 = \cdots = k_n = 1$.

When $k_1 = \cdots = k_n = 1$ so that $\mathbf{k} = (1, 1, \ldots, 1)$, then the system in Fig. 1. reduces to that in Fig. 2, where the lifelengths of components $A_1', \ldots, A_n'$ have independent exponential distributions with a common mean $(\lambda_1 + \lambda_2)^{-1}$.

Example 2.2. *Shock models.* In a fatal shock model suppose that the components of an n-component system die after receiving a fatal shock from one of several sources and that independent Poisson processes govern the occurrence of shocks. If (i) the failure times for components of type A, B, C, governed by shocks, are independent exponential variables with means λ_1^{-1}, λ_2^{-1}, λ_0^{-1}, respectively; (ii) a separate type A shock applies to each of the n components; (iii) shocks of type B affect groups of $k_1, k_2, \ldots, k_r$ components; and (iv) a shock to type C applies to all the n components, then the survival probability function of concern is

$$\begin{aligned}\overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(x_1, \ldots, x_n) &= P_{\mathbf{k},\boldsymbol{\lambda}}\left[\bigcap_{i=1}^{n}\{X_i > x_i\}\right] \\ &= \exp\left[-\lambda_1 \sum_{i=1}^{n} x_i - \lambda_2 \sum_{s=1}^{r} \max(x_{K_{s-1}+1}, \ldots, x_{K_s}) - \lambda_0 \max_{1\le i\le n} x_i\right],\end{aligned} \tag{2.7}$$

where $K_0 \equiv 0$. When $x_1 = \cdots = x_n = t$, (2.7) reduces to

$$\overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(t, \ldots, t) = \exp[-n\lambda_1 t - r\lambda_2 t - \lambda_0 t]. \tag{2.8}$$

It should be noted that in general we need not restrict our attention to just three types of components A, B, and C. Although our model given in (2.4) can be modified for any number of types of components, we have not obtained manageable analytical results for the general case. (In the general case there are several W variables, and a corresponding conditioning argument, given their values, becomes complicated.) Thus, throughout this paper we consider only three types of components as modeled in (2.4).

3. Positive dependence properties. In this section we present analytic results concerning how the positive dependence of the random variables $X_1(\mathbf{k}), \ldots, X_n(\mathbf{k})$ depends on $\mathbf{k}$ and

$\boldsymbol{\lambda}$. In particular, we show that when the random variables are more positively dependent in a fashion to be defined, then they are more concentrated in the sense that both $F_{\mathbf{k},\boldsymbol{\lambda}}(t,\dots,t)$ and $\overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(t,\dots,t)$ defined in (2.5) become larger.

Positive dependence of the components $\mathbf{X}(\mathbf{k})$ depend on $\mathbf{k}$ and $\boldsymbol{\lambda}$, and in particular on the number of variables r used in the generation (2.4) of the joint distribution. We first show that if the parameters remain fixed, then we obtain more positive dependence if $\mathbf{k}$ is ordered by majorization. (For details of majorization, see [Marshall and Olkin, (1979)].)

THEOREM 3.1. *Let n, $\boldsymbol{\lambda}$, and t be arbitrary but fixed, and let $\mathbf{k}, \mathbf{k}'$ be two vectors satisfying* (2.3). *If $\mathbf{k} \succ \mathbf{k}'$, where $\succ$ denotes the majorization ordering, then*

$$\overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(t,\dots,t) \geq \overline{F}_{\mathbf{k}',\boldsymbol{\lambda}}(t,\dots,t). \tag{3.1}$$

Proof. The result follows from (2.8) and the fact that if

$$\mathbf{k} = (k_1,\dots,k_r,k_{r+1},\dots,k_n) \succ (k_1',\dots,k_{r'}',k_{r'+1}',\dots,k_n') = \mathbf{k}',$$

where $k_1 \geq \cdots \geq k_r \geq 1$, $k_{r+1} = \cdots = k_n = 0$, $k_1' \geq \cdots \geq k_{r'}' \geq 1$, $k_{r'+1}' = \cdots = k_n' = 0$, then $r \leq r'$. □

We next compare two distributions for which $\mathbf{k}$ remains fixed but the parameters $\boldsymbol{\lambda}$ change. To do this we require the definition of a "decreasing transformation."

DEFINITION 3.2. *Let $\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\lambda_0)$ and $\boldsymbol{\lambda}^* = (\lambda_1^*,\lambda_2^*,\lambda_0^*)$ denote two vectors of parameters. The vector $\boldsymbol{\lambda}^*$ is said to be a* decreasing transformation *of $\boldsymbol{\lambda}$ (denoted $\boldsymbol{\lambda} \overset{t}{>} \boldsymbol{\lambda}^*$) if $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}^*$ and*

$$\lambda_1 \leq \lambda_1^*, \quad \lambda_1+\lambda_2 \leq \lambda_1^*+\lambda_2^*, \quad \textit{and} \quad \lambda_1+\lambda_2+\lambda_0 = \lambda_1^*+\lambda_2^*+\lambda_0^*.$$

Note that the distinction between decreasing transformation and majorization is that the former does not require an ordering of the elements. An important aspect of a decreasing transformation is that the marginal distributions remain unchanged. We show in Theorem 3.4 that the random variables are more positively orthant dependent (for definition see, e.g., [Tong (1990), p. 102]) under $\boldsymbol{\lambda}^*$. The following example may serve to illustrate the structure.

Example 3.3. Suppose $n = 4$, $\mathbf{k} = (2,2,0,0)$, and $\boldsymbol{\lambda} = (2,5,6) \overset{t}{>} (4,3,6) = \boldsymbol{\lambda}^*$. Then the distribution of the lifelength of the system under $\boldsymbol{\lambda}$ and under $\boldsymbol{\lambda}^*$ is as in Figs. 3 and 4, respectively. Here the lifelengths of the A, B, and C components are independent exponential variables with means $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{6}$, respectively. In Fig. 3, X_1 and X_2 involve the same component A_{12} and X_3 and X_4 involve the same component A_{32}, so that it is intuitively clear that the system of Fig. 3 is more positively dependent than that of Fig. 4.

THEOREM 3.4. *Let n, $\mathbf{k}$, and t be arbitrary but fixed, and let*

$$\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\lambda_0) \quad \textit{and} \quad \boldsymbol{\lambda}^* = (\lambda_1^*,\lambda_2^*,\lambda_0^*) \tag{3.2}$$

be two parameter vectors. If $\lambda \overset{t}{>} \lambda^*$, then

$$\overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(x_1,\dots,x_n) \geq \overline{F}_{\mathbf{k},\boldsymbol{\lambda}^*}(x_1,\dots,x_n) \quad \textit{for all } (x_1,\dots,x_n) \in \mathbb{R}_+^n. \tag{3.3}$$

Proof. The proof follows from (2.7) and the fact that

$$\sum_{i=1}^{n} x_i \geq \sum_{s=1}^{r} \max(x_{K_{s-1}+1},\dots,x_{K_s}) \geq \max_{1\leq i\leq n} x_i. \qquad \square$$

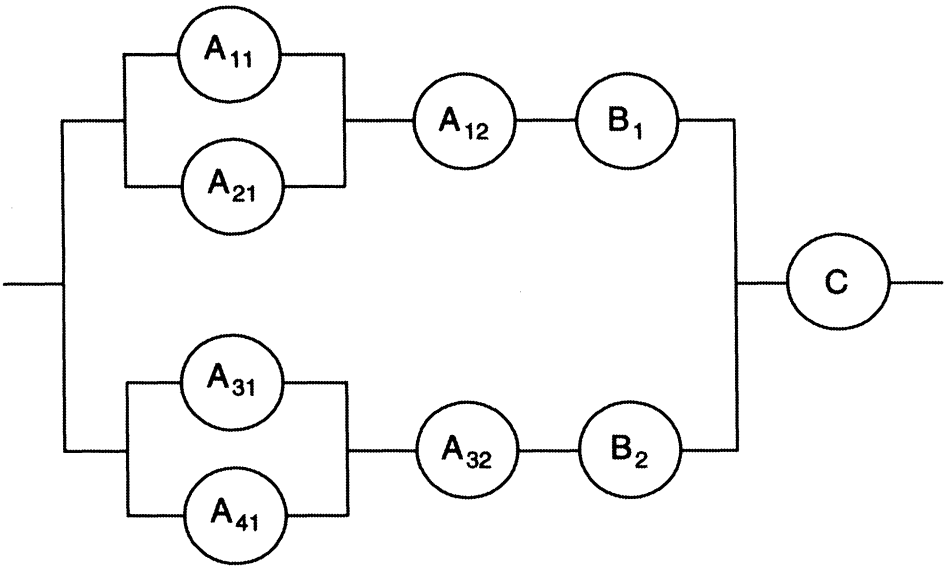


FIG. 3. *A parallel-series system.*

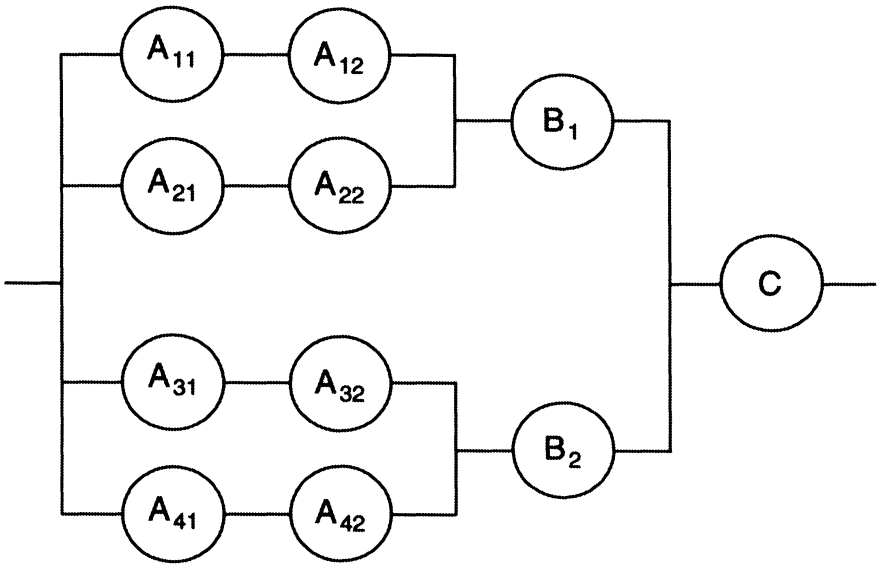


FIG. 4. *A parallel-series system.*

It follows from Theorem 3.4 that

$$\overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(t,\ldots,t) \geq \overline{F}_{\mathbf{k},\boldsymbol{\lambda}^*}(t,\ldots,t) \quad \text{for all } t \in (0,\infty). \tag{3.4}$$

We now examine in more detail the function $F_{\mathbf{k},\boldsymbol{\lambda}}(t,\ldots,t)$ and its reliability function $R_{\mathbf{k},\boldsymbol{\lambda}}(t) =$

$1 - F_{\mathbf{k},\boldsymbol{\lambda}}(t,\ldots,t)$. In particular, if (2.5) prevails, then

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) = \overline{G}_0(t)\left[1 - \prod_{s=1}^{r}\{1 - \overline{G}_2(t)[1 - G_1^{k_s}(t)]\}\right], \tag{3.5}$$
$$\overline{G}_j(t) = e^{-\lambda_j t}, \quad G_j(t) = 1 - \overline{G}_j(t) \quad \text{for } j = 0, 1, 2.$$

THEOREM 3.5. *Let n, $\boldsymbol{\lambda}$, and t be arbitrary but fixed, and let $\mathbf{k}, \mathbf{k}'$ satisfy* (2.3). *If $\mathbf{k} \succ \mathbf{k}'$, then*

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) \le R_{\mathbf{k}',\boldsymbol{\lambda}}(t). \tag{3.6}$$

Proof. Although this theorem can be derived by applying a more general result in [Tong (1989)], here we give an independent proof that more clearly illustrates the structure using the specific expression in (3.5). Consider the function

$$h(\mathbf{k}) \equiv \log \prod_{s=1}^{r}\{1 - c_2(1 - c_2^{k_s})\}$$
$$= \sum_{s=1}^{r} \log\{\bar{c}_2 + c_2 e^{k_s \log c_1}\} \equiv \sum_{s=1}^{r} \phi(k_s),$$

where $c_1 = G_1(t)$, $c_2 = 1 - \bar{c}_2 = \overline{G}_2(t)$ are in $(0,1)$, $\phi(u) = \log\{\bar{c}_2 + c_2 e^{\rho u}\}$. Because

$$\phi''(u) = c_2\rho^2\bar{c}_2 e^{\rho u}(\bar{c}_2 + c_2 e^{\rho u})^{-2} > 0$$

for all $\rho = \log c_1 < 0$ and $u \ge 0$, the function $h(\mathbf{k})$, and hence $\prod_{s=1}^{r}\{1 - \overline{G}_2(t)(1 - G_1^{k_s}(t))\}$, is a Schur-convex function of $\mathbf{k}$. (See [Marshall and Olkin (1979), p. 11].) Thus, $R_{\mathbf{k},\boldsymbol{\lambda}}(t)$ is a Schur-concave function of $\mathbf{k}$. □

Note that $F_{\mathbf{k},\boldsymbol{\lambda}}(t,\ldots,t) = 1 - R_{\mathbf{k},\boldsymbol{\lambda}}(t)$ is a Schur-convex function of $\mathbf{k}$.

The practical implication of Theorem 3.5 is that the reliability increases as $\mathbf{k}$ moves toward the uniform distribution in the sense of majorization. This has a direct application in the optimal assembly of systems, as is illustrate further in §4.

The next result is an analog to Theorem 3.5.

THEOREM 3.6. *Let n, $\mathbf{k}$, and t be arbitrary but fixed. If $\boldsymbol{\lambda} \overset{t}{>} \boldsymbol{\lambda}^*$, then*

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) \le R_{\mathbf{k},\boldsymbol{\lambda}^*}(t). \tag{3.7}$$

Proof. Define a parameter vector $\boldsymbol{\lambda}'$ of positive elements given by

$$\boldsymbol{\lambda}' = (\lambda_1', \lambda_2', \lambda_0') \equiv (\lambda_1, \lambda_1^* + \lambda_2^* - \lambda_1, \lambda_0^*).$$

We show that

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) \le R_{\mathbf{k},\boldsymbol{\lambda}'}(t) \le R_{\mathbf{k},\boldsymbol{\lambda}^*}(t).$$

The second inequality follows as a special case from the discussion in [Tong (1989), Ex. 3.1] with $B = \{x : x \le t\}$ and $g(u,v,w) = \min(u,v,w)$. To prove the first inequality, let $\delta = \lambda_2' - \lambda_2 \ge 0$. Then

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) = e^{-\lambda_0 t}\left[1 - \prod_{s=1}^{r}(1 - z_s e^{-\lambda_2 t})\right] \tag{3.8}$$
$$= e^{-\delta t}e^{-\lambda_0' t}\left[1 - \prod_{s=1}^{r}(1 - z_s e^{\delta t}e^{-\lambda_2' t})\right],$$

where $z_s \equiv z_s(t) = 1 - G_1^{k_s}(t)\,(s = 1, \ldots, r)$ are arbitrary but fixed, $z_s \in [0, 1]$. For notational convenience, let $c \equiv e^{\delta t} \geq 1$, $\omega \equiv e^{-\lambda_2' t} < 1$, so that

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) = \frac{1}{c}\, e^{-\lambda_0' t} \left[1 - \prod_{s=1}^{r} (1 - c\omega z_s)\right]. \tag{3.9}$$

In (3.9) let $y_s \equiv \omega z_s \leq 1$, $s = 1, \ldots, r$, so that inequality (3.7) becomes

$$\frac{1}{c}\left[1 - \prod_{j=1}^{r} (1 - cy_j)\right] \leq \left[1 - \prod_{j=1}^{r} (1 - y_j)\right]. \tag{3.10}$$

That (3.10) holds follows by an inductive argument: it holds for $r = 1$ and 2 (for $r = 1$ it becomes an equality). Suppose that it holds for $r - 1$. Then

$$\begin{aligned}
\frac{1}{c}\left[1 - \prod_{j=1}^{r} (1 - cy_j)\right] &= \frac{1}{c}\left[1 - \prod_{j=1}^{r-1} (1 - cy_j)\right](1 - cy_r) + y_r \\
&\leq \left[1 - \prod_{j=1}^{r-1} (1 - y_j)\right](1 - cy_r) + y_r \\
&\leq \left[1 - \prod_{j=1}^{r-1} (1 - y_j)\right](1 - y_r) + y_r \\
&= 1 - \prod_{j=1}^{r} (1 - y_j),
\end{aligned}$$

where the first inequality is based on the induction hypothesis and the second inequality holds from the reliability context in that $1 - e^{\lambda_2 t} z_r > 0$. □

To illustrate an application of Theorems 3.5 and 3.6, consider the following specific example.

Example 3.7. For fixed n and $\mathbf{k}$ let $\boldsymbol{\lambda} = (3, 2, 8)$, $\boldsymbol{\lambda}^* = (5, 2, 6)$. From

$$\boldsymbol{\lambda} \overset{t}{>} \boldsymbol{\lambda}' = (3, 4, 6) \overset{t}{>} \boldsymbol{\lambda}^*,$$

it follows that

$$R_{\mathbf{k},\boldsymbol{\lambda}}(t) \leq R_{\mathbf{k},\boldsymbol{\lambda}'}(t) \leq R_{\mathbf{k},\boldsymbol{\lambda}^*}(t).$$

In the special case that $\mathbf{k}$ is equal to $\mathbf{k}^{(1)} \equiv (n, 0, \ldots, 0)$ or $\mathbf{k}^{(2)} \equiv (1, 1, \ldots, 1)$, it is clear from Figs. 1 and 2 that

$$R_{\mathbf{k}^{(1)},\boldsymbol{\lambda}}(t) \leq R_{\mathbf{k}^{(2)},\boldsymbol{\lambda}}(t) = R_{\mathbf{k}^{(1)},\boldsymbol{\lambda}^*}(t).$$

Thus, in this special case, the effect on the reliability function when $\mathbf{k}$ is changed from $(n, 0, \ldots, 0)$ to $(1, 1, \ldots, 1)$ for fixed $\boldsymbol{\lambda} = (3, 2, 8)$ is identical to that when $\boldsymbol{\lambda}$ is changed from $(3, 2, 8)$ to $(5, 2, 6)$ when $\mathbf{k}$ is fixed to be $(n, 0, \ldots, 0)$.

4. Applications in reliability and shock models. The results of §3 can be used to study a variety of applications, as in the case of optimal allocation of components in a system or network as shown in Fig. 1. In particular, Theorem 3.5 provides a solution for the optimal assembly of systems by choosing an optimal design vector $\mathbf{k}$, and Theorem 3.6 illustrates how a configuration of $\boldsymbol{\lambda}$ (for fixed $\lambda_0 + \lambda_1 + \lambda_2$) affects the performance of a system when $\mathbf{k}$ is given.

The optimal assembly of systems, as shown in the earlier work such as [Derman, Lieberman, and Ross (1974)], extensively involves applications of stochastic inequalities. (For references on recent developments in this area, see [Boland, Proschan, and Tong (1993)].) In the present application we note that if the lifelengths of the components are exponentially distributed, then for a fixed number r of type-B components, the system reliability is maximized when the $\mathbf{k}$ vector is such that $|k_i - k_{i'}| \leq 1$ for all $1 \leq i, i' \leq r$. Furthermore, if r increases and the majorization ordering $\mathbf{k} \succ \mathbf{k}'$ holds, then the system becomes more reliable.

A similar application can be found in the shock model described in Example 2.2. In that application the probability function $\overline{F}_{\mathbf{k},\boldsymbol{\lambda}}(t)$ is maximized when the vector $\mathbf{k} = (n - r + 1, 1, \ldots, 1, 0, \ldots, 0)$ for all fixed $\boldsymbol{\lambda}$, r, and t.

REFERENCES

S. D. Al-Saadi and D. H. Young (1982). *A test for independence in a multivariate exponential distribution with equal correlation coefficients*, J. Statist. Comput. Simulation, 14, pp. 219–227.

B. C. Arnold (1968). *Parameter estimation of a multivariate exponential distribution*, J. Amer. Statist. Assoc., 63, pp. 848–852.

A. P. Basu and H. W. Block (1975). *On characterizing univariate and multivariate exponential distributions with applications*, in Statistical Distributions in Scientific Work, Vol 3, G. P. Patil, S. Kotz, and J. K. Ord, eds., D. Reidel, Boston, MA, pp. 399–422.

H. W. Block (1985). *Multivariate exponential distribution*, in Encyclopedia of Statistical Sciences, Vol 6, S. Kotz and N. L. Johnson, eds., John Wiley, New York, pp. 55–59.

H. W. Block and A. P. Basu (1974). *A continuous bivariate exponential extension*, J. Amer. Statist. Assoc., 69, pp. 1031–1037.

P. J. Boland, F. Proschan, and Y. L. Tong (1993). *Some recent applications of stochastic inequalities in system reliability theory*, in Advances in Reliability Theory, A. P. Basu, ed., North-Holland, Amsterdam, pp. 29–41.

C. Derman, G. J. Lieberman, and S. M. Ross (1974). *Assembly of systems having maximum reliability*, Naval Res. Logist. Quart., 21, pp. 1–12.

F. Downton (1970). *Bivariate exponential distributions in reliability theory*, J. Roy. Statist. Soc. Ser. B, 32, pp. 408–417.

J. E. Freund (1961). *A bivariate extension of the exponential distribution*, J. Amer. Statist. Assoc., 56, pp. 971–977.

D. S. Friday (1976). *A new multivariate life distribution*, Ph.D. thesis, Department of Statistics, Pennsylvania State University, University Park, PA.

D. S. Friday and G. P. Patil (1977). *A bivariate exponential model with applications to reliability and computer generation of random variables*, in The Theory and Application of Reliability, Vol I, With Emphasis on Bayesian and Nonparametric Methods; Proceeding of a Conference, University of South Florida, Tampa, FL, December 15–18, 1975, C. Tsokos and I. N. Shimi, eds., Academic Press, New York, pp. 527–549.

N. L. Johnson and S. Kotz (1972). *Distributions in Statistics: Continuous multivariate Distributions*, John Wiley, New York.

A. W. Marshall and I. Olkin (1967a). *A multivariate exponential distribution*, J. Amer. Statist. Assoc., 62, pp. 30–44.

——— (1967b). *A generalized bivariate exponential distribution*, J. Appl. Probab., 4, pp. 291–302.

——— (1979). *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

——— (1985). *Multivariate exponential distributions, Marshall–Olkin*, in Encyclopedia of Statistical Sciences, Vol 6, S. Kotz and N. L. Johnson, eds., John Wiley, New York, pp. 59–62.

F. Proschan and P. Sullo (1974). *Estimating the parameters of a bivariate exponential distribution in several sampling situations*, in Reliability and Biometry, F. Proschan and R. J. Serfling, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 423–440.

——— (1976). *Estimating the parameters of a multivariate exponential distribution*, J. Amer. Statist. Assoc., 71, pp. 465–472.

M. SHAKED AND Y. L. TONG (1985). *Some partial orderings of exchangeable random variables by positive dependence*, J. Multivariate Anal., 17, pp. 333–349.

Y. L. TONG (1989). *Inequalities for a class of positively dependent random variables with a common marginal*, Ann. Statist., 17, pp. 429–435.

——— (1990). *The Multivariate Normal Distribution*, Springer-Verlag, New York and Berlin.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 975–994, July 1994

005

OBSERVABILITY AND OBSERVERS FOR NONLINEAR SYSTEMS*

J. P. GAUTHIER† AND I. A. K. KUPKA‡

Abstract. This paper deals with observability and observers for general nonlinear systems. In the non-control-affine case, we characterize systems that are observable independently of the inputs. An exponential observer for these systems is also exhibited.

Key words. nonlinear systems, observability

AMS subject classifications. 93A99, 93B07, 93B10

1. Introduction.

In this paper, we deal with general, single-output, nonlinear systems:

$$\Sigma \begin{cases} \dot{x} = f_u(x), \\ y = h_u(x), \end{cases}$$

where x belongs to an analytic connected manifold X and the vector fields $f_u(x)$ and the map $h_u(x)$ are analytic with respect to x.

It will be assumed that Σ is smooth with respect to $u \in U$ (an analytic manifold) in a sense that will be made precise in §3.

Our purpose is the following.

(1) We want to characterize those systems Σ that are observable independently of the input ("uniform" observability).

(2) For these systems, we want to construct an exponential observer.

These results have been obtained previously in the case of *control-affine systems*. In this latter case the following facts hold:

(3) There is a local necessary and sufficient condition of "uniform" observability. This local condition leads to a local canonical form for uniform observability.

(4) Assuming that this canonical form is global and assuming some global regularity conditions (some functions have to be globally Lipschitz), an exponential observer is exhibited with arbitrary exponential decay of the error.

Point (3) was dealt with first by Williamson [W] in the case of bilinear systems. It was considered for general control-affine nonlinear systems in [GB], [NI], and [GHO]. Observers for these systems were exhibited in [GHO] and, as stated previously, our aim is to generalize these results to the non-control-affine case, which seems to be much more complicated, as we will see.

Observable nonlinear systems generically have bad inputs [S] (inputs that make them unobservable). However, it seems that in a number of practical cases, systems are observable whatever the input is. These systems are designed to satisfy this property of uniform observability, which is nongeneric.

In other cases where there are sufficiently few bad inputs, our method can be adapted. See [DG] for a very interesting practical case treated in detail.

The observers that we construct have a shortcoming: they are "high-gain" observers (in a sense that will be made clear later). Therefore, they can be very sensitive to noise, though in

* Received by the editors November 6, 1991; accepted for publication (in revised form) February 19, 1993.

† Institut Universitaire de France, Department of Mathematics, Institut National des Sciences Appliquées de Rouen, BP no. 8, Place Emile Blondel, AMS, URA CNRS, D1378, 76131 Mont Saint Aignan Cedex, France. This research was done while the author visited the Department of Mathematics, University of Toronto, and was supported by Natural Sciences and Engineering Research Council of Canada grant OGP 0036498.

‡ Department of Mathematics, University of Toronto, 100, St. Georges Street, Toronto, Ontario, M5S-1A1, Canada.

a number of cases it appears that they are not. On the other hand, it was recently proved (see [D]) that, for these systems, a version of the extended Kalman filter in a special coordinate system converges. The proof of this last fact is more or less a reformulation of our basic proof [GHO].

The type of methodology we use was also applied in [NTT] and [T] for a class of mechanical systems (rigid robots, namely), leading to different results.

Finally, let us point out that our method, although nongeneric (because of the nongenericity of uniform observability), is more general than other classical approaches, such as linearization or bilinearization by output injection. Basic papers on linearization by output injection are [KI] and [KR]; for bilinearization, see [HG1] and [HG2]. Notice, however, that in this last case, observers can also be obtained for systems with bad inputs.

This paper is organized as follows. In §2, we introduce a new concept of observability called infinitesimal observability, which is slightly different from standard observability and more tractable for our purposes. We study these two concepts and prove some of their properties.

In §3, we give the characterization of those systems that are observable independently of the inputs: It turns out that a certain flag of distributions (related to the system considered for every fixed value of the input) has to be independent of the value of this input on the complement of some subanalytic set of codimension 1.

This generalizes the control-affine case, but is much more difficult to prove. In particular, in the control-affine case, standard observability can be dealt with. In the non-control-affine case, the infinitesimal observability assumption is needed.

A canonical form follows from these considerations, generalizing that of the control-affine case, which appears somewhat special. We use this canonical form to construct a "high-gain" observer in §4. Under some mild additional regularity assumptions, the error can be made to decay at an arbitrary exponential rate. This observer generalizes the observer of the control-affine case, but, again, it is more difficult to prove the arbitrary exponential convergence.

2. Infinitesimal observability. We consider the input-output system of the Introduction:

$$\Sigma \begin{cases} \dfrac{dx}{dt} = f_u(x) = f(u,x), \\ y = h_u(x) = h(u,x), \end{cases} \qquad y \in R,$$

and we assume that f and h are analytic in x and jointly continuous in both u and x.

The output depends directly on the input, which is not very relevant in practice. However, the sake of mathematical generalization is not our primary motive: the reason for assuming this dependence is to make the first step of the proof of our main theorems, Theorems 3.0 and 3.1, similar to that of the following steps. Moreover, the canonical form is much more symmetric. Finally, a very special property appears in the two-dimensional case (see Remark 3.3) if the output does not depend on the input.

Let us recall the concept of input-output mapping of the system Σ. Our space of inputs will be the space $L^\infty[U]$ of all measurable and bounded functions $u : [O, T_u[\to U$ defined on a semi-open interval $[O, T_u[$ (depending on u). Usually the inputs are defined on closed intervals $[O, T_u]$, but this is totally unimportant. We take the domains of our inputs to be semi-open, mainly for the sake of symmetry in our definitions.

The space of our output functions will be the space $L[R]$ of all measurable functions $y : [O, T_y[\to R$ defined on the semi-open intervals $[O, T_y[$.

For any input $\hat{u} \in L^\infty[U]$, $\hat{u} : [0, T\hat{u}[\to U$, and any initial state x_0, the maximal solution

(for positive times) $\hat{x}$ of the Cauchy problem

$$\frac{d\hat{x}}{dt}(t) = f(u(t), \hat{x}(t)), \qquad \hat{x}(0) = x_0$$

is defined on a semi-open interval $[0, e(\hat{u}, x_0)[$, where $0 < e(\hat{u}, x_0) \leq T_{\hat{u}}$. That $e(\hat{u}, x_0)$ can be at most equal to $T_{\hat{u}}$ is obvious. If $e(\hat{u}, x_0) < T_{\hat{u}}$, then $e(\hat{u}, x_0)$ is the positive escape time of x_0. It is characterized by the fact that the function $t \to \hat{x}(t) \in X$ has no accumulation point as t tends to $e(\hat{u}, x_0)$; i.e., no sequence

$$\{\hat{x}(t_n) \mid t_n \in [0, e(\hat{u}, x_0)[,\ n \in N,\ t_n \to e(\hat{u}, x_0) \text{ as } n \to +\infty\}$$

has a limit in X.

Let us now recall the main property of $e(\hat{u}, x_0)$.

LEMMA 2.0. *For any input $\hat{u} \in L^\infty[U]$ the function $x_0 \in X \to e(\hat{u}, x_0) \in \overline{R}_+^*$ is lower semicontinuous ($\overline{R}_+^* = \{a \mid 0 < a \leq \infty\}$).*

Now we can define the input-output mapping of Σ.

DEFINITION 2.0. *The input-output mapping P_Σ is defined by the following*:

$$L^\infty[U] \times X \to L[R], \qquad (\hat{u}, x_0) \to P_\Sigma(\hat{u}, x_0),$$

where $P_\Sigma(\hat{u}, x_0)$ is the function $\hat{y} : [0, e(\hat{u}, x_0)[\to L[R]$ defined by $\hat{y}(t) = h(\hat{u}(t), \hat{x}(t))$, $\hat{x}(t)$ as above.

Remark 2.0. For any $\hat{u} \in L^\infty[U]$, $P_{\Sigma,\hat{u}} : X \to L[R]$ will denote the mapping $x_0 \to P_\Sigma(\hat{u}, x_0)$.

DEFINITION 2.1. *A system is called observable if for any triple*

$$(\hat{u}, x_s, x_r) \in L^\infty[U] \times X \times X, \qquad x_s \neq x_r,$$

the set of all $t \in [0, \min(e(\hat{u}, x_s), e(\hat{u}, x_r))[$ such that $P_\Sigma(\hat{u}, x_s)(t) \neq P_\Sigma(\hat{u}, x_r)(t)$ is not of measure zero.

This means that any input separates any two distinct states. But what we need in this study is an apparently weaker condition: we require that for any input $\hat{u}$, the associated input-output mapping does not map any infinitesimal point (i.e., tangent vector) of X into 0, or in other words, that it distinguishes between any infinitesimal point of X and the associated point of X. To make this precise, we need the concept of the lifting of a system Σ to the tangent-space TX (the infinitesimal point space) of X.

Lifting of a system Σ to TX. The mapping $f : U \times X \to TX$ associated with the parametrized vector field f on X induces the tangent mapping $T_X f : U \times TX \to TTX$ (tangent bundle of TX). This mapping defines a parametrized vector field (also denoted by $T_X f$) on TX, since for any $(u, \xi) \in U \times TX$, $T_X f(u, \xi)$ belongs to $T_\xi TX$. Similarly, the function $h : U \times X \to R$ induces a differential $d_X h : U \times TX \to R$. Now we can define the lifting of Σ to TX.

DEFINITION 2.2. *The lifting $T\Sigma$ of Σ to TX is the input-output system*

$$T\Sigma \begin{cases} \dfrac{d\xi}{dt} = T_X f(u, \xi) = T_X f_u(\xi), \\ \eta = d_X h(u, \xi) = d_X h_u(\xi). \end{cases}$$

The trajectories of Σ and $T\Sigma$ are related as follows: let $\pi : TX \to X$ be the canonical projection of the tangent bundle TX. If

$$\xi : [0, t_\xi[\to TX$$

is a trajectory of $T\Sigma$ associated with the input $\hat{u}$, its projection

$$\pi\xi : [0, t_\xi] \to X$$

is a trajectory of Σ associated with the same input $\hat{u}$. But if we know the trajectories of Σ, we can find those of $T\Sigma$. More precisely, let us, for any $x \in X$ and any input $\hat{u} \in L^\infty[U]$, denote by

$$t \in [0, e(\hat{u}, x)[\to \varphi_t(\hat{u}, x) \in X$$

the maximal trajectory of the system Σ corresponding to the input $\hat{u}$ and starting at x (denoted by $\hat{x}$ before). By Lemma 2.0, for any input $\hat{u} \in L[U]$, any $x \in X$, and any time $\tau, 0 \leq \tau < e(\hat{u}, x)$, there exists an open neighborhood $V_{x,\tau}$ of x in X such that $e(\hat{u}, x') > \tau$ for all $x' \in V_{x,\tau}$, and the mapping

$$x' \in V_{x,\tau} \to \varphi_\tau(\hat{u}, x')$$

is a diffeomorphism.

Let $T_X\varphi_\tau : T_{x'}X \to T_Z X$, $z = \varphi_\tau(\hat{u}, x')$, be the induced tangent mapping. For any input $\hat{u} \in L^\infty[U]$, any $\xi_0 \in TX$, $e(\hat{u}, \xi_0) = e(\hat{u}, \pi(\xi_0))$, and the maximal solution $\hat{\xi} : [0, e(\hat{u}, \xi_0)[\to TX$ of $T\Sigma$ corresponding to the input $\hat{u}$ and starting at ξ_0 is given by

$$\hat{\xi}(t) = T_X\varphi_t(\hat{u}, \xi_0), \qquad 0 \leq t < e(\hat{u}, \pi(\xi_0)). \tag{1}$$

Let $x \in X$ and let $t \in [0, e(\hat{u}, x)[$; then $P_\Sigma(\hat{u}, x)(t) = h(\hat{u}(t), \varphi_t(\hat{u}, x))$.
Let $\xi \in TX$ and $t \in [0, e(\hat{u}, \xi)[$; then

$$P_{T\Sigma}(\hat{u}, \xi)(t) = d_X h(\hat{u}(t), T_X\varphi_t(\hat{u}, \xi)).$$

We also introduce the mappings $\Phi_{\Sigma,\hat{u}} : B(\hat{u}) \to X$ and $\Phi_{T\Sigma,\hat{u}} : BT(\hat{u}) \to TX$ defined by

$$\Phi_{\Sigma,\hat{u}}(x, t) = \varphi_t(\hat{u}, x),$$
$$\Phi_{T\Sigma,\hat{u}}(\xi, t) = T_X\varphi_t(\hat{u}, \pi\xi)\xi,$$

where $B(\hat{u})$ and $BT(\hat{u})$ are the subdomains of $X \times R_+$, $TX \times R_+$, respectively, such that

$$B(\hat{u}) = \{(x, t) \mid x \in X,\ t \in R_+,\ 0 \leq t < e(\hat{u}, x)\},$$
$$BT(\hat{u}) = \{(\xi, t) \mid \xi \in TX,\ t \in R_+,\ 0 \leq t < e(\hat{u}, \xi)\}.$$

Then $\Phi_{\Sigma,\hat{u}}$ is differentiable in the X variables and its tangent mapping $T_X\Phi_{\Sigma,\hat{u}} : BT(\hat{u}) \to TX$ is given by

$$T_X\Phi_{\Sigma,\hat{u}} = \Phi_{T\Sigma,\hat{u}}.$$

Remark 2.1. The considerations above show that for any input $\hat{u} \in L^\infty_{[U]}$ and any state $x \in X$, the restriction of $P_{T\Sigma}$ to $\{\hat{u}\} \times T_x X$ defines a linear mapping

$$P_{T\Sigma,\hat{u},x} : T_x X \to L^\infty_c([0, e(\hat{u}, x)[; R)$$
$$\text{given by } P_{T\Sigma,\hat{u},x}(\xi) = P_{T\Sigma}(\hat{u}, \xi)(t),$$

where $L^\infty_c([0, e(\hat{u}, x)[; R)$ is the space of all equivalence classes of measurable functions $g : [0, e(\hat{u}, x)[\to R$ such that for any $T, 0 < T < e(\hat{u}, x)$, $g_{|[0,T]}$ belongs to L^∞.

Now we can introduce the concept of infinitesimal observability.

DEFINITION 2.3. *A system Σ is called infinitesimally observable at $(\hat{u}, x) \in L^\infty_{[U]} \times X$ if the linear mapping*

$$P_{T\Sigma,\hat{u},x} : T_xX \to L^\infty_c([0, e(\hat{u}, x)[; R), \qquad \xi \in T_xX \to P_{T\Sigma}(\hat{u}, \xi)$$

is injective.

DEFINITION 2.4. *A system Σ is called* infinitesimally observable at $\hat{u} \in L^\infty_{[U]}$ *if it is infinitesimally observable at all pairs $(\hat{u}, x)$, $x \in X$. It is called* uniformly infinitesimally observable *if it is infinitesimally observable at all inputs $\hat{u} \in L^\infty_{[U]}$.*

Remark 2.2. One has for any $\xi \in TX$ and almost any $t \in [0, e(\hat{u}, \xi)[$,

$$P_{T\Sigma}(\hat{u}, \xi)(t) = d_x P^t_{\Sigma,\hat{u}}(\xi), \tag{2}$$

where the right-hand side is the differential $TX \to R$ induced by the function $P^t_{\Sigma,\hat{u}} : V \to R$, and V is the open set

$$\{x \in X \mid 0 < t < e(\hat{u}, x)\} \quad \text{and} \quad P^t_{\Sigma,\hat{u}}(x) = P_\Sigma(\hat{u}, x)(t).$$

In view of the relation (2) above, the fact that a system is *infinitesimally observable* at $\hat{u} \in L^\infty[U]$ means that the mapping $P_{\Sigma,\hat{u}} : X \to L_{[R]}$ is an *immersion* of X into $L_{[R]}$. (As was stated above, $P_{\Sigma,\hat{u}}$ is differentiable in the following sense: we know that $e(\hat{u}, x) \geq e(\hat{u}, x_0) - \varepsilon$ in a neighborhood $U\varepsilon$ of x_0. Then $P_{\Sigma,\hat{u}}$ is differentiable in the classical sense from $U\varepsilon$ into $L^\infty([0, e(\hat{u}, x_0) - \varepsilon[; R)$. $P_{\Sigma,\hat{u}}$ is an immersion in the sense that these differential maps are injective.)

Example 2.0. Consider the (uncontrolled) system on R^2:

$$\begin{aligned} \dot{X}_1 &= X_2, \\ \dot{X}_2 &= 0, \\ Y &= X_1^3. \end{aligned}$$

This system is not infinitesimally observable at $x_0 = 0$, but clearly it is observable on R^2.

Next we shall study the relations between observability and infinitesimal observability.

THEOREM 2.0. (i) *For any system Σ and any input $\hat{u}$, the set $\theta(\hat{u})$ of all states $x \in X$ such that Σ is infinitesimally observable at $(\hat{u}, x)$ is open in X (could be empty, of course).*

(ii) *If Σ is observable for an input $\hat{u}$, then $\theta(\hat{u})$ is everywhere dense in X.*

(iii) *If Σ is infinitesimally observable at $(\hat{u}, x)$, then there exists an open neighborhood V of X such that the restriction $P_{\Sigma,\hat{u}|V}$ is injective.*

Proof. Since the output function h of Σ depends on the inputs, the outputs are measurable functions only and hence are not uniquely defined pointwise. To palliate this difficulty, we define "regularized" outputs. For any input $\hat{u} \in L^\infty_{[U]}$ and any $x \in X$, the restriction of $P_\Sigma(\hat{u}, x)$ to any interval $[0, T]$ is bounded for any T, $0 \leq T < e(\hat{u}, x)$, and hence integrable. So we can define the regularized output by:

$$\overline{P_\Sigma}(\hat{u}, x)(t) = \int_0^t P\Sigma(\hat{u}, x)(s)\, ds \quad \text{for all } t \in [0, e(\hat{u}, x)[.$$

$\overline{P_\Sigma}(\hat{u}, x)(t)$ is an absolutely continuous function. Obviously, we can do the same for the output $P_{T\Sigma}$ of $T\Sigma$ and define for $\hat{u} \in L^\infty_{[U]}$ and $\xi \in TX$:

$$\overline{P}_{T\Sigma}(\hat{u}, \xi)(t) = \int_0^t P_{T\Sigma}(\hat{u}, \xi)(s)\, ds.$$

If we introduce the mappings

$$\overline{P}_{\Sigma,\hat{u}} : \Omega(\hat{u}) = \{(x,t) \in X \times R \,|\, 0 \leq t < e(\hat{u},x)\} \to X,$$
$$\overline{P}_{T\Sigma,\hat{u}} : T\Omega(\hat{u}) = \{(\xi,t) \in TX \times R \,|\, 0 \leq t < e(\hat{u},\xi)\} \to TX,$$

then $\overline{P}_{\Sigma,\hat{u}}$, $\overline{P}_{T\Sigma,\hat{u}}$ are both continuous in x and t, analytic in x, and

$$T_X \overline{P}_{\Sigma,\hat{u}} = \overline{P}_{T\Sigma,\hat{u}}.$$

Let us introduce for every input $\hat{u} \in L^\infty_{[U]}$ the following distribution $D(\hat{u})$: $D(\hat{u})_x = \operatorname{Ker} P_{T\Sigma,\hat{u},x}$, $x \in X$ (for the notation, see Definition 2.3). Then Σ is infinitesimally observable at $(\hat{u},x)$ if and only if $D(\hat{u})_x = 0$.

To be able to handle $D(\hat{u})$ conveniently, we shall give a new definition of $D(\hat{u})$. It is obvious, from the definition of $\overline{P}$, that

$$D(\hat{u})_x = \bigcap_{0 \leq t < e(\hat{u},x)} \operatorname{Ker} \omega(x,t) \quad \text{where } \omega(x,t) : T_xX \to R$$

is the linear form $\xi \in T_xX \to \omega(x,t)\,(\xi) = \overline{P}_{T\Sigma}(\hat{u},\xi)\,(t)$.

Since $\overline{P}_{T\Sigma,\hat{u}}$ is continuous in x and t, so is the form ω. If $D(\hat{u})_x = 0$ at some x, then there exists $d = \dim X$ times $0 < t_1 < t_r < \cdots < t_d < e(\hat{u},x)$ such that the forms $\omega(x,t_i)$, $1 \leq i \leq d$, are linearly independent in T_xX.

Using Lemma 2.0 we can find an open neighborhood W_x of x in X such that for any $z \in W_x$, $e(\hat{u},z) > t_d$ and the forms $\omega(z,t_i)$, $1 \leq i \leq d$, are linearly independent on T_zX. Hence $D(\hat{u})_z \subset \cap_{i=1}^d \operatorname{Ker} \omega(z,t_i) = 0$. This proves (i). To prove (ii) we have to show that if Σ is observable for $\hat{u}$, then any open subset Y of X contains a point such that $D(\hat{u})_x = 0$.

Let $c = \sup\{\operatorname{codim} D(\hat{u})_x \,|\, x \in Y\}$. We have to prove that $c = d = \dim X$. But first we shall prove that the set of all $x \in Y$, such that $\operatorname{codim} D(\hat{u})_x = c$, is open.

In fact, for any such $\overline{x} \in Y$, we can find c times $0 < t_1 < \cdots < t_c < e(\hat{u},\overline{x})$ such that the forms $\omega(\overline{x},t_1),\ldots,\omega(\overline{x},t_c)$ are linearly independent. Then, using Lemma 2.0 and the continuity of $\omega(x,t)$ in x, we can find an open neighborhood W of $\overline{x}$ in Y such that for all $x \in W$, $e(\hat{u},x) > t_c$ and the forms $\omega(x,t_1),\ldots,\omega(x,t_c)$ are linearly independent.

Since for any $x \in W$, $D(\hat{u})_x \subset \cap_{i=1}^c \operatorname{Ker} \omega(z,t_i)$, $\operatorname{codim} D(\hat{u})_x \geq c$, but by the definition of c, $\operatorname{codim} D(\hat{u})_x \leq c$. Hence, $\operatorname{codim} D(\hat{u})_x = c$ for all $x \in W$ and $D(\hat{u}) = \cap_{i=1}^c \operatorname{Ker} \omega(.,t_i)$.

For any $\xi \in TW$, we have that

$$\omega(z_i,t_i)\,(\xi) = \int_0^{t_i} d_X h(\hat{u}(s), T_X\varphi_s(\hat{u},\xi))\,ds;$$

hence $\omega(.,t_i) = d_X F_i$, where $F_i : W \to R$ is the function

$$F_i(x) = \int_0^{t_i} h(\hat{u}(s),\, \varphi_s(\hat{u},x))\,ds = \overline{P}_\Sigma(\hat{u},x)(t_i).$$

This shows that $D(\hat{u})_{|W}$ is an integrable distribution whose leaves are the connected components of the level manifold of the function $(F_1,\ldots,F_c)$.

Now we show that $c = d$. Assume that $c < d$ and take any compact connected set K containing more than one point and contained in a leaf L of $D(\hat{u})_{|W}$. This is possible since $\dim L = d - c > 0$. The infimum of the function $x \in K \to e(\hat{u},x)$ is attained at some point $x_0 \in K$.

For any time T, $0 \leq T < e(\hat{u}, x_0)$, there exists an open connected neighborhood V_T of K in W such that $e(\hat{u}, x) > T$ for all $x \in V_T$.

Since $\omega(.,T)_{|V_T} = d_X F_T$ where $F_T : V_T \to R$ is the function

$$F_T(x) = \int_0^T h(\hat{u}(s), \varphi_s(\hat{u}, x))\, ds$$
$$= \overline{P}_\Sigma(\hat{u}, x)(T),$$

and Ker $\omega(.,T) \supset D(\hat{u})_{|W}$, F_T is constant on any connected component of the intersection $L \cap V_T$, in particular, on the one containing K. Hence F_T is constant on K for all T, $0 \leq T < e(\hat{u}, x_0)$.

Since $F_T(x) = \overline{P}_\Sigma(\hat{u}, x)(T)$, we have that $\overline{P}_\Sigma(\hat{u}, x)(t) = \overline{P}_\Sigma(\hat{u}, x_0)(t)$ for all $x \in K$ and all $t \in [0, e(\hat{u}, x_0)[$.

Differentiating with respect to t, we get $P_\Sigma(\hat{u}, x)(t) = P_\Sigma(\hat{u}, x_0)(t)$ for all $x \in K$ and almost all $t \in [0, e(\hat{u}, x_0)[$. This shows that $\hat{u}$ does not distinguish x_0 and x for any $x \in K$—a contradiction. Hence $c = 0$ and this proves (ii).

(iii) is easy to prove and is left to the reader. □

3. Analytic systems that are observable for any input. In this section, we need additional assumptions about the system. We shall assume that one of the following holds.

Assumption H1. U is a compact connected analytic manifold, possibly with a boundary, and f and h are analytic in x and u.

Assumption H2. U is a vector-space R^p and f and h are analytic in x and u and *polynomial* in u.

We now introduce a flag of distributions.

DEFINITION 3.0. *For each integer* n, $0 \leq n < \dim X = d$, *and for any* $u \in U$, *let us denote by* $D_n(u)$ *the analytic distribution on* X *defined as follows*:

$$D_n(u) = \bigcap_{i=0}^{n} \text{Ker } dh_u^i,$$

where $h_u^i : X \to R$ *is the function* $= \theta(f_u)^i(h_u)$ *and* θ *denotes the Lie derivative.*

It is clear that $D_0(u) \supset D_1(u) \supset \cdots \supset D_{d-1}(u)$. Let M denote the projection on X of the subset

$$\tilde{M} = \{(u, x) \mid dh_u^0(x) \wedge dh_u^1 \wedge \cdots \wedge dh_u^{d-1}(x) = 0\}$$

of $U \times X$. We can now state our main theorem.

THEOREM 3.0. *Assume that* Σ *is uniformly infinitesimally observable and that it satisfies Assumption* H1 (*respectively*, H2). *Then we have the following.*

(i) *The set* M *is a subanalytic* (*respectively, semi-analytic in the case of* H2) *set of codimension at least* 1. *In the case* H1, M *is closed. In any case, denote by* $\overline{M}$ *its closure.*

(ii) *On* $X - \overline{M}$, $D_n(u)$ *has constant rank equal to* $d - n - 1$ *for all* $n, 0 \leq n \leq d - 1$.

(iii) *On* $X - \overline{M}$, $D_n(u)$ *is independent of* $u \in U$ *for* $0 \leq n \leq d - 1$.

First let us make some remarks.

Remark 3.0. Since $\tilde{M}$ is an analytic subset of $U \times X$, it is clear that under Assumption H1 (respectively, H2) its projection M is subanalytic (respectively, semi-analytic). What is not so obvious is that its codimension is at least 1.

Remark 3.1. The points $(u, x) \in U \times X \backslash \tilde{M}$ can be characterized as those $x \in X$ having a neighborhood V_0 in X such that the restriction to V_0 of the functions $h_u^0, h_u^1, \ldots, h_u^{d-1}$ form a system of analytic coordinates on V_0.

It is easy to see that Theorem 3.0 is equivalent to the apparently stronger Theorem 3.1, stated now.

THEOREM 3.1. *Let Σ be uniformly infinitesimally observable and let it satisfy either Assumption* H1 *or* H2. *Then*

(i) *M is a subanalytic (respectively, semi-analytic in the case of* H2) *subset of codimension* 1 *in X;*

(ii) *For any $a \in X - \overline{M}$ and any $v \in U$, there exists an open neighborhood V_a of a, $V_a \subset X - \overline{M}$, such that the functions $x^0 = h_v^0|_{V_a}$, $x^1 = h_v^1|_{V_a}, \dots, x^{d-1} = h_v^{d-1}|_{V_a}$, form a coordinate system on V_a, and on $U \times V_a$, each h^i is a function of u, $x^0, \dots, x^i$ only, $0 \le i \le d-1$:*

$$h^i = H^i(u, x^0, \dots, x^i).$$

COROLLARY. *With the assumptions and notations of Theorem* 3.1, *in the coordinates $(x^0, \dots, x^{d-1})$, the system Σ has the following expression:*

$$\text{(C)} \qquad \begin{aligned} \frac{dx^0}{dt} &= F^0(u, x^0, x^1), \\ &\vdots \\ \frac{dx^i}{dt} &= F^i(u, x^0, x^1, \dots, x^{i+1}), \\ &\vdots \\ \frac{dx^{d-1}}{dt} &= F^{d-1}(u, x^0, x^1, \dots, x^{d-1}), \\ y &= H^0(u, x^0), \end{aligned}$$

and the functions

$$\frac{\partial H^0}{\partial x^0}, \frac{\partial F^i}{\partial x^{i+1}}, \qquad 0 \le i \le d-2$$

are nowhere zero on $U \times Va$.

Proof of Theorem 3.1. We shall prove, by induction on n, the following assertion:

(A_n). Let M_n be the projection on X of the semi-analytic (respectively, analytic, partially algebraic) subset $\tilde{M}_n = \{(u, x) \mid dh_u^0(x) \wedge \cdots \wedge dh_u^n(x) = 0\}$.

Then M_n is a subanalytic (respectively, semi-analytic) subset of X of codimension ≥ 1, and for any $a \in X \backslash \overline{M}_n$ and any $v \in U$, there exists an open neighborhood V of a such that the restriction of h^i to $U \times V$ is a function of u and of the restrictions $h_v^0|_V, \dots, h_v^i|_V$ of $h_v^0, \dots, h_v^i$ to V only, for all i, $0 \le i \le n$ (this last property we shall denote by (P_n)).

It is clear that A_{d-1} implies Theorem 3.1, in fact, since the functions $h_v^0, \dots, h_v^{d-1}$ are independent on V, there exists a, perhaps smaller, neighborhood of a such that $h_v^0, \dots, h_v^{d-1}$ form a system of analytic coordinates on V_a. Also, $M_{d-1} = M$ and $M_{d-1} \supset \cdots \supset M_1 \supset M_0$.

Assume we have proved $A_0, \dots, A_{n-1}$, and let us prove A_n. This will be done in four steps. In order to prove Steps 1, 2, and 4, we construct feedback laws contradicting infinitesimal observability. In Step 1, this feedback is a constant control. In Step 2, it is a general feedback depending on ξ. In Step 4, it only depends on x.

Let Z_n be the set of all $x \in X$ such that $dh_u^0(x) \wedge \cdots \wedge dh_u^n(x) = 0$ for all $u \in U$. Since $Z_n = \cap_{u \in U} Z_n(u)$, where $Z_n(u) = \{x \in X \mid dh_u^0(x) \wedge \cdots \wedge dh_u^n(x) = 0\}$ is an analytic subset of X, it follows that Z_n is also analytic ([NA, Corollary 2, p. 100]).

Step 1. We claim that the codimension of Z_n is at least 1.

Were it otherwise, Z_n would contain an open set ω. Then $\omega - \overline{M_{n-1}}$ is also open and nonempty. Since for any $u \in U$, $dh_u^0(x) \wedge \cdots \wedge dh_u^n(x) = 0$ on $\omega - \overline{M_{n-1}}$, any point $a \in \omega - \overline{M_{n-1}}$ has, for any given $v \in U$, an open neighborhood W in $\omega - \overline{M_{n-1}}$ such that h_v^n is a function of $h_v^0, \ldots, h_v^{n-1}$ in W.

If $\xi : [0, e[\to TW$ is any trajectory of $T\Sigma$ corresponding to the constant control $u(t) = v$, and such that

$$\xi(0) \in T_a X, \quad \xi(0) \neq 0, \quad dh_v^0(\xi(0)) = \cdots = dh_v^{n-1}(\xi(0)) = 0,$$

then $dh_v^0(\xi) = \cdots = dh_v^{n-1}(\xi) = 0$. In fact, $\left(\frac{d}{dt}\right)(dh_v^i(\xi)) = d(\theta(f_v)h_v^i)(\xi) = dh_v^{i+1}(\xi)$ for all i, $0 \leq i \leq n-1$.

Since $dh_v^n(\xi)$ is a linear combination of the $dh_v^0(\xi)$, $dh_v^{n-1}(\xi)$ it follows from the uniqueness part of Cauchy's theorem for linear differential equations that $dh_v^0(\xi) = \cdots = dh_v^{n-1}(\xi) = 0$. Thus Σ is not uniformly infinitesimally observable—a contradiction.

Step 2. On $TU \times T(X - \overline{M_{n-1}})$, $d_x h^0 \wedge \cdots \wedge d_x h^n \wedge d_u d_x h^n = 0$.

Here d_u (respectively, d_x) denotes the differential with respect to u variables (respectively, x variables) only. If $x_1, \ldots, x_d$ is some coordinate system on some open set X' of X, and $u^1, \ldots, u^c$ is a system of coordinates on some open set U' of U, then

$$d_X h^i = \sum_{j=1}^{d} \frac{\partial h^i}{\partial x^j}(u, x)\, dx^j, \qquad d_u d_x h^n = \sum_{k=1}^{c}\sum_{j=1}^{d} \frac{\partial^2}{\partial u^k \partial x^j} h^n\, du^k \wedge dx^j.$$

Assume the assertion of Step 2 is not true. Then there exists a pair $(u_0, \xi_0) \in \operatorname{int}(U) \times T(X - \overline{M_{n-1}})$ such that $d_x h_{u_0}^i(\xi_0) = 0$ for $1 \leq i \leq n$, but $d_u d_x h^n(., \xi_0)$ is not identically zero on $T_{u_0}U$. By the implicit function theorem, there exists an open neighborhood N of ξ_0 in $T(X - \overline{M_{n-1}})$ and an analytic mapping $\overline{u} : N \to U$ such that $d_x h_{\overline{u}(\xi)}^n(\xi) = 0$ for all ξ in N and $\overline{u}(\xi_0) = u_0$. Let $a = \pi(\xi_0)$ be the base of ξ_0. Since $a \in X - \overline{M}_{n-1}$, we can apply assertion A_{n-1} to a and u_0. Call V_a the corresponding neighborhood; restrict both N and V_a so that $\pi(N) \subset V_a$, and so that for any $v \in \overline{u}(N)$ the statement of A_{n-1} applies to $h_v^0, \ldots, h_v^{n-1}$, i.e., h^{n-1} is a function of u and $h_v^0, \ldots, h_v^{n-1}$ only in $U \times V_a$.

Let $\hat{\xi} : [0, e[\to N$ be the solution of the feedback system $\frac{d\xi}{dt} = T_x f(\overline{u}(\xi), \xi)$ such that $\xi(0) = \xi_0$. We claim that $d_x h_{\hat{u}(t)}^i(\hat{\xi}(t)) = 0$ for all $t \in [0, e[$ and all $0 \leq i \leq n-1$, where $\hat{u}(t) = \overline{u}(\hat{\xi}(t))$. In fact,

$$1 \leq i \leq n-2 \begin{cases} \dfrac{d}{dt}\left(d_x h_{\hat{u}(t)}^0(\hat{\xi}(t))\right) = d_x h_{\hat{u}(t)}^1(\hat{\xi}(t)) + d_u d_x h^0\left(\dfrac{d\hat{u}(t)}{dt}, \hat{\xi}(t)\right), \\ \dfrac{d}{dt}\left(d_x h_{\hat{u}(t)}^i(\hat{\xi}(t))\right) = d_x h_{\hat{u}(t)}^{i+1}(\hat{\xi}(t)) + d_u d_x h^i\left(\dfrac{d\hat{u}(t)}{dt}\hat{\xi}(t)\right), \\ \dfrac{d}{dt}\left(d_x h_{\hat{u}(t)}^{n-1}(\hat{\xi}(t))\right) = d_u d_x h^{n-1}\left(\dfrac{d\hat{u}(t)}{dt}, \hat{\xi}(t)\right), \end{cases}$$

since by construction,

$$d_x h_{\overline{u}(\hat{\xi}(t))}^n(\hat{\xi}(t)) = d_x h_{\hat{u}(t)}^n(\hat{\xi}(t)) = 0.$$

By the choice of V_a, each h^i is a function of u and$h_{\hat{u}(t)}^0, \ldots, h_{\hat{u}(t)}^i$ only in $U \times V_a$, and this is true for all t in $[0, e[$. Hence

$$d_u d_x h^i\left(\frac{d\hat{u}(t)}{dt}, \hat{\xi}(t)\right)$$

is a linear combination, for all $1 \leq i \leq n$ and all t, of $d_x h^0_{\hat{u}(t)}(\hat{\xi}(t)), \ldots, d_x h^i_{\hat{u}(t)}(\hat{\xi}(t))$.

By Cauchy's uniqueness theorem applied to the linear system above, we get that $d_x h^i_{\hat{u}(t)}(\hat{\xi}(t)) = 0$ for all t. This violates the fact that Σ is uniformly infinitesimally observable.

Step 3. *Proof of* (P_n). Take any point a in $X\backslash(Zn \cup \overline{M_{n-1}})$. There exists a $v \in \mathrm{Int}(U)$ such that (v, a) is not in $\tilde{M}_n$. We know that $d_x h^0 \wedge \cdots \wedge d_x h^n \wedge d_u d_x h^n = 0$ everywhere on $X\backslash\overline{M_{n-1}}$. Now apply A_{n-1} to a and v. Since (v, a) is not in $\tilde{M}_n$, $d_x h^0_v(a) \wedge \cdots \wedge d_x h^n_v(a) \neq 0$. Restricting the neighborhood V_a given by A_{n-1}, we can assume that the set $\{h^0_v, \ldots, h^n_v\}$ can be extended to a coordinate system in V_a.

Applying Lemma 3.0 below to $Y = U$, $Z = V_a$, $F^i = h^i$, we get that h^n is a function of u and $h^0_v, \ldots, h^n_v$ only in $U \times V_a$.

Step 4. *Proof of the fact that* M_n *has codimension* 1. A and V_a are chosen as in Step 3.

For simplicity let us denote the restrictions $h^0_v|V_a, \ldots, h^n_v|V_a$ by $x^0, \ldots, X^n$. Then $h^i = H^i(u, x^0, \ldots, x^i)$ for $0 \leq i \leq n$ in $U \times V_a$. Then,

$$d_x h^0 \wedge \cdots \wedge d_x h^n = \frac{\partial H^0}{\partial x^0}\frac{\partial H^1}{\partial x^1} \cdots \frac{\partial H^n}{\partial x^n}\, dx^0 \wedge \cdots \wedge dx^n,$$

$$d_x h^0 \wedge \cdots \wedge d_x h^{n-1} = \frac{\partial H^0}{\partial x^0}\frac{\partial H^1}{\partial x^1} \cdots \frac{\partial H^{n-1}}{\partial x^{n-1}}\, dx^0 \wedge \cdots \wedge dx^{n-1}.$$

Since

$$V_a \cap \overline{M_{n-1}} = \emptyset, \qquad \frac{\partial H^0}{\partial x^0}, \ldots, \frac{\partial H^{n-1}}{\partial x^{n-1}}$$

are all everywhere nonzero in $U \times V_a$. Since $\tilde{M}_n = \{(u, x) \in U \times X | dh^0_u(x) \wedge \cdots \wedge dh^n_u(x) = 0\}$, we see that

$$\tilde{M}_n \cap (U \times V_a) = \left\{ (u, x) \in U \times V_a \;\middle|\; \frac{\partial H^n}{\partial x^n}(u, x) = 0 \right\}.$$

What remains to be proved is that M_n has empty interior. If not, M_n would contain an open set θ, and $\theta\backslash(Z_n \cup \overline{M_{n-1}})$ would be a nonempty open set. Take a point $a \in \theta\backslash(Z_n \cup \overline{M_{n-1}})$. Apply the considerations just developed to a and restrict the neighborhood V_a we have constructed to $V_a \cap (\theta\backslash(Z_n \cup \overline{M}_{n-1}))$.

Denote by $P : \tilde{M}_n \cap (U \times V_a) \to V_a$ the restriction of the projection $U \times V_a \to V_a$ to $\tilde{M}_n$. Since P is surjective, Sard's theorem and the implicit function theorem show that there is an open subset W of V_a and an analytic mapping $\overline{u} : W \to U$ such that

$$\frac{\partial H^n}{\partial x^n}(\overline{u}(x), x) = 0 \quad \text{for all } x \in W.$$

The same reasoning as before shows that Σ is not uniformly infinitesimally observable: let $\hat{\xi} : [0, e[\to TW$ be any maximal (for positive times) solution of the feedback system

$$\frac{d\hat{\xi}}{dt} = T_x f(\overline{u}(\pi(\hat{\xi})), \hat{\xi}) \quad \text{in } TW \text{ such that } \hat{\xi}(0) \neq 0$$

but

$$d_x h^i_{\overline{u}(x_0)}(\hat{\xi}(0)) = 0 \quad \text{for } 0 \leq i \leq n-1, \quad x_0 = \pi(\hat{\xi}(0)).$$

As before, we have

$$\frac{d}{dt} d_x h^i_{\hat{u}}(\hat{\xi}) = d_x h^{i+1}_{\hat{u}}(\hat{\xi}) + d_u d_x h^i \left(\frac{d\hat{u}}{dt}, \hat{\xi}\right), \qquad 0 \le i \le n-1,$$

where $\hat{u}(t) = \overline{u}(\pi(\hat{\xi}(t)))$, $\hat{u} : [0, e[\to U$.

Since

$$\frac{\partial H^i}{\partial x^i} \not\equiv 0 \quad \text{in } U \times V_a, \quad 0 \le i \le n-1,$$

$$d_u d_x h^i \left(\frac{d\hat{u}}{dt}, \hat{\xi}\right)$$

is a linear combination of $d_x h^0_{\hat{u}}(\hat{\xi}), \dots, d_x h^i_{\hat{u}}(\hat{\xi})$. Also,

$$d_x h^n_{\hat{u}}(\hat{\xi}) = \sum_{j=0}^{n} \frac{\partial H^n}{\partial x^j}(\hat{u}, \pi(\hat{\xi}))\, dx^j(\hat{\xi}).$$

But

$$\frac{\partial H^n}{\partial x^n}(\hat{u}, \pi(\hat{\xi})) = \frac{\partial H^n}{\partial x^n}(\overline{u}(\pi(\hat{\xi})), \hat{\xi}) = 0.$$

So $d_x h^n_{\hat{u}}(\hat{\xi})$ is again a linear combination of $d_x h^0_{\hat{u}}(\hat{\xi}), \dots, d_x h^{n-1}_{\hat{u}}(\hat{\xi})$. Again, we can apply Cauchy's uniqueness theorem and get a contradiction. Thus M_n and hence $\overline{M_n}$ are of codimension 1, since the interior of M_n is empty. □

Proof of the corollary to Theorem 3.1. Since $V_a \subset X \backslash \overline{M}$, $d_x h^0 \wedge \cdots \wedge d_x h^{d-1} \not\equiv 0$ everywhere on $U \times V_a$, this is equivalent to the derivatives

$$\frac{\partial H^0}{\partial x^0}, \dots, \frac{\partial H^{d-1}}{\partial x^{d-1}}$$

being everywhere nonzero in $U \times V_a$.

But we can compute the components $F^0, \dots, F^{d-2}$ by induction using the formula

$$F^i = \frac{1}{\dfrac{\partial H^i}{\partial x^i}} \left[H^{i+1} - \sum_{j=0}^{i+1} \frac{\partial H^i}{\partial x^j} F^j \right], \qquad 0 \le i \le d-2. \qquad \square$$

Example 3.0. We give an example of an analytic case, with $u \in R$, where the condition

$$\frac{\partial H^0}{\partial x^i} \text{ never vanishes}$$

is false:

$$\Sigma \begin{cases} \dot{x} = 1, & x \in R, \\ y = \dfrac{1}{2}\left(x - \dfrac{\sin(2x(1+u^2)^{\frac{1}{2}})}{2(1+u^2)^{\frac{1}{2}}}\right) + x \sin^2 u = h(u,x), \end{cases}$$

$$\frac{\partial h}{\partial x} = \sin^2 u + \sin^2(x(1+u^2)^{\frac{1}{2}}).$$

Therefore,

$$\frac{\partial h}{\partial x} = 0 \quad \text{iff } u = k\pi \quad \text{and} \quad x(1+u^2)^{\frac{1}{2}} = m\pi.$$

For the set M, we get

$$x = \frac{m\pi}{(1+(k\pi)^2)^{\frac{1}{2}}},$$

which provides a dense set on R. Now, Σ is uniformly infinitesimally observable. $T\Sigma$ takes the form

$$T\Sigma \begin{cases} \dot{x} = 1, \\ \dot{\xi} = 0, \end{cases}$$
$$\eta = [\sin^2 u + \sin^2(x(1+u^2)^{\frac{1}{2}})]\xi.$$

For $\xi(0) \neq 0$, $\eta = 0$ implies

$$u = k\pi, \qquad x(t) = \frac{m\pi}{(1+(k\pi)^2)^{\frac{1}{2}}},$$

which does not define a set invariant by the dynamics $\dot{x} = 1$.

Remark 3.2. The "usual" control-affine case is the following:

$$\Sigma \begin{cases} \dot{x} = f(x) + u\,g(x), \\ y = h(x). \end{cases}$$

This case is linear with respect to u; therefore, it satisfies Assumption H2. The canonical form (C) reduces to

$$\begin{aligned} \dot{x}_1 &= \varphi_1(x_1, x_2) + u\psi_1(x_1, x_2), \\ &\vdots \\ \dot{x}_{n-1} &= \varphi_{n-1}(x) + u\psi_{n-1}(x), \\ \dot{x}_n &= \varphi_n(x) + u\psi_n(x), \\ y &= h(x_1). \end{aligned}$$

$\partial h/\partial x_1$ nonzero means that we can replace the coordinate x_1 by $h(x_1)$. $\partial F^i/\partial x^{i+1}$ nonzero implies

$$\frac{\partial \varphi_i}{\partial x_{i+1}} + u\frac{\partial \psi_i}{\partial x_{i+1}}$$

are nonzero and hence

$$\frac{\partial \psi_i}{\partial x_{i+1}} = 0, \qquad \frac{\partial \varphi_i}{\partial x_{i+1}} \neq 0.$$

Now, by taking the appropriate coordinate change, we get the usual canonical form of [GB], [GHO] in this case:

$$\begin{aligned} \dot{x}_1 &= x_2 + u\widetilde{\psi_1}(x_1), \\ \dot{x}_{d-1} &= x_d + u\widetilde{\psi_{d-1}}(x_1, \ldots, x_{d-1}), \\ \dot{x}_d &= \tilde{\varphi}_d(x) + u\widetilde{\psi_d}(x), \\ y &= x_1. \end{aligned}$$

Remark 3.3. In the "usual" general case, engineers assume that the output function h does not depend on u.

$$\Sigma \begin{cases} \dot{x} = f(u,x), \\ y = h(x), \end{cases}$$

or, near a generic point for h,

$$\Sigma' \begin{cases} \dot{x} = f(u,x), \\ y = x_1. \end{cases}$$

In the two-dimensional case, the canonical form (C) reduces to

$$C' \begin{cases} \dot{x}_1 = f_1(u,x_1,x_2), \\ \dot{x}_2 = f_2(u,x_1,x_2), \\ y = x_1, \end{cases}$$

which implies no additional restriction on Σ'.

The condition $\partial f_1/\partial x_2$ nonzero is open (at least locally). Therefore (at least locally), in the two-dimensional case, there is an *open set* of systems that are observable independently of the input. This is rather unexpected (and it is not the case in the control-affine situation).

Now our theorems (3.0, 3.1) have an obvious converse in the following.

THEOREM 3.2 (converse of Theorems 3.0, 3.1). *Assume that, around some point x_0 of X, the equivalent conditions of Theorems* 3.0 *and* 3.1 *are met. Then there is a neighborhood V_{x_0} of x_0 such that $\Sigma_{|V_{x_0}}$ (Σ restricted to V_{x_0}) is* uniformly infinitesimally observable (*and then also observable in the usual sense, independently of the input*).

Proof. Choose a V_{x_0} such that $\Sigma_{|V_{x_0}}$ has the canonical form (C), with

$$\frac{\partial H^0}{\partial x^0}, \qquad \frac{\partial F^i}{\partial x^{i+1}} \neq 0.$$

Using the expression $T\Sigma_1$ of $T\Sigma$ in local coordinates, we get

$$\left\{ \begin{array}{c} \begin{bmatrix} \dot{\xi}_1 \\ \dot{\xi}_2 \\ \vdots \\ \dot{\xi}_{d-1} \\ \dot{\xi}_d \end{bmatrix} = \begin{bmatrix} d_x\varphi_1(u,x_1,x_2) \\ d_x\varphi_2(u,x_1,x_2,x_3) \\ \vdots \\ d_x\varphi_{d-1}(u,x) \\ d_x\varphi_d(u,x) \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_d \end{bmatrix}, \\ \eta = d_x\varphi_0(u,x_1)\xi_1. \end{array} \right\}$$

Since

$$\frac{\partial \varphi_0}{\partial x_1} \neq 0,$$

$\eta(t) = 0$ implies $\xi_1(t) = 0$, $\dot{\xi}_1(t) = 0$, which implies

$$0 = \frac{\partial \varphi_1}{\partial x_2}\,\xi_2(t),$$

and $\xi_2(t)$ is zero because

$$\frac{\partial \varphi_1}{\partial x_2} \neq 0.$$

By induction, $\xi(t) = 0$ and $\Sigma_{|V_{x_0}}$ is uniformly infinitesimally observable. □

LEMMA 3.0. *Let Y, Z be two connected analytic manifolds and let $f^0, \dots, f^n : Y \times Z \to R$ be $n+1$ analytic functions such that*

(i) $d_Z f^0 \wedge \cdots \wedge d_Z f^n \wedge d_Y d_Z f^n = 0$ *on* $Y \times Z$.

(ii) *There exists $y_0 \in Y$ such that : $f^0_{y_0}, \dots, f^n_{y_0} : Z \to R$ can be extended to a global coordinate system on Z.*

(iii) *There exists an analytic function $h_i : Y \times \theta_i \to R$, θ_i open in R^n, $0 \le i \le n-1$, such that*

$$f^i(y,z) = h^i(y, f^0_{y_0}, \dots, f^{n-1}_{y_0}(z)) \quad \textit{for all } (y,z) \in Y \times Z.$$

Then there exists a function $h^n : Y \times \theta_n \to R, \theta_n \subset R^{n+1}$ such that for all $(y,z) \in Y \times Z$,

$$f^n(y,z) = h^n(y, f^0_{y_0}(z), \dots, f^n_{y_0}(z)).$$

Proof. Let $x^0, \dots, x^d : Z \to R$ be a global analytic coordinate system on Z such that $x^i = f^i_{y_0}$, $0 \le i \le n$.

Then for any relatively compact $Z' \subset Z$, there exists an open neighborhood Y_0 of y_0 in Y such that for any $y \in Y_0$, $f^0_y, \dots, f^n_y, x^{n+1}, \dots, x^d : Z' \to R$ form a global coordinate system on Z'. By restricting Y_0, we can assume that it carries a coordinate system $\tilde{y} = y^1, \dots, y^m : Y_0 \to R$ such that $\tilde{y}(Y_0) \subset R^m$ is convex and $\tilde{y}(y_0) = 0$.

Then $d_Z f^0 \wedge \cdots \wedge d_Z f^n \wedge d_Y d_Z f^n = 0$ on $Y \times Z$ implies that

$$d_Z f^0 \wedge \cdots \wedge d_Z f^n \wedge d_Z \left(\frac{\partial f^n}{\partial y^i} \right) = 0 \quad \text{on } Y_0 \times Z \quad \text{for all } 1 \le i \le m.$$

This in turn implies that

$$d_Z \left(\frac{\partial f^n}{\partial y^i} \right) = \sum_{j=0}^{n} A^n_{i,j}\, d_Z f^j_y \quad \text{on } Z' \quad \text{for all } y \in Y_0,$$

where $A^n_{i,j}$ are analytic functions on $Y_0 \times Z'$.

So, there are analytic functions

$$\tilde{h}^n_k : Y_0 \times \tilde{\theta} \to R, \quad \tilde{\theta} \subset R^{n+1} \quad \text{open}, \ 1 \le k \le m,$$

such that for all $(y,z) \in Y_0 \times Z'$ and all k, $1 \le k \le m$,

$$\frac{\partial f^n}{\partial y^k}(y,z) = \tilde{h}^n_k(y, f^0_y(z), \dots, f^{n-1}_y(z),\ f^n_y(z)).$$

Using assumption (iii) there are analytic functions $h^n_k : Y_0 \times \theta \to R$, $1 \le k \le m$, θ open $\subset R^{n+1}$ such that

$$\frac{\partial f^n}{\partial y^k}(y,z) = h^n_k(y, f^0_{y0}(z), \dots, f^{n-1}_{y0}(z),\ f^n_y(z))$$

or

$$\frac{\partial f^n}{\partial y^k}(y,z) = h^n_k(y, x^0(z), \dots, x^{n-1}(z),\ f^n_y(z))$$

for all $(y,z) \in Y_0 \times Z'$ and all k, $1 \le k \le m$.

But for all $t \in [0, 1]$ and all $(y, z) \in Y_0 \times Z'$ we have

$$f^n(ty, z) = \int_0^t \sum_{k=1}^{m} \frac{\partial f^n}{\partial y^k}(sy, z) y^k \, ds + f^n(y_0, z),$$

where ty is the point with coordinates $ty^1(y), \ldots, ty^m(y)$, which yields

$$\text{(E)} \qquad \begin{aligned} f^n(ty, z) \; &= \int_0^t \sum_{k=1}^{m} h_k^n(sy, x^0(z), \ldots, x^{n-1}(z),\ f^n(sy, z)) y^k \, ds \\ &\quad + f^n(y_0, z). \end{aligned}$$

Now there exists a neighborhood $Y_0' \subset Y_0$ of y_0 such that the equation (E) has a unique solution (given $f^n(y^0, z)$) defined on $Y_0' \times Z'$ and for all $t \in [0, 1]$. This solution is an analytic function of t, $x_0, \ldots, x_{n-1}, f^n_{Y_0'}$. Hence, if we take $t = 1$, we obtain

$$f^n(y, z) = H^n(y, x^0(z), \ldots, x^{n-1}(z), f^n_{y_0}) \quad \text{on } Y_0' \times Z'$$

or

$$f^n_y = H^n(y, x^0, \ldots, x^n) \quad \text{on } Y_0' \times Z'.$$

But since $x^0, \ldots, x^d : Z \to R$ is a global coordinate system on Z, we have $f^n = G(y, x^0, \ldots, x^d)$ on $Y \times Z$. But on the open subset $Y_0' \times Z'$,

$$\frac{\partial G}{\partial x^j} = \frac{\partial H^n}{\partial x^j} = 0 \quad \text{if } n + 1 \le j \le d.$$

Hence, $\partial G / \partial x^j = 0$ everywhere and $f^n = G(y, x^0, \ldots, x^n)$. □

4. An observer with arbitrary exponential decay for systems that are uniformly infinitesimally observable. As was stated in the Introduction, the observer will be a generalization of the "high-gain" observer that we used in the control-affine case ([GHO]; see also [D], [DG], and [GHK]).

To prove our theorem, we need a technical lemma, which is based upon the same idea as a result in [DA] for stabilization purposes.

LEMMA 4.0. *Consider the time-dependent real matrices $A(t)$ and $C(t)$:*

$$A(t) = \begin{bmatrix} 0 & \varphi_2(t) & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \varphi_3(t) & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \varphi_d(t) \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix},$$

$$C(t) = [\varphi_1(t), 0, \ldots, 0],$$

with $\varphi_1(t)$ such that $0 < \alpha \le \varphi_1(t) \le \beta < \infty$; $i = 1, \ldots, d$.

Then there is a $\lambda > 0$ and there is a vector $\overline{K} \in R^d$ and a symmetric, positive definite $d \times d$ matrix S, depending on α, β only such that

$$(A(t) - \overline{K}\, C)' S + S(A(t) - \overline{K} C) \le -\lambda Id$$

(in which $(A(t) - \overline{K}C)'$ means the transpose of $(A(t) - \overline{K}C)$).

Proof. The proof is an induction on the dimension. For $d = 1$, we consider the quadratic form $S(x_1, x_2)$,

$$S(x, x) = \frac{x^2}{2}.$$

Then $S(x, (A - \overline{K}C)x) = -k\,\varphi_1(t)(x^2/2)$ with a sufficiently large k does the job.

Step d. We look for $\overline{K} = (k, K)$, $k \in R$, $K \in R^{d-1}$ in

$$\left[A(t) - \begin{pmatrix} k \\ K \end{pmatrix} C(t)\right] x = \begin{bmatrix} -k\varphi_1 & C_1(t) \\ -K\varphi_1 & A_1(t) \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad \begin{matrix} x_1 \in R, \\ x_2 \in R^{d-1}, \end{matrix} \tag{3}$$

with $C_1(t) = (\varphi_2(t), 0, \ldots, 0)$ and

$$A_1 = \begin{bmatrix} 0 & \varphi_3 & 0 \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdots & 0 & \varphi_d \\ 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

First we make the following coordinates change:

$$Z_1 = x_{1'}, \qquad Z_2 = x_2 + \Omega x_1.$$

The matrix in (3) becomes

$$\begin{bmatrix} -k\varphi_1 - C_1\Omega & C_1 \\ (-K - \Omega k)\varphi_1 - (A_1 + \Omega C_1)\Omega & (A_1 + \Omega C_1) \end{bmatrix} = B(t). \tag{4}$$

By the induction hypothesis relative to α, β, there is a λ, an Ω, and a quadratic Lyapunov function $Z_2' S_{n-1} Z_2$ such that

$$-(A_1 + \Omega C_1)\, S_{n-1} - S_{n-1}\,(A_1 + \Omega C_1) \leq -\lambda Id.$$

We will look for S_n of the form

$$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & S_{n-1} \end{bmatrix}.$$

Setting $V(Z, Y) = Z' S_n Y$, we get

$$\begin{aligned} 2V(z, B(t)z) &= (-k\varphi_1 - C_1\Omega)z_1^2 + C_1 z_2 z_1 + 2z_2' S_{n-1}(A_1 + \Omega C_1)z_2 \\ &\quad + 2z_2' S_{n-1}[(-K - \Omega k)\,\varphi_1 - (\Omega C_1 + A_1)\Omega]z_1, \\ 2V(z, B(t)z) &\leq (-k\varphi_1 - C_1\Omega)z_1^2 - \lambda\|z_2\|^2 \\ &\quad + (\varphi_2 + 2\|S_{n-1}\|\|(K + \Omega k)\varphi_1 + (A_1 + \Omega C_1)\Omega\|)|z_1|\|z_2\|. \end{aligned}$$

Ω and k being given, K is chosen such that $(K + \Omega k) = 0$. On the other hand,

$$|z_1|\|z_2\| = |z_1/\varepsilon|\|\varepsilon z_2\| \leq \frac{1}{2}\left(\varepsilon^2\|z_2\|^2 + \frac{1}{\varepsilon^2}|z_1|^2\right).$$

Hence

$$2V(z, B(t)z) \leq \left((-k\varphi_1 - C_1\Omega) + \delta(t)\frac{1}{2\varepsilon^2}\right)|z_1|^2 + \left(-\lambda + \delta(t)\frac{\varepsilon^2}{2}\right)\|z_2\|^2,$$

where $\delta(t) = \varphi_2 + 2\|S_{n-1}\|\|(A_1 + \Omega C_1)\Omega\|$.

Since $\delta(t)$ is bounded from above, ε can be chosen small enough for $(-\lambda + \delta(t)(\varepsilon^2/2)) < -\frac{\lambda}{2}$.

Since φ_1 is bounded from below, k can be chosen large enough for

$$-k\varphi_1 - C_1\Omega + \delta(t)\frac{1}{2\varepsilon^2} < -\frac{\lambda}{2}.$$

Hence, $2V(z, B(t)z) \leq -\frac{\lambda}{2}\|z\|^2$.

Setting

$$\theta = \begin{bmatrix} 1 & 0 \\ \Omega & I \end{bmatrix},$$

we have $z = \theta x$, $2x'\theta' S_n B(t)\theta x \leq -\frac{\lambda}{2}x'\theta'\theta x$. Hence with $\tilde{s}_n = \theta' Sn\theta$, $2x'\tilde{s}_n[A - \binom{k}{K}C]x \leq -\frac{\lambda}{2}x'\theta'\theta x$, we obtain

$$2\,x'\tilde{s}_n\left[A(t) - \binom{k}{K}C(t)\right]x \leq -\frac{\lambda}{2}\gamma\|x\|^2$$

for some $\gamma > 0$, which gives the result. □

Now, to exhibit the observer, we need a number of additional technical assumptions.

We will assume

—A1. $X = R^d$ and the canonical form (C) given by Theorems 3.0 and 3.1 is global, i.e., Σ can be globally written on R^d in the form (C).

—A2. Each of the maps

$$\frac{\partial F^i}{\partial x^{i+1}}, \qquad 0 \leq i \leq d-2,$$

is globally Lipschitz with respect to $\underline{x^i}$ uniformly with respect to u and x^{i+1} (denoting $\underline{x^i} = (x^0, \ldots, x^i)$) in (C).

—A3. There exists α, β real, $0 < \alpha < \beta$, such that

$$\alpha \leq \left|\frac{\partial H^0}{\partial x^0}\right|, \quad \left|\frac{\partial F^i}{\partial x^{i+1}}\right| \leq \beta, \quad 0 \leq i \leq d-2 \quad \text{everywhere.}$$

We can assume, in fact, after making an obvious change of coordinates (if necessary), that

$$\alpha \leq \frac{\partial H^0}{\partial x^0}, \frac{\partial F^i}{\partial x^{i+1}} \leq \beta, \qquad 0 \leq i \leq d-2.$$

These assumptions are very strong. However,

—If Σ is uniformly infinitesimally observable, then it can be put locally under the form (C) (by Theorems 3.0 and 3.1). Assume that it is under the form (C) on some sufficiently regular bounded subset B of R^d (a ball, for instance). Then the functions H^0, F^i can be C^k extended to all of R^d in such a way that these assumptions are met.

—In a number of practical cases, they are satisfied. (See [DG] for a nice case and [GHO] for another one. Potential applications to mechanical systems can be obtained from [T] and [NTT].)

Under these assumptions, with the system being rewritten as

$$(\Sigma) \qquad \left.\begin{aligned} \dot{x}_1 &= \varphi_1(u, x_1, x_2) \\ &\vdots \\ \dot{x}_d &= \varphi_d(u, x) \\ y &= \varphi_0(u, x_1) \end{aligned}\right\} = F(u, x),$$

a candidate observer system is (as in [GHO]) just the "high-gain extended Luenberger observer":

$$(\mathrm{O}) \qquad \dot{\hat{x}} = F(u, \hat{x}) - K_\theta(\varphi_0(u, \hat{x}) - y(t)),$$

in which $K_\theta = \Delta_\theta K$, $\Delta_\theta = \mathrm{diag}(\theta, \theta^2, \ldots, \theta^d)$ for some $\theta > 0$ and K, together with S, come from Lemma 4.0, relative to α, β in Assumption A3 above.

We have the following theorem.

THEOREM 4.0. *Assume A_1, A_2, and A_3. Then* (O) *is an exponential observer for* (Σ) *in the sense that*

$$\|\hat{x}(t) - x(t)\| \leq ke^{-at}\|\hat{x}(0) - x(0)\|, \qquad t \geq 0,$$

for every initial condition $\hat{x}(0)$, $x(0)$. Moreover, θ can be chosen large enough for a to be arbitrary.

Proof. Setting $\varepsilon = \hat{x} - x$, we get

$$(5) \qquad \dot{\varepsilon} = F(u, \hat{x}) - F(u, x) - K_\theta(\varphi_0(u, \hat{x}) - y(t)).$$

Consider a trajectory solution of the equations (Σ), (O): $x(t)$, $\hat{x}(t)$, $\varepsilon(t)$, $u(t)$. We have, for $1 \leq i \leq d$ and at each time t, denoting $(x_1, \ldots, x_i, 0, \ldots, 0)$ by $\underline{x^i}$:

$$\begin{aligned} \varphi_i(u, \hat{x}) - \varphi_i(u, x) &= \varphi_i(u, \underline{\hat{x}^i}, \hat{x}_{i+1}) - \varphi_i(u, \underline{x^i}, x_{i+1}) \\ &= \varphi_i(u(t), \underline{\hat{x}^i}(t), \hat{x}_{i+1}(t)) - \varphi_i(u, \underline{x^i}(t), \hat{x}_{i+1}(t)) \\ &\quad + \varphi_i(u, \underline{x^i}(t), \hat{x}_{i+1}(t)) - \varphi_i(u, \underline{x^i}(t), x_{i+1}(t)) \\ &= \varphi_i(u, \underline{\hat{x}^i}, \hat{x}_{i+1}) - \varphi_i(u, \underline{x^i}, \hat{x}_{i+1}) \\ &\quad + \frac{\partial \varphi_i}{\partial x_{i+1}}(u(t), \underline{x^i}(t), \delta_i(t))\varepsilon_{i+1}(t) \end{aligned}$$

for some $\delta_i(t)$.

We set

$$g_{i+1}(t) = \frac{\partial \varphi_i}{\partial x_{i+1}}(u(t), \underline{x^i}(t), \delta_i(t))$$

to obtain

$$\dot{\varepsilon}_i = \varphi_i(u, \underline{\hat{x}^i}, \hat{x}_{i+1}) - \varphi_i(u, \underline{x^i}, \hat{x}_{i+1}) + g_{i+1}(t)\varepsilon_{i+1}(t) - (K_\theta)_i \quad g_1(t)\varepsilon_1,$$

where

$$g_1(t) = \frac{\partial \varphi_0}{\partial x_1}(u(t), \delta_0(t)).$$

This yields

$$\dot{\varepsilon} = (A(t) - K_\theta C(t))\varepsilon + \overline{F}. \tag{6}$$

With $C(t) = (g_1(t), 0, \ldots, 0)$,

$$A(t) = \begin{bmatrix} 0 & g_2(t) & 0, \ldots, & 0 \\ 0 & 0 & g_3(t) & 0, \ldots, 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & g_d(t) \\ 0 & \cdot & \cdot & 0 \end{bmatrix}$$

and

$$\overline{F} = \begin{pmatrix} \vdots \\ \varphi_i(u, \underline{\hat{x}}^i, \hat{x}_{i+1}) - \varphi_i(u, \underline{x}^i, \hat{x}_{i+1}) \\ \vdots \end{pmatrix}.$$

Because the φ_i are Lipschitz with respect to $\underline{x_i}$, by Assumption A2, the end of the proof is as in [GHO] for the control-affine case: Set

$$x^0 = \Delta_\theta^{-1} x, \quad \hat{x}^0 = \Delta_\theta^{-1}\hat{x}, \quad \varepsilon^0 = \Delta_\theta^{-1}\varepsilon,$$
$$\dot{\varepsilon}^0 = \Delta_\theta^{-1}\dot{\varepsilon} = \Delta_\theta^{-1}(A(t) - K_\theta C(t))\Delta_\theta \varepsilon^0 + \Delta_\theta^{-1}\overline{F}.$$

Otherwise,

$$\Delta_\theta^{-1}\overline{F} = \Delta_\theta^{-1} \begin{pmatrix} \vdots \\ \varphi_i(u, \Delta_\theta \underline{\hat{x}^{i,0}}, \hat{x}_{i+1}) - \varphi_i(u, \Delta_\theta \underline{x^{i,0}}, \hat{x}_{i+1}) \\ \vdots \end{pmatrix}.$$

The main fact is that $\|\Delta_\theta^{-1}\overline{F}\| \le L\|\varepsilon^0\|$, where $L/\sqrt{d}$ is the Lipschitz constant of the φ_i' s with respect to $\underline{x_i}$, as is easily verified. Hence

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}(\varepsilon^{0\prime} S \varepsilon^0) = \varepsilon^{0\prime} S \dot{\varepsilon}^0 &= \varepsilon^{0\prime} S \Delta_\theta^{-1}(A(t) - K_\theta C(t))\Delta_\theta \varepsilon^0 + \varepsilon^{0\prime} S \Delta_\theta^{-1}\overline{F} \\ &\le \theta\, \varepsilon^{0\prime} S(A(t) - KC(t))\varepsilon^0 + L\|S\|\|\varepsilon^0\|^2 \\ &\le \left(-\theta\frac{\lambda}{2} + L\|S\|\right)\|\varepsilon^0\|^2. \end{aligned}$$

Hence, for θ sufficiently large, $\gamma > 0$ being arbitrary, we get

$$\frac{d}{dt}(\varepsilon^{0\prime} S \varepsilon^0) \le -\|S\|\|\varepsilon^0\|^2 \le -\gamma\, \varepsilon^{0\prime} S \varepsilon^0,$$

which yields $\varepsilon^{0\prime} S \varepsilon^0 \le \varepsilon^{0\prime} S \varepsilon^0 e^{-\gamma t}$. □

Acknowledgment. We thank Prof. W. Respondek for several very interesting discussions on the subject. We especially thank him for pointing out an unpublished result by W. Dayawansa that gave the idea for the proof of Lemma 4.0. We thank Prof. Dayawansa for allowing us to communicate this result.

REFERENCES

[BZ] D. Bestle and M. Zeitz, *Canonical form design for nonlinear observers with linearizable error dynamics*, Internat. J. Control, 23 (1981), pp. 419–431.

[D] F. Deza, *Contribution to the Synthesis of Exponential Observers*, Ph.D. Thesis, INSA de Rouen, France, 1991.

[DA] W. Dayawansa, personnal communication.

[DG] F. Deza and J. P. Gauthier, *Exponentially converging observers for distillation columns and internal stability of the dynamic output feedback*, Chem. Engrg. Sci., 47 (1992), pp. 3935–3941.

[GB] J. P. Gauthier and G. Bornard, *Observability for any $u(t)$ of a class of nonlinear systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 922–926.

[GHK] J. P. Gauthier, H. Hammouri, and I. Kupka, *Observers for nonlinear systems*, IEEE CDC Conference, Brighton, England, December 1991, pp. 1483–1489.

[GHO] J. P. Gauthier, H. Hammouri, and S. Othman, *A simple observer for nonlinear systems, application to bioreactors*, IEEE Trans. Automat Control, 37 (1992), pp. 875–880.

[H] H. Hironaka, *Subanalytic sets, number theory*, in Algebraic Geometry and Commutative Algebra (in honor of Y. Akizuki), Kinokuniya, Tokyo, 1973, pp. 453–493.

[HG1] H. Hammouri and J. P. Gauthier, *Bilinearization up to output injection*, Systems Control Lett., 11 (1988), pp. 139–149.

[HG2] ———, *Global time varying linearization up to output injection*, SIAM J. Control, 6 (1992), pp. 1295–1310.

[KI] A. Krener and A. Isidori, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47–52.

[KR] A. Krener and W. Respondek, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197–216.

[L] S. Lojasiewicz, *Triangulation of semianalytic sets*, Ann. Sci. Ecol. Norm. Sup. PISA, 1965, pp. 449–474.

[LU] D. G. Luenberger, *Observers for multivariable systems*, IEEE Trans. Automat. Control, 11 (1966), pp. 190–197.

[NA] R. Narasimhan, *Introduction to the theory of analytic spaces*, Lecture Notes in Mathematics 25, Springer-Verlag, Berlin, 1966.

[NI] H. Nijmeijer, *Observability of a class of nonlinear systems, a geometric approach*, Report, University of Twente, Enschede, the Netherlands, 1982.

[NTT] S. Nicosia, P. Tomei, and A. Tornambe, *A nonlinear observer for elastic robots*, IEEE J. Robotics Automat., RA-4 (1988), pp. 45–52.

[S] H. J. Sussmann, *Single input observability of continuous-time systems*, Math. Systems Theory, 12 (1979), pp. 371–393.

[T] A. Tornambe, *Use of asymptotic observers having high gains in the state and parameter estimation*, 28th IEEE CDC Conference, Tampa, FL, December 1989.

[W] D. Williamson, *Observability of bilinear systems, with applications to biological control*, Automatica, 13 (1977), pp. 243–254.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 995–1007, July 1994

006

DECOMPOSITION AND PARAMETRIZATION OF SEMIDEFINITE SOLUTIONS OF THE CONTINUOUS-TIME ALGEBRAIC RICCATI EQUATION*

HARALD K. WIMMER†

Abstract. Negative-semidefinite solutions of the ARE $\mathcal{R}(X) = A^*X + XA + XBB^*X - C^*C = 0$ are studied. With respect to an appropriate basis the ARE breaks up into a Lyapunov equation $A_0^*X_0 + X_0A_0 = 0$, where A_0 has only purely imaginary eigenvalues, and an indecomposable Riccati equation $\mathcal{R}_r(X_r) = A_r^*X_r + X_rA_r + X_rB_rB_r^*X_r - C_r^*C_r = 0$ such that each solution $X \leq 0$ is of the form $X = \operatorname{diag}(X_0, X_r)$. The focus is on the solutions $\mathcal{S} = \{X \mid X = \operatorname{diag}(0, X_r), \mathcal{R}_r(X_r) = 0, X_r \leq 0\}$. The set $\mathcal{S}$ has as an order-isomorphic image a well-defined set $\mathcal{N}$ of A-invariant subspaces. The characterization of $\mathcal{N}$ involves the stabilizable and the uncontrollable subspace of (A, B, C).

Key words. algebraic Riccati equation, semidefinite solutions, parametrization by invariant subspaces

AMS subject classifications. 15A24, 93C45

1. Introduction.

We consider the algebraic Riccati equation (ARE)

$$\mathcal{R}(X) = A^*X + XA + XBB^*X - C^*C = 0, \tag{1.1}$$

where A, B, C are complex matrices of sizes $n \times n, n \times p, q \times n$, respectively. It is the purpose of this paper to give a complete description of the set

$$\mathcal{T} = \{X \mid \mathcal{R}(X) = 0, X \leq 0\}$$

of negative-semidefinite solutions of (1.1).

The following notation will be used. In the partitions

$$\mathbb{C} = \mathbb{C}_{\leq} \cup \mathbb{C}_{>} = C_{<} \cup \mathbb{C}_{=} \cup \mathbb{C}_{>}, \tag{1.2}$$

the subscripts refer to real parts such that $\mathbb{C}_{\leq} = \{\lambda \mid \lambda \in \mathbb{C}, \operatorname{Re} \lambda \leq 0\}$, etc. Put

$$E_\lambda(A) = \operatorname{Ker}(A - \lambda I)^n.$$

To (1.2) correspond the decompositions $\mathbb{C}^n = E_{\leq}(A) \oplus E_{>}(A)$ and

$$\mathbb{C}^n = E_{<}(A) \oplus E_{=}(A) \oplus E_{>}(A), \tag{1.3}$$

where $E_{\leq}(A) = \oplus\{E_\lambda(A), \lambda \in \mathbb{C}_{\leq}\}$, etc. With our choice of notation we also have in mind its use for the discrete-time algebraic Riccati equation [14] where the subscripts refer to $|\lambda| \leq 1$, etc. Let $\operatorname{Inv} A$ denote the lattice of A-invariant subspaces of $\mathbb{C}^n$. To the triple (A, B, C) we associate the controllable subspace

$$R(A, B) = \operatorname{Im}(B, AB, \ldots, A^{n-1}B)$$

and the unobservable subspace

$$V(A, C) = \operatorname{Ker} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

* Received by the editors April 13, 1992; accepted for publication (in revised form) February 22, 1993. This research was supported by the SCIENCE-Programme of the European Communities (SC1*/0126-C) and by Deutsche Forschungsgemeinschaft (Kn 164/3-1).

† Mathematisches Institut, Universität Würzburg, D-8700 Würzburg, Germany.

It will be convenient to define

$$V_{\leq}(A,C) = V(A,C) \cap E_{\leq}(A),$$

and similarly $V_{=}(A,C), V_{<}(A,C)$, etc. Let

$$\sigma(A - sI, B) = \{\lambda \mid \lambda \in \mathbb{C},\ \operatorname{rank}(A - \lambda I, B) < n\}$$

denote the set of uncontrollable eigenvalues of A, and similarly define

$$\sigma\begin{pmatrix} A - sI \\ C \end{pmatrix} = \left\{\lambda \mid \lambda \in \mathbb{C},\ \operatorname{rank}\begin{pmatrix} A - sI \\ C \end{pmatrix} < n\right\}.$$

Then $V_{=}(A,C) = 0$ if and only if

$$\sigma\begin{pmatrix} A - sI \\ C \end{pmatrix} \cap \mathbb{C}_{=} = \emptyset.$$

Note that $R(A,B) + E_{<}(A) = \mathbb{C}^n$ or equivalently $\sigma(A - sI, B) \cap \mathbb{C}_{\geq} = \emptyset$ means that (A,B) is stabilizable.

We know from [5], [7] that $\mathcal{T} \neq \emptyset$ is equivalent to

$$V(A,C) + R(A,B) + E_{<}(A) = \mathbb{C}^n. \tag{1.4}$$

If we write (1.4) as

$$V_{=}(A,C) + V_{>}(A,C) + R(A,B) + E_{<}(A) = \mathbb{C}^n$$

and put

$$U_r = V_{>}(A,C) + R(A,B) + E_{<}(A), \tag{1.5}$$

then (1.4) holds if and only if there exists a subspace U_0 such that $\mathbb{C}^n = U_0 \oplus U_r$ and $U_0 \subseteq V_{=}(A,C)$.

DEFINITION 1.1. *We call* $\mathbb{C}^n = U_0 \oplus U_r$ *an LR*-decomposition *of* $\mathbb{C}^n$ *if* $U_0 \subseteq V_{=}(A,C)$. *The subspace* U_r *is the* Riccati component *and* U_0 *is a* Lyapunov complement.

The following decomposition theorem is the main tool for our investigation. It will be proved in §3.

THEOREM 1.2. *Let* $\mathbb{C}^n = U_0 \oplus U_r$ *be an LR-decomposition. Assume*

$$U_0 = \left\{\begin{pmatrix} x_0 \\ 0 \end{pmatrix}\right\} \quad \textit{and} \quad U_r = \left\{\begin{pmatrix} 0 \\ x_r \end{pmatrix}, x_r \in \mathbb{C}^{n_r}\right\}. \tag{1.6}$$

Then

$$A = \begin{pmatrix} A_0 & 0 \\ A_{r0} & A_r \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ B_r \end{pmatrix}, \qquad C = (0, C_r), \tag{1.7}$$

and

$$\sigma(A_0) \subseteq \mathbb{C}_{=}, \tag{1.8}$$

and

$$V_{>}(A_r, C_r) + R(A_r, B_r) + E_{<}(A_r) = \mathbb{C}^{n_r}. \tag{1.9}$$

We have $X \in \mathcal{T}$ if and only if

$$X = \operatorname{diag}(X_0, X_r) \tag{1.10}$$

and X_0 and X_r satisfy

$$\mathcal{L}_0(X_0) = A_0^* X_0 + X_0 A_0 = 0, \qquad X_0 \leq 0 \tag{1.11}$$

and

$$\mathcal{R}_r(X_r) = A_r^* X_r + X_r A_r + X_r B_r B_r^* X_r - C_r^* C_r = 0, \qquad X_r \leq 0. \tag{1.12}$$

Since (1.11) has the trivial solution $X_0 = 0$, we can associate to each $X = \operatorname{diag}(X_0, X_r) \in \mathcal{T}$ a solution $\rho X \in \mathcal{T}$ given by $\rho X = \operatorname{diag}(0, X_r)$. The basis-free definition of ρ in the theorem below will be proved in §3.

THEOREM 1.3. *Let $\mathbb{C}^n = U_0 \oplus U_r$ be an LR-decomposition and let Π be the projection on U_r along U_0. Then we have*

$$(I - \Pi)^* X \Pi = 0 \tag{1.13}$$

for $X \in \mathcal{T}$. The map $\rho : \mathcal{T} \to \mathcal{T}$ defined by $\rho X = X\Pi$ is independent of the Lyapunov complement U_0.

Put $\mathcal{S} = \rho\mathcal{T}$. The partially ordered set $\mathcal{S}$—or rather an order-isomorphic image $\mathcal{N}$ of $\mathcal{S}$—will become the main object of our study. Define

$$\mathcal{N} = \{N \mid N \in \operatorname{Inv} A, V_{\leq}(A, C) \subseteq N \subseteq V(A, C), N + R(A, B) + E_{<}(A) = \mathbb{C}^n\}.$$

In §4 we adopt a point of view of [3] and [12] and focus on the kernels of solutions to obtain a parametrization of $\mathcal{S}$. The following result will be proved.

THEOREM 1.4. *The map $\gamma : \mathcal{S} \to \mathcal{N}$ given by $\gamma X = \operatorname{Ker} X$ is a bijection, and γ and γ^{-1} are order preserving.*

The remaining part of the paper contains applications of the preceding theorems. Section 5 deals with the existence of negative-definite solutions and §6 with solutions X where the spectrum of the associated closed-loop matrix $A + BB^*$ lies in a prescribed set Λ such that $\mathbb{C}_{\leq} \subseteq \Lambda$. Most of what has been known about semidefinite solutions of (1.1) can be found in the survey article [9] of Kučera and in Ando's monograph [1]. In addition we refer to [6], [7], and [12].

2. Auxiliary results. Put $A_X = A + BB^* X$. Then (1.1) can be written as

$$\mathcal{R}(X) = A_X^* X + X A_X - X B B^* X - C^* C = 0. \tag{2.1}$$

In many instances it will be of advantage to regard (2.1) as a Lyapunov matrix equation

$$A_X^* X + X A_X = X B B^* X + C^* C$$

with semidefinite right-hand side. The following facts are well known.

LEMMA 2.1. *Suppose X satisfies $A^* X + X A = R^* R$ and $X \leq 0$.*

(1) *If $\sigma(A) \subseteq \mathbb{C}_=$, then $R = 0$.*

(2) *If*

$$\sigma \begin{pmatrix} A - sI \\ R \end{pmatrix} \cap \mathbb{C}_= = \emptyset, \tag{2.2}$$

then $\sigma(A) \subseteq \mathbb{C}_<$.

(3) *If* $A^*X + XA = 0$ *and* $X < 0$, *then* $\sigma(A) \subseteq \mathbb{C}_=$ *and* A *is diagonalizable.*

For matrices A_i, B_i, C_i that will appear in partitions of A, B, C we define Riccati operators

$$\mathcal{R}_i(X) = A_i^*X + XA_i + XB_iB_i^*X - C_i^*C_i$$

and Lyapunov operators

$$\mathcal{L}_i(X) = A_i^*X + XA_i.$$

In what follows we encounter more than once a subspace $N \in \mathrm{Inv}A$ such that $N \subseteq \mathrm{Ker}\, C$. Therefore we fix the following set-up. It is understood that in a given context matrices and vectors shall be partitioned in a conforming manner. Let a basis of $\mathbb{C}^n$ be given such that

$$N = \left\{ \begin{pmatrix} x_1 \\ 0 \end{pmatrix}, x_1 \in \mathbb{C}^{n-n_2} \right\}. \tag{2.3}$$

Then

$$A = \begin{pmatrix} A_1 & A_{12} \\ 0 & A_2 \end{pmatrix} \quad \text{and} \quad C = (0, C_2). \tag{2.4}$$

Assume

$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}. \tag{2.5}$$

If

$$X = \begin{pmatrix} X_1 & X_{12} \\ X_{12}^* & X_2 \end{pmatrix}$$

is a solution of (1.1), then we have

$$A_1^*X_1 + X_1A_1 = -(X_1 \quad X_{12})BB^*(X_1 \quad X_{12})^*. \tag{2.6}$$

In the case where X is of the form $X = \mathrm{diag}(X_1, X_2)$, the block X_2 satisfies $\mathcal{R}_2(X_2) = 0$. The following observations are well known and easy to prove.

LEMMA 2.2. *Let* N *and* A, B, C *be as in* (2.3)–(2.5). *Then*

$$V_\leq(A, C) \subseteq N \tag{2.7}$$

is equivalent to

$$\sigma \begin{pmatrix} A_2 - sI \\ C_2 \end{pmatrix} \cap \mathbb{C}_\leq = \emptyset, \tag{2.8}$$

and

$$N + R(A, B) + E_<(A) = \mathbb{C}^n \tag{2.9}$$

is equivalent to

$$\sigma(A_2 - sI, B_2) \cap \mathbb{C}_\geq = \emptyset, \tag{2.10}$$

i.e., to stabilizability of (A_2, B_2).

The conditions (2.8) and (2.10) have implications for definiteness of solutions.

THEOREM 2.3 (see, e.g., [8]). *There exists a unique solution $W_2 < 0$ of $\mathcal{R}_2(X_2) = 0$ if and only if both* (2.8) *and* (2.10) *are satisfied. The matrix W_2 is the least solution, i.e., $\mathcal{R}_2(X_2) = 0$ implies $W_2 \leq X_2$.*

Apart from the use of Theorem 2.3 we want the proof of the parametrization theorem in §4 to be self-contained. For that purpose we include already at this stage an existence result that appears in much more general form in Theorem 1.4.

COROLLARY 2.4 [5], [7]. *If $V(A,C) + R(A,B) + E_<(A) = \mathbb{C}^n$ holds, then $\mathcal{T} \neq \emptyset$.*

Proof. Take $N = V(A,C)$ in (2.3). Then according to Lemma 2.2, the matrices A_2, B_2, C_2 in (2.4) and (2.5) satisfy (2.8) and (2.10). Hence there exists a solution $X_2 < 0$ of $\mathcal{R}_2(X_2) = 0$, and $X = \operatorname{diag}(0, X_2) \leq 0$ satisfies $\mathcal{R}(X) = 0$. □

Let

$$H = \begin{pmatrix} A & BB^* \\ C^*C & -A^* \end{pmatrix}$$

be the Hamiltonian matrix associated to (1.1). There is a link between $V_=(A,C)$ and the space $E_=(H)$.

LEMMA 2.5. *We have*

$$\dim E_=(H) = 2 \dim V_=(A,C)$$

if and only if

$$E_=(A) \subseteq V(A,C) + R(A,B). \tag{2.11}$$

Proof. Note that (2.11) is equivalent to

$$E_=(A) \subseteq V_=(A,C) + R(A,B). \tag{2.12}$$

Assume

$$V_=(A,C) = \operatorname{Im} \begin{pmatrix} I_{n_1} \\ 0 \end{pmatrix}$$

or equivalently (2.4) with $\sigma(A_1) \subseteq \mathbb{C}_=$ and

$$\sigma \begin{pmatrix} A_2 - sI \\ C_2 \end{pmatrix} \cap \mathbb{C}_= = \emptyset. \tag{2.13}$$

If B is given by (2.5), then (2.12) means $E_=(A_2) \subseteq R(A_2, B_2)$, i.e.,

$$\sigma(A_2 - sI, B_2) \cap \mathbb{C}_= = \emptyset. \tag{2.14}$$

On the other hand (2.4) yields

$$H = \begin{pmatrix} A_1 & A_{12} & * & * \\ 0 & A_2 & * & B_2B_2^* \\ 0 & 0 & -A_1^* & 0 \\ 0 & C_2^*C_2 & -A_{12}^* & -A_2^* \end{pmatrix}.$$

Hence we have $\dim E_=(H) = 2n_1$ if and only if

$$\sigma(H_2) \cap \mathbb{C}_= = \emptyset, \tag{2.15}$$

where

$$H_2 = \begin{pmatrix} A_2 & B_2B_2^* \\ C_2^*C_2 & -A_2^* \end{pmatrix}.$$

It is well known that

$$\sigma(H) \cap \mathbb{C}_= = [\sigma(A - sI, B) \cap \mathbb{C}_=] \cup \left[\sigma \begin{pmatrix} A - sI \\ C \end{pmatrix} \cap \mathbb{C}_=\right].$$

Because of (2.13) the property (2.15) is equivalent to (2.14), which completes the proof. □

3. The decomposition theorem. Recall U_r in (1.5). The space

$$U_1 = V_=(A, C) \cap U_r = V(A, C) \cap R(A, B) \cap E_=(A)$$

is crucial for the proof of Theorem 1.2. We shall see that

$$U_1 \subseteq \text{Ker } X \quad \text{if } X \in \mathcal{T}. \tag{3.1}$$

The following statement is equivalent to (3.1).

LEMMA 3.1. *Assume*

$$U_r = V_>(A, C) + R(A, B) + E_<(A) = \mathbb{C}^n. \tag{3.2}$$

Then we have

$$V_=(A, C) \subseteq \text{Ker } X \tag{3.3}$$

for all $X \in \mathcal{T}$.

Proof. Because of (1.3), condition (3.2) implies $E_=(A) \subseteq R(A, B)$. Hence we have $V_=(A, C) \subseteq R(A, B)$, or equivalently

$$R(A, B)^\perp = V(A^*, B^*) \subseteq V_=(A, C)^\perp. \tag{3.4}$$

Now assume $N = V_=(A, C)$ in (2.3) such that $\sigma(A_1) \subseteq \mathbb{C}_=$. From (2.6) and Lemma 2.1 we obtain

$$B^* \begin{pmatrix} X_1 \\ X_{12}^* \end{pmatrix} = 0, \tag{3.5}$$

which implies

$$A^* \begin{pmatrix} X_1 \\ X_{12}^* \end{pmatrix} + \begin{pmatrix} X_1 \\ X_{12}^* \end{pmatrix} A_1 = 0. \tag{3.6}$$

Hence (3.5) and (3.6) yield

$$B^*(A^*)^i \begin{pmatrix} X_1 \\ X_{12}^* \end{pmatrix} = 0, \quad i = 0, 1, \ldots, n-1,$$

or equivalently

$$\text{Im} \begin{pmatrix} X_1 \\ X_{12}^* \end{pmatrix} \subseteq V(A^*, B^*).$$

From (3.4) and $V_=(A,C)^\perp = \operatorname{Im}\binom{0}{I}$ follows $X_1 = 0$. And because of $X \le 0$ we have $X = \operatorname{diag}(0, X_2)$, which proves (3.3). □

Proof of Theorem 1.2. Let $\mathbb{C}^n = U_0 \oplus U_r$ be a given *LR*-decomposition and assume (1.6). Then it is obvious that (1.7)–(1.9) hold. If

$$X = \begin{pmatrix} X_0 & X_{0r} \\ X_{0r}^* & X_r \end{pmatrix} \in \mathcal{T}$$

is partitioned accordingly, then the block X_r satisfies (1.12). From Lemma 3.1 we obtain

$$V_=(A_r, C_r) \subseteq \operatorname{Ker} X_r. \tag{3.7}$$

Note that

$$V_=(A,C) = U_0 \oplus [V_=(A,C) \cap U_r] = U_0 \oplus U_1.$$

Let U_2 be such that $U_r = U_1 \oplus U_2$. If we choose an appropriate basis and take $U_0 \oplus U_1 \in \operatorname{Inv} A$ into account then we have

$$A = \begin{pmatrix} A_0 & 0 & 0 \\ A_{10} & A_1 & A_{12} \\ 0 & 0 & A_2 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ B_1 \\ B_2 \end{pmatrix}, \qquad C = (0, 0, C_2),$$

where

$$\sigma(A_0) \cup \sigma(A_1) \subseteq \mathbb{C}_= \tag{3.8}$$

and

$$\sigma \begin{pmatrix} A_2 - sI \\ C_2 \end{pmatrix} \cap \mathbb{C}_= = \emptyset. \tag{3.9}$$

From

$$U_1 = \{(0, x_1^T, 0)^T, (x_1^T, 0)^T \in V_=(A_r, C_r)\} \tag{3.10}$$

and (3.7) follows (3.1), i.e., $U_1 \subseteq \operatorname{Ker} X$. Hence

$$X = \begin{pmatrix} X_0 & 0 & X_{02} \\ 0 & 0 & 0 \\ X_{02}^* & 0 & X_2 \end{pmatrix}, \tag{3.11}$$

and (1.1) is equivalent to the following set of three equations: $\mathcal{L}_0(X_0) = A_0^* X_0 + X_0 A_0 = 0$,

$$A_0^* X_{02} + X_{02}(A_2 + B_2 B_2^* X_2) = 0, \tag{3.12}$$

and

$$\mathcal{R}_2(X_2) = A_2^* X_2 + X_2 A_2 + X_2 B_2 B_2^* X_2 - C_2^* C_2 = 0. \tag{3.13}$$

Put $\hat{A}_2 = A_2 + B_2 B_2^* X_2$ such that (3.13) can be written as $\hat{A}_2^* X_2 + X_2 \hat{A}_2 = R_2^* R_2$ with $R_2^* = (X_2 B_2, C_2^*)$. Since (3.9) implies

$$\sigma \begin{pmatrix} \hat{A}_2 - sI \\ R_2 \end{pmatrix} \cap \mathbb{C}_= = \emptyset,$$

Lemma 2.1 yields $\sigma(\hat{A}_2) \cap \mathbb{C}_= = \emptyset$. Because of (3.8), (3.12) has only the trivial solution $X_{02} = 0$. Hence the matrix (3.11) is further reduced to $X = \operatorname{diag}(X_0, 0, X_2)$, which is a decomposition of the form (1.10).

Conversely if $X = \operatorname{diag}(X_0, X_r)$ is such that (1.11) and (1.12) hold, then Lemma 3.1 yields $X_r = \operatorname{diag}(0, X_2)$ and X is a solution of $\mathcal{R}(X) = 0$. □

Proof of Theorem 1.3. Fix U_2 such that $U_r = U_1 \oplus U_2$. Let a basis of $\mathbb{C}^n$ be given such that (1.6), (3.10), and $U_2 = \{(0, 0, x_2^T)^T\}$ hold. Then we have

$$X = \operatorname{diag}(X_0, X_r) \tag{3.14}$$

and

$$X_r = \operatorname{diag}(0, X_2) \tag{3.15}$$

for $X \in \mathcal{T}$. Therefore $\Pi = \operatorname{diag}(0, I)$; hence (3.14) an equivalent to (1.13). Let Ψ be the projection onto U_2 along $V_=(A, C) = U_0 \oplus U_1$. Obviously (3.15) implies $X\Psi = X\Pi$, which shows that $\rho X = X\Pi$ is independent of the choice of the Lyapunov component U_0. □

The map ρ has the properties of a closure operator [2] on the partially ordered set $\mathcal{T}$, namely: (1) $X \leq \rho X$, (2) $X \leq Y \Rightarrow \rho X \leq \rho Y$, (3) $\rho^2 X = \rho X$ for all $X, Y \in \mathcal{T}$.

In the next section we are concerned with the set $\mathcal{S} = \rho\mathcal{T}$. It is obvious that $X \in \mathcal{T}$ is in $\mathcal{S}$ if and only if $V_=(A, C) \subseteq \operatorname{Ker} X$. Similarly, we have $\mathcal{T} \neq \emptyset$ together with $\mathcal{T} = \mathcal{S}$ if and only if

$$V_>(A, C) + R(A, B) + E_<(A) = \mathbb{C}^n. \tag{3.16}$$

LEMMA 3.2. *Assume $\mathcal{T} \neq \emptyset$. Then $\mathcal{T} = \mathcal{S}$ is equivalent to*

$$E_=(A) \subseteq R(A, B) \tag{3.17}$$

and also to

$$V_=(A, C) \subseteq R(A, B). \tag{3.18}$$

Proof. If we intersect both sides of (3.18) with $E_=(A)$ we obtain (3.17) in the form $R(A, B) \cap E_=(A) = E_=(A)$. Assume now (3.18). Then $\mathcal{T} \neq \emptyset$, i.e., $\mathbb{C}^n = V(A, C) + R(A, B) + E_<(A) = V_>(A, C) + [V_=(A, C) + R(A, B)] + [V_<(A, C) + E_<(A)]$ implies (3.16). □

4. Kernels of solutions. Scherer's approach in [12], which is based on the map $\gamma : X \mapsto \operatorname{Ker} X$, can be adapted to our analysis and leads to a description of the solution set $\mathcal{S}$.

LEMMA 4.1. (1) *If X is a solution of* (1.1) *then* Ker X *is A-invariant and satisfies*

$$\operatorname{Ker} X \subseteq V(A, C) \tag{4.1}$$

(2) *For $X \in \mathcal{T}$ we have*

$$V_<(A, C) \subseteq \operatorname{Ker} X. \tag{4.2}$$

(3) *A solution $X \in \mathcal{S}$ satisfies*

$$V_\leq(A, C) \subseteq \operatorname{Ker} X \tag{4.3}$$

and

$$\operatorname{Ker} X + R(A, B) + E_<(A) = \mathbb{C}^n. \tag{4.4}$$

Proof. (1) If $Xy = 0$, then $y^*\mathcal{R}(X)y = -y^*C^*Cy = 0$ yields Ker $X \subseteq$ Ker C. From $Cy = 0$ and $\mathcal{R}(X)y = 0$ we obtain $XAy = 0$. Now (4.1) follows from the fact that $V(A, C)$ is the largest A-invariant subspace in Ker C.

(2) Take $N =$ Ker X in (2.3) such that $X = \text{diag}(0, X_2)$, $X_2 < 0$. Put $Y = X_2^{-1}$ and $\tilde{A}_2 = A_2 - YC_2^*C_2$. Then $\mathcal{R}_2(X_2) = 0$ can be written as

$$Y\tilde{A}_2^* + \tilde{A}_2 Y = -(B_2B_2^* + YC_2^*C_2Y).$$

Obviously $Y < 0$ implies $\sigma(\tilde{A}_2) \cap \mathbb{C}_< = \emptyset$, which is equivalent to

$$\sigma\begin{pmatrix} A_2 - sI \\ C_2 \end{pmatrix} \cap \mathbb{C}_< = \emptyset$$

or to (4.2).

(3) For $X \in \mathcal{S}$ we have $V_=(A, C) \subseteq$ Ker X, which establishes (4.3). If we take $N =$ Ker X as above then we conclude that

$$V_=(A_2, C_2) = 0. \tag{4.5}$$

Now put $\hat{A}_2 = A_2 + B_2B_2^*X_2$. Again write $\mathcal{R}_2(X_2) = 0$ as

$$\hat{A}_2^*X_2 + X_2\hat{A}_2 = X_2B_2B_2^*X_2 + C_2^*C_2.$$

Then $X_2 < 0$ implies $\sigma(\hat{A}_2) \subseteq \mathbb{C}_\leq$. Suppose $\hat{A}_2 y = \lambda y$ and $\lambda \in \mathbb{C}_=$. Then the preceding Lyapunov equation yields $B_2^*X_2y = 0$ and $C_2y = 0$. Hence $A_2y = \lambda y$, and $y \in V_=(A_2, C_2)$. From (4.5) follows $y = 0$. Hence

$$\sigma(\hat{A}_2) \subseteq \mathbb{C}_<, \tag{4.6}$$

which means (A_2, B_2) is stabilizable, i.e., we have $R(A_2, B_2) + E_<(A_2) = \mathbb{C}^{n_2}$, which in turn proves (4.4). □

COROLLARY 4.2. *For $X \in \mathcal{S}$ we have*

$$\text{Ker } X = V_\leq(A, C) \oplus E_>(A_X). \tag{4.7}$$

Furthermore $E_>(A_X) \in$ Inv A and $A = A_X$ on $E_>(A_X)$.

Proof. As before assume $X = \text{diag}(0, X_2)$, $X_2 < 0$. Then

$$A_X = \begin{pmatrix} A_1 & \hat{A}_{12} \\ 0 & \hat{A}_2 \end{pmatrix} \quad \text{and} \quad C = (0, C_2).$$

Hence (4.6) and (4.3) imply

$$\text{Ker } X = \{(x_1^T, 0)^T, x_1 \in E_\leq(A_1) \oplus E_>(A_1)\} = V_\leq(A, C) \oplus E_>(A_X).$$

The remaining statements are easy to verify. □

With (4.1), (4.3), and (4.4) we have the properties which characterize the set $\{\text{Ker } X \mid X \in \mathcal{S}\}$. Recall the definition

$$\mathcal{N} = \{N \mid N \in \text{Inv } A, V_\leq(A, C) \subseteq N \subseteq V(A, C), N + R(A, B) + E_<(A) = \mathbb{C}^n\}.$$

Proof of Theorem 1.4. From Lemma 4.1 it is clear that $X \in \mathcal{S}$ implies $\gamma X =$ Ker $X \in \mathcal{N}$. To show that $\gamma : \mathcal{S} \to \mathcal{N}$ is bijection, we fix a subspace $N \in \mathcal{N}$. We want to show

that there exists a unique $Y \in \mathcal{T}$ with $\operatorname{Ker} Y = N$. Because of $V_=(A,C) \subseteq N$ such a Y is necessarily in $\mathcal{S}$. As usual we work in the set-up (2.3)–(2.5). Note that $\operatorname{Ker} X = N$ together with $X \leq 0$ is equivalent to $X = \operatorname{diag}(0, X_2), X_2 < 0$. Furthermore X is a solution of $\mathcal{R}(X) = 0$ if and only if X_2 satisfies $\mathcal{R}_2(X_2) = 0$. According to Lemma 2.2 the properties (2.7) and (2.9) of N can be expressed by (2.8) and (2.10). Hence Theorem 2.3 yields a unique solution $Y_2 < 0$ of $\mathcal{R}_2(X_2) = 0$. Then $Y = \operatorname{diag}(0, Y_2)$ is the uniquely determined solution in $\mathcal{T}$ with $\operatorname{Ker} Y = N$. It is obvious that γ is order preserving, since $X \leq Y \leq 0$ implies $\operatorname{Ker} X \subseteq \operatorname{Ker} Y$. Now assume that $N, M \in \mathcal{N}$ satisfy $N \subseteq M$. Let N be given as in (2.3). Put $W = \gamma^{-1}(N), Y = \gamma^{-1}(M)$. Then $\operatorname{Ker} W = N \subseteq \operatorname{Ker} Y$ implies $W = \operatorname{diag}(0, W_2), W_2 < 0$, and $Y = \operatorname{diag}(0, Y_2), Y_2 \leq 0$. Both W_2 and Y_2 are solutions of $\mathcal{R}_2(X_2) = 0$. According to Theorem 2.3 the definite matrix W_2 is the least solution of $\mathcal{R}_2(X_2) = 0$. Hence $W_2 \leq Y_2$ and $W \leq Y$, which shows that also γ^{-1} is order preserving. □

Remark 4.3. The following statements are equivalent:

(4.8)
(1) $\mathcal{N} \neq \emptyset$,
(2) $V(A,C) + R(A,B) + E_<(A) = \mathbb{C}^n$,

(4.9)
(3) $V(A,C) \in \mathcal{N}$,
(4) $E_>(A) \subseteq V(A,C) + R(A,B)$,

and

$$\dim E_=(H) = 2 \dim V_=(A,C). \tag{4.10}$$

Proof. It is easy to see that the definition of $\mathcal{N}$ implies the equivalence of (1), (2), and (3). Because of (1.3) we can state (4.8) as

$$E_>(A) + E_=(A) \subseteq V(A,C) + R(A,B) \tag{4.11}$$

or equivalently as a pair of two inclusions, namely (4.9) together with

$$E_=(A) \subseteq V(A,C) + R(A,B). \tag{4.12}$$

According to Lemma 2.5 the conditions (4.12) and (4.10) are equivalent, which implies the equivalence of (2) and (4). □

Since $\mathcal{T} \neq \emptyset$ if and only if $\mathcal{S} \neq \emptyset$ (i.e., $\mathcal{N} \neq \emptyset$) the preceding remark yields the known necessary and sufficient conditions for the existence of a solution $X \leq 0$ of (1.1). Condition (2) is contained in [5], [7] whereas (4) can be found in [10]. The fact $\mathcal{T}$ has a greatest element [5], [7] is another immediate consequence of Theorem 1.4. Since $X \leq \rho X$ and $V = V(A,C) = \sup \mathcal{N}$, we see that $\gamma^{-1}(V)$ is greatest negative-semidefinite solution of (1.1).

In [13], [8], [4], [1] solutions of (1.1) are parametrized under more restrictive hypotheses such as controllability or stabilizability, and the parametrization is based on the subspaces $E_>(A_X)$. The subsequent observation makes a connection to those results. From (4.7) follows that a solution $X \in \mathcal{S}$ is uniquely determined by $E_>(A_X)$. Define

$$\mathcal{G} = \{G \mid G \in \operatorname{Inv} A, G \subseteq V_>(A,C), G + V_\leq(A,C) + R(A,B) + E_<(A) = \mathbb{C}^n\}.$$

Then the map $\mu : \mathcal{S} \to \mathcal{G}$ given by $\mu X = E_>(A_X)$ is a bijection, and μ and μ^{-1} are order preserving. □

5. Negative-definite solutions. From Theorem 1.2 we obtain a condition for the existence of negative-definite solutions of (1.1). The following result is attributed to Richardson and Kwong.

THEOREM 5.1 [11]. *The* ARE (1.1) *has a solution* $W < 0$ *if and only if with respect to an appropriate basis the matrices* A, B, C *take the form*

$$A = \begin{pmatrix} A_0 & 0 \\ 0 & A_r \end{pmatrix}, \qquad B = \begin{pmatrix} 0 \\ B_r \end{pmatrix}, \qquad C = (0, C_r), \tag{5.1}$$

where

$$\sigma(A_0) \subseteq \mathbb{C}_= \quad \text{and } A_0 \text{ is diagonalizable}, \tag{5.2}$$

and

$$\sigma(A_r - sI, B_r) \cap \mathbb{C}_{\geq} = \emptyset, \qquad \sigma \begin{pmatrix} A_r - sI \\ C_r \end{pmatrix} \cap \mathbb{C}_{\leq} = \emptyset. \tag{5.3}$$

Assume (5.1)–(5.3). *Then* $W < 0$ *is a solution of* (1.1) *if and only if*

$$W = \operatorname{diag}(W_0, W_r), \tag{5.4}$$

and $W_0 < 0$, $\mathcal{L}_0(W_0) = 0$, *and* W_r *is the unique negative-definite solution of* $\mathcal{R}_r(X_r) = 0$.

The concept of LR-decomposition yields a more specific existence condition.

THEOREM 5.2. *The* ARE (1.1) *has a negative-definite solution if and only if the following conditions are satisfied*:

(5.5) (1) $\mathbb{C}^n = V_=(A, C) \oplus U_r$,

(5.6) (2) $V_<(A, C) = 0$,

(3) $A_{|V_=(A,C)}$ *is diagonalizable.*

If

$$V_=(A, C) = \left\{ \begin{pmatrix} x_0 \\ 0 \end{pmatrix} \right\} \quad \text{and} \quad U_r = \left\{ \begin{pmatrix} 0 \\ x_r \end{pmatrix} \right\}, \tag{5.7}$$

then the matrices A, B, C *take the form* (5.1) *and have the properties* (5.2) *and* (5.3).

Proof. Suppose (1.1) has a solution $W < 0$. Let $\mathbb{C}^n = U_0 \oplus U_r$ be an LR-decomposition such that $U_0 \subseteq V_=(A, C)$ and $U_r = V_>(A, C) + R(A, B) + E_<(A)$. We know that $U_1 = V_=(A, C) \cap U_r \subseteq \operatorname{Ker} W$. Therefore $V_=(A, C) \cap U_r = 0$, and from $V_=(A, C) + U_r = \mathbb{C}^n$ follows (5.5). From (4.2) we obtain (5.6). Let a basis of $\mathbb{C}^n$ be chosen such that (5.7) holds. Then (5.1) is obvious, and (5.4) follows from Theorem 1.2. Since $W_0 < 0$ satisfies $A_0^* W_0 + W_0 A_0 = 0$, we obtain (5.2) from Lemma 2.1. The matrix $W_r < 0$ is a solution of $\mathcal{R}_r(X_r) = 0$ and because of $V_=(A_r, C_r) = 0$, it is the only negative-definite solution. Hence Theorem 2.3 yields (5.3). The sufficiency part of the theorem is obvious. □

6. Location of $\sigma(A_X)$. In addition to definiteness or semidefiniteness of solutions, the location of $\sigma(A_X)$ is of interest. For a subset $\Lambda \subseteq \mathbb{C}$ put $E_\Lambda(A) = \oplus\{E_\lambda(A), \lambda \in \Lambda\}$ and $V_\Lambda(A) = V(A, C) \cap E_\Lambda(A)$. In the important case where $\Lambda = \mathbb{C}_{\leq}$, the equivalence of (1) and (2) of the subsequent theorem can be found in [4].

THEOREM 6.1. *Let* $\Lambda \subseteq \mathbb{C}$, *be given, and assume* $\mathbb{C}_{\leq} \subseteq \Lambda$. *Then the following statements are equivalent*:

(1) *There exists an* $X \in \mathcal{T}$ *such that*

$$\sigma(A_X) \subseteq \Lambda. \tag{6.1}$$

(2) *Both*

$$V(A,C) + R(A,B) + E_<(A) = \mathbb{C}^n \tag{6.2}$$

and

$$R(A,B) + E_\Lambda(A) = \mathbb{C}^n \tag{6.3}$$

hold.

$$\text{(3) } V_\Lambda(A,C) + R(A,B) + E_<(A) = \mathbb{C}^n. \tag{6.4}$$

Proof. (1) $\Rightarrow$ (2). Because of $\sigma(A - sI, B) = \sigma(A_X - sI, B)$, the inclusion (6.1) implies $\sigma(A - sI, B) \subseteq \Lambda$, which is equivalent to (6.3).

(2) $\Rightarrow$ (3). Put $K = \mathbb{C}\backslash\Lambda$. Because of

$$\begin{aligned} R(A,B) + E_\Lambda(A) &= [R(A,B) \cap E_\Lambda(A)] + [R(A,B) \cap E_K(A)] + E_\Lambda(A) \\ &= [R(A,B) \cap E_K(A)] + E_\Lambda(A), \end{aligned}$$

condition (6.3) is equivalent to $R(A,B) \cap E_K(A) = E_K(A)$, i.e., to $E_K(A) \subseteq R(A,B)$. If we write (6.2) as

$$V_\Lambda(A,C) + [V_K(A,C) + R(A,B)] + E_<(A) = \mathbb{C}^n,$$

then $V_K(A,C) \subseteq E_K(A) \subseteq R(A,B)$ yields (6.4).

(3) $\Rightarrow$ (1). Clearly (6.1) is equivalent to $E_\Lambda(A_X) = \mathbb{C}^n$. Put $M = \Lambda\backslash\mathbb{C}_\leq$ such that $E_\Lambda(A_X) = E_\leq(A_X) \oplus E_M(A_X)$. Then (6.1) takes the equivalent form

$$E_>(A_X) = E_M(A_X). \tag{6.5}$$

Note that $\mathbb{C}_\leq \subseteq \Lambda$ yields $V_\leq(A,C) \subseteq V_\Lambda(A,C)$. Hence (6.4) implies $V_\Lambda(A,C) \in \mathcal{N}$, and there exists a solution $X \in \mathcal{S}$ such that $\operatorname{Ker} X = V_\Lambda(A,C) = V_\leq(A,C) \oplus V_M(A,C)$. From (4.7), i.e., $\operatorname{Ker} X = V_\leq(A,C) \oplus E_>(A_X)$, we obtain $E_>(A_X) = V_M(A,C)$. Therefore we have $\sigma(A_X) \cap \mathbb{C}_> \subseteq M$ or $E_>(A_X) \subseteq E_M(A_X)$, which yields (6.5). □

Acknowledgment. I am indebted to a referee for valuable comments and suggestions, which substantially strengthened this paper.

REFERENCES

[1] T. ANDO, *Matrix Quadratic Equations*, Lecture notes, Hokkaido University, Sapporo, Japan, 1988.

[2] G. BIRKHOFF, *Lattice Theory*, Third ed., Vol 25, American Mathematical Colloquium Publications, Providence, RI, 1967.

[3] F. M. CALLIER AND J. L. WILLEMS, *Criterion for the convergence of the solution of the Riccati differential equation*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1232–1242.

[4] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.

[5] A. H. W. GEERTS, *A necessary and sufficient condition for solvability of the linear-quadratic control problem without stability*, Systems Control Lett., 11 (1988), pp. 47–51.

[6] T. GEERTS, *Another method for determining all positive-semidefinite solutions of the algebraic Riccati equation*, Report 88-WSK-03, Eindhoven University of Technology, Eindhoven, the Netherlands, 1988.

[7] A. H. W. GEERTS AND M. L. J. HAUTUS, *The output-stabilizable subspace and linear optimal control*, Proc. Internat. Symposium MTNS-89, Vol II, Birkhäuser, Boston, 1990, pp. 113–120.

[8] V. KUČERA, *On nonnegative definite solutions to matrix quadratic equations*, Automatica, 8 (1972), pp. 413–423.

[9] ———, *Algebraic Riccati equations: Hermitian and definite solutions*, in The Riccati Equation, S. Bittani et al., eds., Springer-Verlag, Berlin, 1991, pp. 53–88.

[10] A. PASTOR AND V. HERNÁNDEZ, *The algebraic Riccati equation: existence and uniqueness of nonnegative definite solutions*, Internal report GAMA/1/91, Dipartamento de Sistemas Informáticos, Universidad Politécnica de Valencia, Valencia, Spain, 1991.

[11] T. J. RICHARDSON AND R. H. KWONG, *On positive definite solutions to the algebraic Riccati equation*, Systems Control Lett., 7 (1986), pp. 99–104.

[12] C. SCHERER, *The solution set of the algebraic Riccati equation and the algebraic Riccati inequality*, Linear Algebra Appl., 153 (1991), pp. 99–122.

[13] J. C. WILLEMS, *Least-squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.

[14] H. K. WIMMER, *Existence of positive-definite and semidefinite solutions of discrete-time algebraic Riccati equations*, Internat. J. Control, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1008–1020, July 1994

007

A STRONG SEPARATION PRINCIPLE FOR STOCHASTIC CONTROL SYSTEMS DRIVEN BY A HIDDEN MARKOV MODEL*

RAYMOND RISHEL†

Abstract. For a linear quadratic system driven by the output of a hidden Markov model, it is shown that the optimal control is obtained by computing the optimal control as if this output was a known deterministic function, and then substituting the best current estimates of the future values of this output for the known function in this control.

Key words. separation principle, hidden Markov model, non-Gaussian noise, linear quadratic, partially observed, optimal stochastic control

AMS subject classification. 93E20

1. Introduction. Consider a linear quadratic control problem in which it is desired to choose the control u_t for the system

$$dx_t = (Ax_t + Bu_t)\,dt + dz_t \tag{1}$$

with given initial state $x_0 = a$, to minimize the quadratic criteria

$$E\left[\int_0^T (x_t' M x_t + u_t' N u_t)\,dt + x_T' Q x_T\right], \tag{2}$$

where the driving noise z_t has the form

$$z_t = \int_0^t H(y_s)\,ds + W_t, \tag{3}$$

in which y is a Markov process and W is a Wiener process independent of y. That is, the driving noise z is the sum of the nonlinear output

$$\int_0^t H(y_s)\,ds \tag{4}$$

of the Markov process y and the Wiener process W. The state x_t of the system is observed and the control u_t is to be chosen based on the past observations of x to minimize the criteria (2).

This problem is a linear quadratic control problem driven by a type of non-Gaussian noise. It can also be thought of as a partially observed stochastic control problem with observed component x and unobserved component y. It is a system driven by a hidden Markov model in that the Markov process y is hidden from the controller, but x is observed.

If in (3), $H(y_s)$ is replaced by a known deterministic function H_s, standard linear quadratic arguments show that the optimal control has the form

$$u_t = -N^{-1}B'[K_t x_t + J_t]. \tag{5}$$

In (5), K_t is the solution of the matrix Ricatti equation

$$\dot{K}_t = -A'K_t - K_t A + K_t B N^{-1} B' K_t - M \tag{6}$$

* Received by the editors June 10, 1992; accepted for publication (in revised form) March 12, 1993. This research was partially supported by National Science Foundation grant DMS-9105649.

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

with terminal condition

$$K_T = Q, \tag{7}$$

and J_t is given by

$$J_t = \int_t^T \phi(t,s) K_s H_s \, ds \tag{8}$$

where $\phi(t,s)$ is the solution of the matrix differential equation

$$\frac{d}{dt}\phi(t,s) = (-A' + K_t B N^{-1} B')\phi(t,s) \tag{9}$$

with boundary condition

$$\phi(s,s) = I \tag{10}$$

where I is the identity matrix.

We wish to show that a strong separation principle holds in that for the original stochastic problem the same control (5) is optimal, if in (8), H_s is replaced by the best current estimate of the future values of $H(y_s)$ given the observations of the process x up to the current time t. That is, the optimal control for the stochastic problem is obtained by replacing H_s in (8) by the conditional expectation

$$E[H(y_s)|x_r;\ 0 \le r \le t]. \tag{11}$$

Special cases of the problem stated in (1)–(3) have been discussed by Helmes and Rishel in [3] and [4] and by Beneš in [1]. In these papers there is no quadratic term involving the state in the performance criteria and in [3] and [4] the Markov process y_t is either a jump Markov process or a diffusion process. When there is no quadratic term involving the state in the performance criteria, the optimal control is linear in a quantity called the *predicted miss*. In the present case, with a quadratic state term in the performance, this is no longer true. A type of separation was mentioned in [3], but the current type of strong separation principle was not pointed out in [1], [3], or [4].

Separation principles for discrete time linear quadratic non-Gaussian systems are well known. Root in [9] shows that the separation principle holds for partially observed linear quadratic discrete time systems driven by independently and identically distributed (i.i.d.) zero mean non-Gaussian random variables. However, the discrete time version of the current problem would be more general than that of [9].

2. Preliminary considerations. Let us begin our discussion by defining the quantities in the problem and the class of controls more precisely. In (1) x_t is the n-dimensional state; u_t is the m-dimensional control; and A, B, M, N, Q are, respectively, $(n \times n)$-, $(n \times m)$-, $(n \times n)$-, $(m \times m)$-, and $(n \times n)$-dimensional matrices. The matrix N is positive definite symmetric and M and Q are nonnegative definite symmetric matrices.

The Markov process y has values in a measurable space $\mathcal{S}$ with σ-field Σ. It will be assumed to have stationary transition probabilities

$$P(t,y,E) \tag{12}$$

and an initial probability distribution $P_0(E)$. Let B denote the space of bounded measurable functions f on $\mathcal{S}$. The transition probability (12) defines a semigroup P_t of operators on B by

$$P_t[f](y) \triangleq \int_{\mathcal{S}} P(t,y,dz) f(z). \tag{13}$$

The function $H(y)$ is a bounded m-dimensional vector-valued measurable function defined on $\mathcal{S}$.

Let

$$\mathcal{F}_t^x \triangleq \sigma[x_r;\ 0 \le r \le t] \tag{14}$$

and

$$\mathcal{F}_t^z \triangleq \sigma[z_r;\ 0 \le r \le t] \tag{15}$$

be, respectively, the σ-fields generated by the past of x and the past of z up to time t. Let β denote the Borel field on $[0, T]$.

The class of admissible controls must satisfy the following two properties:

(i) For each control u there is a corresponding strong solution of (1). By a strong solution of (1) we mean a solution that is a nonanticipative functional on the driving noise z, that is, a solution x, so that x_t can be represented as

$$x_t = F(t, (z_r;\ 0 \le r \le t)), \tag{16}$$

where F is a $(B \times \mathcal{F}_T^z)$-measurable function so that for each t, $F(t, (z_t;\ 0 \le r \le t))$ is $\mathcal{F}_t^z$-measurable.

(ii) Each control u is a measurable stochastic process and for each t, u_t is $\mathcal{F}_t^x$-measurable, where x is the solution of (1) corresponding to u. That is, the control u_t used at time t depends only on measurements of its corresponding process x made up to time t.

Remark. Our assumptions on the class of admissible controls imply that for each t

$$\mathcal{F}_t^x = \mathcal{F}_t^z. \tag{17}$$

To see this, notice that property (i) implies

$$\mathcal{F}_t^x \subset \mathcal{F}_t^z. \tag{18}$$

Equation (1) implies

$$x_t - a - \int_0^t (Ax_s + Bu_s)\, ds = \int_0^t H(y_s)\, ds + W_t = z_t. \tag{19}$$

This and property (ii) imply

$$\mathcal{F}_t^z \subset \mathcal{F}_t^x. \tag{20}$$

Thus the two inequalities (18) and (20) imply (17).

3. Estimation and extrapolation. First we will need some results on nonlinear filtering and nonlinear extrapolation. These follow rather directly from results of Kunita [5] and are summarized in Lemma 1.

LEMMA 1. *The conditional probability distribution π_t of y_t given $\mathcal{F}_t^x$ exists as a probability measure-valued random process such that for each f in B,*

$$\int_{\mathcal{S}} f(y)\pi_t(dy) = E[f(y_t)|\mathcal{F}_t^x]. \tag{21}$$

It is independent of the control used. If we use the notation

$$\pi_t(f) = \int_{\mathcal{S}} f(y)\pi_t(dy), \tag{22}$$

then π_t is the unique solution of the equations

$$\pi_t(f) = P_0(P_t f) + \int_0^t [\pi_r((P_{t-r}f)H') - \pi_r(P_{t-r}f)\pi_r(H')]\, d\nu_r, \tag{23}$$

which hold for each $f \in B$. In (23), *ν is the innovations Wiener process defined by*

$$\nu_t \triangleq \int_0^t [H(y_r) - \pi_r(H)]\, dr + W_t. \tag{24}$$

For $s > t$ the extrapolated value

$$E[H(y_s)|\mathcal{F}_t^x] \tag{25}$$

of $H(y(s))$, given the past observations of x up to time t, can be expressed in terms of π_t by either

$$E[H(y_s)|\mathcal{F}_t^x] = \pi_t(P_{s-t}H) \tag{26}$$

or

$$E[H(y_s)|\mathcal{F}_t^x] = P_0(P_s(H)) + \int_0^t [\pi_r((P_{s-r}H)H') - \pi_r(P_{s-r}H)\pi_r(H')]\, d\nu_r. \tag{27}$$

Proof of Lemma 1. Since the σ-field equality (17) holds,

$$E[f(y_t)|\mathcal{F}_t^x] = E[f(y_t)|\mathcal{F}_t^z]. \tag{28}$$

Thus, since both y and z do not depend on the control, (28) is independent of the control. It now follows from Kunita [5][1] that a unique probability measure-valued process π_t exists, so

$$\int_{\mathcal{S}} f(y)\pi_t(dy) = E[f(y_t)|\mathcal{F}_t^z], \tag{29}$$

π_t is the unique solution of (23) and (24), and (24) is a Wiener process.

To obtain the formula (26) for the extrapolated value (25) of $H(y_s)$, notice that

$$\mathcal{F}_t^{yW} \supset \mathcal{F}_t^z; \tag{30}$$

thus the law of iterated conditional expectations gives

$$E[H(y_s)|\mathcal{F}_t^z] = E[E[H(y_s)|\mathcal{F}_t^{yW}]|\mathcal{F}_t^z]. \tag{31}$$

Since y and W are mutually independent,

$$E[H(y_s)|\mathcal{F}_t^{yW}] = E[H(y_s)|\mathcal{F}_t^y] = P_{s-t}(H). \tag{32}$$

Thus (26) follows from (31) and (32). Formula (27) now follows from (26) and (23) by substituting $P_{s-t}(H)$ for f in (23) and using the semigroup property of P_t.

[1] Actually the results of Kunita [5] are stated with $\mathcal{S}$ a compact space, and f and H continuous functions, but these results of [5] hold under the current assumptions.

4. The linear quadratic optimality argument. Let us show that if J_t^π is defined by

$$J_t^\pi \triangleq \int_t^T \phi(t,s) K_s \pi_t(P_{s-t}H)\, ds, \tag{33}$$

then

$$u_t = -N^{-1}B'[K_t x_t + J_t^\pi] \tag{34}$$

is an optimal control. We show in Appendix I that (34) satisfies the conditions (i) and (ii) for admissibility.

We shall need to take the stochastic differential of J_t^π. In Appendix II we show that

$$dJ_t^\pi = -[(A - BN^{-1}B'K_t)'J_t^\pi + K_t\pi_t(H)]\, dt + \Gamma_t\, d\nu_t, \tag{35}$$

where Γ_t is defined by

$$\Gamma_t \triangleq \int_t^T \phi(t,s) K_s[\pi_t((P_{s-t}H)H') - \pi_t(P_{s-t}H)\pi_t(H')]\, ds. \tag{36}$$

THEOREM 1. *If x is the solution of* (1) *with initial state a corresponding to any admissible control u, then*

$$E\left[\int_0^T (x_t'Mx_t + u_t'Nu_t)\, dt + x_T'Qx_T\right] \geq a'K_0a + J_0^{\pi'}a + R_0 \tag{37}$$

where

$$R_t = \int_t^T [2J_s^{\pi'}\pi_s(H) - J_s^{\pi'}BN^{-1}B'J_s^\pi + \text{ trace } K_s + \text{ trace } \Gamma_s]\, ds. \tag{38}$$

If x is the solution of (1) *with initial state a corresponding to the control given by* (34) *and* (33), *then equality holds in* (37). *Thus* (34) *and* (33) *define the optimal control.*

Proof of Theorem 1. Let u be any admissible control and x be the corresponding solution of (1) with initial state a. Apply Ito's stochastic differential rule and (35) to obtain

$$\begin{aligned} d[x_t'K_tx_t + 2J_t^{\pi'}x_t] &= x_t'\dot{K}_tx_t\, dt + 2(x_t'K_t + J_t^{\pi'})\, dx_t \\ &\quad + 2x_t'dJ_t^\pi + (\text{trace } K_t + \text{ trace } \Gamma_t)\, dt. \end{aligned} \tag{39}$$

From (1) and (24),

$$dx_t = (Ax_t + Bu_t + \pi_t(H))\, dt + d\nu_t. \tag{40}$$

Adding $(x_tMx_t + u_tNu_t)\, dt$ to both sides of (39) and using (40) gives

$$\begin{aligned} &d[x_t'K_tx_t + 2J_t^{\pi'}x_t] + (x_t'Mx_t + u_t'Nu_t)\, dt \\ &\quad = x_t'(\dot{K}_t + M)x_t + 2(x_t'K_t + J_t^{\pi'})[(Ax_t + \pi_t(H))\, dt + d\nu_t] \\ &\qquad + 2(x_t'K_t + J_t^{\pi'})Bu_t\, dt + u_t'Nu_t\, dt \\ &\qquad + 2x_t'dJ_t^\pi + (\text{trace } K_t + \text{ trace } \Gamma(t))\, dt. \end{aligned} \tag{41}$$

Now the minimum over u_t of

$$2(x_t'K_t + J_t^{\pi'})Bu_t + u_t'Nu_t \tag{42}$$

is attained by $u_t = -N^{-1}B'(K_t x_t + J_t^\pi)$ and is given by

$$
\begin{aligned}
&-(x_t'K_t + J_t^{\pi'})BN^{-1}B'(K_t x + J_t^\pi) \\
&\quad = -(x_t'K_tBN^{-1}B'K_t x_t + 2x_t'K_tBN^{-1}B'J_t^\pi + J_t^{\pi'}BN^{-1}B'J_t^\pi).
\end{aligned} \tag{43}
$$

Thus, for any control u,

$$
\begin{aligned}
&d[x_t'K_t x_t + 2J_t^{\pi'}x_t] + (x_t'Mx_t + u_t'Nu_t)\,dt \\
&\quad \geq x_t'(\dot{K}_t + K_tA + A'K_t + M - K_tBN^{-1}B'K_t)x_t \\
&\qquad + 2x_t'[(A'J_t^\pi + K_t\pi_t(H))\,dt - K_tBN^{-1}B'J_t^\pi\,dt + dJ_t^\pi + K_t\,d\nu_t] \\
&\qquad + [2J_t^{\pi'}\pi_t(H) - J_t^{\pi'}BN^{-1}B'J_t^\pi + \text{ trace } K_t + \text{ trace } \Gamma(t)]\,dt + J_t^{\pi'}\,d\nu_t.
\end{aligned} \tag{44}
$$

Now (44), (6), (35), and (38) imply

$$
d[x_t'K_t x_t + 2J_t^{\pi'}x_t] + (x_t'Mx_t + u_t'Nu_t)\,dt \geq 2x_t'(\Gamma_t + K_t)\,d\nu_t + J_t^{\pi'}\,d\nu_t - dR_t. \tag{45}
$$

Integrating both sides of (45) from 0 to T and taking expected values using the fact that $K_T = Q$, $J_T^\pi = 0$, $R_T = 0$, and the expected value of the stochastic integral is zero, gives

$$
E\left[x_T'Qx_T - (a'K_0a + J_0^{\pi'}a) + \int_0^T (x_t'Mx_t + u_t'Nu_t)\,dt\right] \geq R_0. \tag{46}
$$

Thus, on rearranging (90) and using the fact that a and J_0^π are not random,

$$
E\left[x_T'Qx_T + \int_0^T (x_t'Mx_t + u_t'Nu_t)\,dt\right] \geq a'K_0a + J_0^{\pi'}a + R_0 \tag{47}
$$

holds for any control.

If we repeat the argument above for the control (34), then equality will hold in (44). Thus the same steps will show that (47) holds with equality for this control. Thus (34) is the optimal control.

5. An application. Let us illustrate a typical type of application of Theorem 1. Consider the following model for a production planning problem. Suppose that a factory produces goods and will operate most efficiently at a given level of production. Suppose it also wishes to maintain a given level of inventory on hand. The company from time to time replaces their product with a new model and their sales are also dependent on economic conditions. Because of this, the manager chooses a model for his total sales z_t which satisfies

$$
dz_t = y_t\,dt + \sigma dW_t, \tag{48}
$$

in which y_t is a finite state jump Markov process, and W_t is a Wiener process. The term σdW_t represents the short time variations of sales about their "mean rate" y_t. The mean rate y_t depends on the customers' reactions to new models and the economic conditions. Because these change at random times, the manager feels the choice of a jump Markov process, which takes on different constants over different random intervals, is an appropriate model for this mean rate.

Let x_t denote the deviation of the inventory from its desired level, and u_t the deviation of the production rate from its desired level. The inventory equation is

$$
dx_t = u_t\,dt - dz_t. \tag{49}
$$

Consider the problem in which the manager wishes to penalize deviation from his desired inventory and production levels by choosing u_t to minimize

$$E\left[\int_0^T x_s^2 + \lambda u_s^2 \, ds\right], \tag{50}$$

subject to (49) holding. In (50), T is a fixed final time.

When y_t is not random but a known function of time, the optimal control is

$$u_t = \lambda^{-1}(K_t x_t + J_t), \tag{51}$$

where K_t is the solution of

$$\dot{K}_t = \lambda^{-1} K_t^2 - 1; \qquad K_T = 0, \tag{52}$$

and

$$J_t = \int_t^T e^{-\int_t^s \lambda^{-1} K_u \, du} K_s y_s \, ds. \tag{53}$$

Theorem 1 implies, when y_t is modeled as a jump Markov process, that the optimum control is given by (51) where y_s in (53) is replaced by

$$\hat{y}_{ts} = \pi_t P_{s-t}(y). \tag{54}$$

In this case equation (23), defining π_t, is equivalent to the Wonham filter [11] for the vector of conditional probabilities of the states of the jump Markov process, and P_{s-t} is the transition probability matrix of the jump Markov process.

6. Conclusions. For a linear quadratic system driven by the known function of time plus a Wiener process, the optimal control is a linear feedback plus a functional on the future values of this known function. For a linear quadratic system driven by an unknown Markov process plus a Wiener process it was shown that the same linear feedback plus the same functional of the extrapolated future values of the Markov process gives the optimal control. Nonlinear filtering formulas give these extrapolated future values of the Markov process in terms of the conditional distribution of the state of the Markov process given the past measurements and the transition probabilities of the Markov process.

This control problem could be considered as a control problem with state given by (x_t, π_t), where π_t is the conditional probability distribution of the unobserved Markov process. Control problems with a probability distribution as their state were discussed early in the history of stochastic control by Kusher [6] and Mortensen [8]. More recently, Lions [7] has discussed optimal control of Zakai's equation. The current problem gives an example of a control problem whose state involves a probability distribution for which the control law is explicitly computed.

Appendix I. We shall show in this appendix that the control (34) satisfies the conditions (i) and (ii) for admissibility. Let us investigate (i), that is, show that there is a strong solution of (1) corresponding to (34).

To define a solution of (1) corresponding to the control (34) we must have simultaneous solutions of the system of equations

$$dx_t = [Ax_t - BN^{-1}B'(K_t x_t + J_t^{\pi})]\, dt + dz_t, \tag{55}$$

$$J_t^\pi = \int_t^T \phi(t,s) K_s \pi_t(P_{s-t}H)\, ds, \tag{56}$$

$$\pi_t(f) = P_0(P_t f) + \int_0^t [\pi_r((P_{t-r}f)H') - \pi(P_{t-r}f)\pi_r(H')]\, d\nu_r, \tag{57}$$

$$\nu_t = z_t - \int_0^t \pi_r(H)\, dr, \tag{58}$$

where K_t is the solution of (6) and (7). It follows from Kunita [5, Thms. 2.1, 2.2], that there is a unique strong solution of (57) and (58), i.e., a solution π_t, which can be expressed by a $\beta \times \mathcal{F}_t^z$ measure-valued measurable function ψ which is $\mathcal{F}_t^z$-measurable for each t as

$$\pi_t = \psi(t, z_r;\ 0 \le r \le t). \tag{59}$$

Let $\chi(s,t)$ denote the transition matrix of the linear system

$$dx_t = (A - BN^{-1}B'K_t)x_t\, dt. \tag{60}$$

Then the solution of (55) can be expressed as

$$\begin{aligned} x_t &= \chi(0,t)a \\ &\quad + \int_0^t \chi(s,t)BN^{-1}B' \int_s^T \phi(s,v)K_v \psi(v, z_r;\ 0 \le r \le v) P_{v-t}(H)\, dv ds \\ &\quad + \int_0^t \chi(s,t)\, dz_s. \end{aligned} \tag{61}$$

Thus the strong solution of (1) is explicitly exhibited by (61), showing that property (i) is satisfied.

Remark. One might think that property (ii) for the control (34) follows from our previous remark about the equivalence of the σ-fields $\mathcal{F}_t^z$ and $\mathcal{F}_t^x$. However, the remark is valid for controls which are admissible. The fact that property (ii) held was used to show that (19) implied (20) in the proof of that remark. At this point we do not know that property (ii) holds for (34) and must prove it.

To show that (ii) holds, notice that if we solve (55) for dz and substitute this into (58) and (57), we obtain the equation

$$\begin{aligned} \pi_t(f) &= P_0(P_t f) \\ &\quad - \int_0^t [\pi_r(P_{t-r}f)H') - \pi_r(P_{t-r}f)\pi_r(H')][BN^{-1}B'J_r^\pi - \pi_r(H)]\, dr \\ &\quad + \int_0^t [\pi_r((P_{t-r}f)H') - \pi_r(P_{t-r}(f))\pi_r(H')] \\ &\quad \times (dx_r - (Ax_r - BN^{-1}B'K_r x_r)\, dr), \end{aligned} \tag{62}$$

driven by x_t for $\pi_t(f)$. Now, from (33) and (34), it will follow that (ii) holds if we can show for each t that $\pi_t(f)$ is $\mathcal{F}_t^x$-measurable. Let us show that this holds by solving (62) by successive approximations where each of the approximating solutions has this property.

THEOREM 2. *Define*

$$\pi_t^0(f) \triangleq P_0(P_t f), \tag{63}$$

and for $n > 0$,

$$\begin{aligned}\pi_t^n(f) \triangleq\ & P_0(P_t f) \\ & + \int_0^t [\pi^{n-1}((P_{t-r}f)H') - \pi_r^{n-1}(P_{t-r}f)\pi_r^{n-1}(H')] \\ & \quad \times [BN^{-1}B'J_r^{\pi^{n-1}} - \pi_r^{n-1}(H)]\,dr \\ & + \int_0^t [\pi^{n-1}((P_{t-r}f)H') - \pi_r^{n-1}(P_{t-r}f)\pi_r^{n-1}(H')] \\ & \quad \times [dx_r - (Ax_r - BN^{-1}B'K_r x_r)\,dr].\end{aligned} \tag{64}$$

Then for each $n \geq 0$, $\pi_t^n(f)$ *is* $\mathcal{F}_t^x$*-measurable for each* t *and*

$$\lim_{n\to\infty} E[(\pi_t^n(f) - \pi_t(f))^2] = 0, \tag{65}$$

where $\pi_t(f)$ *is the unique solution of* (23) *and* (24).

Proof. First notice that $\pi_t^0(f)$ is a deterministic function of t and hence is $(\mathcal{F}_0^x \subset \mathcal{F}_t^x)$-measurable for each t. Now this and an induction using (64) show $\pi_t^n(f)$ is $\mathcal{F}_t^x$-measurable for each fixed t.

Let us next prove by induction that there is a constant K such that

$$E[(\pi_t(f) - \pi_t^n(f))^2] \leq \frac{K^{n+1}t^{n+1}}{(n+1)!}. \tag{66}$$

If $|f|$ is defined by

$$|f| = \sup_{s\in\mathcal{S}} |f(s)|, \tag{67}$$

notice that

$$|\pi_t(f)| \leq |f| \quad \text{and} \quad |P_{t-r}(f)| \leq |f|. \tag{68}$$

For brevity, define

$$C_{rt}^{\pi} \triangleq \pi_r((P_{t-r}f)H') - \pi_r(P_{t-r}f)\pi_r(H'). \tag{69}$$

Using the inequalities

$$\left(\sum_{i=1}^n a_i\right)^2 \leq n\sum_{i=1}^n a_i^2 \tag{70}$$

and (68), we see that

$$|C_{rt}^{\pi}|^2 \leq 4|f|^2|H|^2. \tag{71}$$

Now using (57), (63), (69), and (71),

$$E[(\pi_t(f) - \pi_t^0(f))^2] = E\left(\int_0^t C_{rt}^{\pi}\,d\nu\right)^2 = \int_0^t E|C_{rt}^{\pi}|^2\,dt \leq 4|f|^2|H|^2 t. \tag{72}$$

Thus (66) will hold for $n = 0$ if

$$K \geq 4|f|^2|H|^2. \tag{73}$$

For $n > 1$, rewriting $\pi_t^n(f)$ in terms of $d\nu$ rather than dx, and using (58) and (55), we have

$$\pi_t^n(f) = P_0(P_t f) + \int_0^t C_{rt}^{\pi^{n-1}}[D_r^{\pi^{n-1}} - D_r^{\pi}]\,dt + \int_0^t C_{rt}^{\pi^{n-1}}\,d\nu, \tag{74}$$

where we have used the abbreviated notation

$$D_r^{\pi} \triangleq BN^{-1}B'J_r^{\pi} - \pi_r(H). \tag{75}$$

Note that we may also write

$$\begin{aligned} D_r^{\pi} &= BN^{-1}B' \int_r^T \phi(r,s)K_s\pi_r(P_{s-r}H)\,ds - \pi_r(H) \\ &= \pi_r BN^{-1}B' \int_r^T \phi(r,s)K_s(P_{s-r}H) - H) \triangleq \pi_r(\gamma_r). \end{aligned} \tag{76}$$

Thus (68), (70), and (76) imply

$$|D_r^{\pi}|^2 \le |\gamma_r|^2 \le 2|B|^4|N^{-1}|^2T^2|\phi|^2|K|^2|H|^2 + 2|H|^2. \tag{77}$$

Subtracting (74) from (23), squaring, taking expected values, and using (70), we have

$$\begin{aligned} E[(\pi_t(f) - \pi_t^n(f))^2] \le\ & 2E\left[\left(\int_0^t C_{rt}^{\pi^{n-1}}[D_r^{\pi} - D_r^{\pi^{n-1}}]\,dt\right)^2\right] \\ & + 2E\left[\left(\int_0^t (C_{rt}^{\pi} - C_{rt}^{\pi^{n-1}})\,d\nu\right)^2\right]. \end{aligned} \tag{78}$$

Using Jensen's inequality on the first integral and properties of stochastic integrals on the second gives

$$\begin{aligned} E[(\pi_t(t) - \pi_t^n(f))^2] \le\ & 2E\int_0^t (C_{rt}^{\pi^{n-1}})^2[D_r^{\pi} - D_r^{\pi^{n-1}}]^2\,dt \\ & + 2E\left(\int_0^t (C_{rt}^{\pi} - C_{rt}^{\pi^{n-1}})^2\,dt\right). \end{aligned} \tag{79}$$

Now

$$\begin{aligned} C_{rt}^{\pi} - C_{rt}^{\pi_n} =\ & \pi_r((P_{t-r}f)H') - \pi_r^n((P_{t-r}f)H') \\ & + \pi_r(P_{t-r}f)(\pi_r(H') - \pi_r^n(H')) \\ & + [\pi_r(P_{t-r}f) - \pi_r^n(P_{t-r}f)]\pi_r^n(H'). \end{aligned} \tag{80}$$

So, using (70) and (68),

$$\begin{aligned} |C_{rt}^{\pi} - C_{rt}^{\pi_n}|^2 \le\ & 3|\pi_r((P_{t-r}f)H') - \pi_r^n(P_{t-r}f)H'|^2 \\ & + 3|f|^2|(\pi_r(H') - \pi_r^n(H'))^2 + 3|H'|^2|\pi_r(P_{t-r}f) - \pi_r^n(P_{t-r}f)|^2. \end{aligned} \tag{81}$$

Thus, from (79), and using (71), (76), (81), and (70),

$$
\begin{aligned}
&E[(\pi_t(f)-\pi_t^n(f))^2] \\
&\quad\le 6\int_0^t E|\pi_t((P_{t-r}f)H')-\pi^{n-1}((P_{t-r}f)H')|^2\,dr \\
&\quad\quad + 6|f|^2\int_0^t E|\pi_r(H')-\pi_r^{n-1}(H')|^2\,dr \\
&\quad\quad + 6|H'|^2\int_0^t |\pi_r(P_{t-r}f)-\pi^{n-1}(P_{t-r}f)|^2\,dr \\
&\quad\quad + 8|f|^2H'|^2\int_0^t |\pi_r(\gamma_r)-\pi_r^{n-1}(\gamma_r)|^2\,dr.
\end{aligned} \tag{82}
$$

Now using our induction hypothesis that (66) holds for $n-1$ and using (68),

$$
E[(\pi_r(f)-\pi_t^n(f))^2] \le (18|f|^2|H'|^2+8|f|^2|H'|^2|\gamma|^2)K^n\int_0^t \frac{r^n}{n!}\,dr. \tag{83}
$$

Thus (83) will imply (66) if

$$
K \ge 18|f|^2|H'|+8|f|^2|H'|^2|\gamma|^2. \tag{84}
$$

Choosing K in this way we see that induction implies that (66) must hold for all n. Now (66) implies (65).

The mean square convergence of $\pi_t^n(f)$ in the sense of (65) implies that there is a subsequence n_j so that $\pi_t^{n_j}(f)$ converges almost surely for Lebesgue measure almost every t to $\pi_t(f)$. This implies that $\pi_t(f)$ is $\mathcal{F}_t$-measurable for almost every t. Since $\pi_t(f)$ is a solution of (23) it is almost surely continuous in t. Thus $\pi_t(f)$ is $\mathcal{F}_t^x$-measurable for each t. Thus property (ii) is satisfied by the control (34).

Appendix II. In this appendix we shall establish a lemma that implies (35) and (36).

LEMMA 2. *Let $\mathcal{F}_t$ be an increasing family of σ-fields and ν an m-dimensional vector-valued $\mathcal{F}_t$-Wiener process. For $0\le t<s\le T$ let γ_{ts} be an $n\times m$-matrix-valued bounded measurable random process, such that γ_{ts} is $\mathcal{F}_t$-measurable for each fixed t, and such that α_s is an n-dimensional vector-valued, and A_t and K_t are $(n\times n)$-dimensional matrix-valued bounded measurable functions defined on $[0,T]$. Let H_{ts}, ϕ_{ts}, J_t satisfy*

$$
H_{ts} = \alpha_s + \int_0^t \gamma_{rs}\,d\nu_r, \tag{85}
$$

$$
\phi_{ts} = I + \int_t^s A_r\phi_{rs}\,dr, \tag{86}
$$

$$
J_t = \int_t^T \phi_{ts}K_sH_{ts}\,ds; \tag{87}
$$

then

$$
dJ_t = -(A_tJ_t - K_tH_{tt})\,dt + \left[\int_t^T \phi_{ts}K_s\gamma_{ts}\,ds\right]d\nu_t. \tag{88}
$$

The proof of Lemma 2 will consist mainly of a succession of interchanges of order of integration. Interchange of order of ordinary integration and stochastic integration is valid. For a proof, see Szpirglass [10]. Actually, Szpirglass's proof is stated for a rectangle but interchanges over more general regions follow by multiplying the integrand by the characteristic function of the region. We shall need the interchange over a triangle. Limits of integration change exactly as in ordinary integration when stochastic and ordinary integration are interchanged over a triangle.

Proof of Lemma 2. Using (86) in (87),

$$J_t = \int_t^T K_s H_{ts}\, ds + \int_t^T \int_t^s A_r \phi_{rs} K_s H_{ts}\, dr ds. \tag{89}$$

Interchanging the order of integration in the second integral of (89), and using that from (85) for $t < r < s$,

$$H_{ts} = H_{rs} - \int_t^r \gamma_{\tau s}\, d\nu_\tau \tag{90}$$

gives

$$\begin{aligned} J_t &= \int_t^T K_s H_{ts}\, ds + \int_t^T \int_r^T A_r \phi_{rs} K_s H_{rs}\, ds dr \\ &\quad - \int_t^T \int_r^T \left[A_r \phi_{rs} K_s \int_t^r \gamma_{\tau s}\, d\nu_\tau \right] ds dr. \end{aligned} \tag{91}$$

Interchanging the outer two orders of integration and then the inner two orders of integration in the third integral of (91) gives

$$J_t = \int_t^T K_s H_{ts}\, ds + \int_t^T A_r J_r\, dr - \int_t^T \int_t^s \left[\int_\tau^s A_r \phi_{rs}\, dr \right] K_s \gamma_{\tau s}\, d\nu_\tau\, ds. \tag{92}$$

Using (86),

$$\begin{aligned} J_t &= \int_t^T K_s H_{ts}\, ds + \int_t^T A_r J_r\, dr \\ &\quad + \int_t^T \int_t^s K_s \gamma_{\tau s}\, d\nu_\tau\, ds - \int_t^T \int_t^s \phi_{\tau s} K_s \gamma_{\tau s}\, d\nu_\tau\, ds. \end{aligned} \tag{93}$$

Using from (85) that

$$\int_t^s \gamma_{\tau s}\, d\nu_\tau = H_{ss} - H_{ts} \tag{94}$$

in the third integral of (93), and interchanging ordinary and stochastic integration in the last integral of (93), gives

$$J_t = \int_t^T K_s H_{ss}\, ds + \int_t^T A_r J_r\, dr - \int_t^T \int_\tau^T \phi_{\tau s} K_s \gamma_{\tau s}\, ds d\nu_\tau. \tag{95}$$

Now (88) follows from (95) by taking the differential.

Remark. It now follows from (33), (23), (9), (10), and Lemma 2 that (34) and (36) hold.

REFERENCES

[1] V. E. Beneš, *Quadratic approximation by linear systems controlled from partial observations*, in Stochastic Analysis: Liber Amicorum of M. Zakai, M. Merzbach, A. Schwartz and E. Meyer–Wolf, eds., Academic Press, New York, 1992.

[2] V. E. Beneš and I. Karatzas, *On the relation of Zakai's and Mortensen's equations*, SIAM J. Control Optim., 21 (1983), pp. 472–489.

[3] K. L. Helmes and R. W. Rishel, *The solution of a partially observed stochastic control problem in terms of predicted miss*, IEEE Trans. Automat. Control, 37 (1992), pp. 1462–1464.

[4] ———, *An optimal control depending on the conditional density of the unobserved state*, Proceedings of US–French Workshop on Applied Stochastic Analysis, D. Ocone and I. Karatzas, eds., pp. 144–150.

[5] H. Kunita, *Asymptotic behavior of non-linear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.

[6] H. J. Kushner, *On the dynamical equations of conditional probability density functions with application to optimal stochastic control theory*, J. Math. Anal. Appl., 8 (1964), pp. 332–344.

[7] P. L. Lions, *Viscosity Solutions of Fully Nonlinear Second Order Equations and Optimal Stochastic Control in Infinite Dimensions. Part* II: *Optimal Control of Zakai's Equation*, in Lecture Notes in Mathematics 1390, Stochastic Partial Differential Equations and Applications II, G. DaPrato and L. Tubaro, eds., Springer-Verlag, Berlin, (1988), pp. 147–170.

[8] R. E. Mortensen, *Stochastic optimal control with noisy observations*, Internat. J. Control, 4 (1966), pp. 455–464.

[9] J. G. Root, *Optimum control of non-Gaussian linear stochastic systems with inaccessible state variables*, SIAM J. Control, 7 (1969), pp. 317–322.

[10] J. Szpirglass, *Sur l'equivalence, d'equations differential a valeurs measures intervenant dans le filterage markovien nonlineare*, Ann. Inst. H. Poincare Section B, 14 (1978), pp. 33–59.

[11] W. M. Wonham, *Some applications of stochastic differential equations to optimal non-linear filtering*, SIAM J. Control, 2 (1964), pp. 371–369.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1021–1036, July 1994

008

OPTIMAL SWITCHING IN AN ECONOMIC ACTIVITY UNDER UNCERTAINTY*

KJELL ARNE BREKKE† AND BERNT ØKSENDAL‡

Abstract. This paper considers the problem of finding the optimal sequence of opening (starting) and closing (stopping) times of a multi-activity production process, given the costs of opening, running, and closing the activities and assuming that the state of the economic system is a stochastic process. The problem is formulated as an extended impulse control problem and solved using stochastic calculus. As an application, the optimal starting and stopping strategy are explicitly found for a resource extraction when the price of the resource is following a geometric Brownian motion.

Key words. impulse control, optimal switching, options

AMS subject classifications. 60G40, 60H30

1. Introduction. The theory of optimal stopping has a wide variety of applications in economics. These applications range over real and financial options, entry to a market, or optimal start of a production process under uncertainty. In other applications it will be natural to consider the possibility of the reverse action, like exiting from a market or shutting down a production. However, optimal stopping theory does not cover situations involving both these actions, such as sequential starting and stopping. For example, there are industries where part of the production process is temporarily shut down when electricity prices are too high; at high prices all workers are relocated to other tasks and when the prices fall below a certain limit, production is restarted. When is the optimal time to shut down, then to restart, then to shut down again, etc.? Problems of this type could be called *starting and stopping problems* or *optimal switching problems*. They may be regarded as special cases of sequential optimal stopping problems.

The starting and stopping problem has been considered in various contexts. It was discussed in connection with taxes and convenience yield by Brennan and Schwartz [BS]. A similar entry and exit model (but without resource extraction) has been studied by Dixit [D]. Neither of these papers gives a rigorous mathematical proof that an optimal starting and stopping strategy exists and that it has the form stated. The more general problem of starting and stopping several activities simultaneously is considered in [MZ], in the context of oil exploration. Optimal switching for alternating processes is studied in [M], which also contains references to other related works.

The purpose of this paper is to formulate problems of this type as *generalized impulse control problems* and to solve them using stochastic calculus. Impulse control problems have been thoroughly studied in [BL]. However, their results do not seem to apply to the situations we are interested in, because the cost function, $-f$, will not be positive in our case (see §2). Nevertheless, our method is inspired by their approach. For concreteness our results are applied to the following sequential stopping problem involving resource depletion with a stochastic price development, studied in [BØ]:

Suppose it costs the amount L to open a field for resource extraction, that the running/rental cost is K per time unit and that the cost of closing down a field is C. If the price of the resource under consideration is varying as a stochastic process (to be specified below), when is the optimal time to open the field and to close it? It seems reasonable that if the field is open, it may be a good strategy to continue the extraction for a while even if the price has gone below the running costs, because there may be a chance that prices could go up again

* Received by the editors April 20, 1992; accepted for publication (in revised form) April 1, 1993.
† Central Bureau of Statistics, Box 8131 Dep, N-0033 Oslo 1, Norway.
‡ Department of Mathematics, University of Oslo, Box 1053, Blindern, N-0316 Oslo 3, Norway.

and closing and re-opening the field is costly. On the other hand, even with such an optimistic point of view there is clearly a limit as to how low the prices can go before closing is the optimal strategy. Similarly, if the field is closed one would wait for a resource price that is higher than the running costs before opening again. But how high?

In [BØ] a candidate ϕ_0 for the solution of the resource extraction problem is found explicitly, as an application of a high contact principle for optimal stopping. But it is not proved there that this candidate actually is the solution. This will be established in this paper. More generally, we consider the problem of optimal starting and stopping of a multi-activity system under uncertainty. We prove that a given function satisfying certain quasivariational inequalities necessarily is the solution of the problem.

This paper is organized as follows: In §2 we formulate a general starting and stopping problem as an impulse control problem. In §3 we give sufficient conditions that a given function and its associated starting and stopping strategy solves the general problem in §2. Then in §4 we apply this to the specific problem of optimal resource extraction mentioned above.

2. A mathematical formulation of the problem. The problems mentioned in the introduction are special cases of the following general problem:

Suppose there are m possible "indicator vectors" $z_1, \dots, z_m$ of the state of the system. Let Z_t denote the indicator vector at time t, so that for all t

$$Z_t \in \{z_1, \dots, z_m\} =: \mathcal{Z} \tag{2.1}$$

We will assume that Z_t is right-continuous with left limits (cadlag).

Remark. If, for example, we consider a firm with k production activities which can be either "on/open" or "off/closed," then each indicator vector $z \in \mathcal{Z}$ can be represented as a k-tuple

$$z = (a_1, a_2, \dots, a_k)$$

where each a_i is either 0 (meaning activity i is closed) or 1 (meaning activity i is open). So in this case there are $m = 2^k$ possible indicator vectors. The components of an indicator vector z are called *indicator values*. In particular, in the resource extraction example there are just 2 indicator values, which we denote by 0 or 1 depending on whether the field is closed or open.

The firm's environment at time t, e.g., prices of output or input goods, is denoted by U_t. We assume that U_t is a stochastic process in $\mathbf{R}^n$ satisfying the following stochastic differential equation

$$dU_t = b(t, U_t, Z_t)dt + \sigma(t, U_t, Z_t)dB_t \tag{2.2}$$

where $b : \mathbf{R}^{n+1} \times \mathcal{Z} \to \mathbf{R}^n, \sigma : \mathbf{R}^{n+1} \times \mathcal{Z} \to \mathbf{R}^{n \times m}$ are Lipschitz functions with at most linear growth in the variables number $2, \dots, n+1$ and B_t denotes m-dimensional Brownian motion. (See e.g., [Ø] for basic information on stochastic differential equations and see [BL] regarding the solution of equations of type (2.2).)

The state of the whole economic system at time t is represented by the stochastic process

$$X_t = \begin{bmatrix} t \\ U_t \\ Z_t \end{bmatrix}. \tag{2.3}$$

The probability law of X_t given that $X_0 = x = (s, u, z)$ is denoted by P^x and expectation with respect to P^x is denoted by E^x.

An *impulse control* w for this system consists of a double (possibly finite) sequence

$$w = (\theta_1, \theta_2, \dots, \theta_k, \dots; \zeta_1, \zeta_2, \dots, \zeta_k, \cdots)_{k \le N} \qquad (N \le \infty) \tag{2.4}$$

where each $\theta_k \le \infty$ is a stopping time (with respect to the filtration $\{\mathcal{F}_t\}$ for the Brownian motion $\{B_t\}$), $\theta_k \le \theta_{k+1}$ and $\theta_k \to \infty$ almost surely (so if N is finite then $\theta_N \equiv \infty$). Associated to the impulse time θ_k is the impulse $\zeta_k \in \mathcal{Z}$, which is the new value of Z_t at time $t = \theta_k$.

We may regard $\theta_1, \theta_2, \cdots$ as the times when we decide to interfere with the system and the corresponding $\zeta_1, \zeta_2 \cdots$ are the new indicator values that we give the system at these times. We often simplify the notation and write $w = (\theta_1, \theta_2, \cdots)$. Let W denote the set of all impulse controls.

If $w \in W$ is applied to the system, it takes the form

$$X_t = X_t^{(w)} = \begin{bmatrix} t \\ U_t \\ \zeta_k \end{bmatrix} \quad \text{if } \theta_k \le t < \theta_{k+1}. \tag{2.5}$$

Note that $X_t^{(w)}$ is right-continuous for all $w \in W$. Let E^x denote the expected value when $X_0 = x = (s, u, z)$.

Let $f(x)$ denote the profit per time unit when the system is in the state x. For $x = (s, u, z) \in \mathbf{R}^{n+1} \times \mathcal{Z}$ and $\zeta \in \mathcal{Z}$ let $H(x, \zeta) \in \mathbf{R}$ be the cost of switching the indicator value from z to ζ when the state is $x = (s, u, z)$. Assume from now on that

$$E^x \left[\int_s^\infty |f(X_t^{(w)})| dt \right] < \infty \tag{2.6}$$

for all x and all $w \in W$. Then the expected total profit of running the system with the impulse control $w = (\theta_1, \theta_2, \cdots; \zeta_1, \zeta_2, \cdots) \in W$ is given by

$$J^w(x) = E^x \left[\int_s^\infty f(X_t^{(w)}) dt - \sum_{j=1}^\infty H(X_{\theta_j -}, \zeta_j) \right], \tag{2.7}$$

where $X_{\theta_j -} = \lim_{t \uparrow \theta_j} X_t$.

We assume that the switching cost function $H : \mathbf{R}^{n+1} \times \mathcal{Z} \times \mathcal{Z} \to \mathbf{R}^+$ satisfies

$$H(x, \zeta) > 0 \quad \text{for all} \quad x \in \mathbf{R}^{n+1} \times \mathcal{Z} \quad \text{and all} \quad \zeta \ne z \tag{2.8}$$

and, if $\mathcal{Z}$ consists of more than 2 elements:

$$H(s, u, z, \zeta_2) \le H(s, u, z, \zeta_1) + H(s, u, \zeta_1, \zeta_2) \quad \text{if } z \ne \zeta_1 \ne \zeta_2 \ne z. \tag{2.9}$$

We also assume that

$$(s, u) \to H(s, u, z, \zeta) \quad \text{is continuous for all } z, \zeta. \tag{2.10}$$

(The values of H when $z = \zeta$ are not used, so we only need to define $H(s, u, z, \zeta)$ for $z \ne \zeta$.)

Remarks. (1) Condition (2.9) states that if we want to switch from indicator state z to indicator state ζ_2, then it is not more expensive to do this directly (in one step) than in two steps, via an intermediate indicator value ζ_1. For example, if

$$H(s, u, z, \zeta) = e^{-\rho s} H_0(z, \zeta) \qquad (\rho \text{ constant}),$$

then (2.9) becomes the "triangle inequality"

$$H_0(z,\zeta_2) \le H_0(z,\zeta_1) + H_0(\zeta_1,\zeta_2) \qquad z \ne \zeta_1 \ne \zeta_2 \ne z.$$

If we are given a function H satisfying (2.8) and (2.10) we can always modify it to satisfy (2.9) as well, by putting

$$\tilde{H}(s,u,z,\zeta) = H(s,u,z,\zeta) \wedge \min_{z\ne\zeta_1\ne\zeta} \{H(s,u,z,\zeta_1) + H(s,u,\zeta_1,\zeta)\}.$$

If H fails to satisfy (2.9) this means that the switching cost of going from z to ζ may be reduced by first jumping to an indicator vector ζ_1 and then immediately jumping to ζ. In this case Z_t cannot be right-continuous. By introducing $\tilde{H}$ we are simply saying that we regard this kind of immediate switch from z to ζ_1 to ζ as a direct switch from z to ζ, neglecting the intermediate indicator vector ζ_1. This makes the problem essentially unchanged and now the corresponding processes Z_t will be cadlag.

(2) In the resource extraction example the switching cost function has the values

$$\begin{aligned} H(s,u,0,1) &= Le^{-\rho s} \qquad \text{(discounted opening cost)} \\ H(s,u,1,0) &= Ce^{-\rho s} \qquad \text{(discounted closing cost),} \end{aligned} \tag{2.11}$$

where $\rho > 0$ is a (constant) discount factor.

We can now formulate the switching problem as follows:

PROBLEM 2.1. Find for all $x = (s,u,z)$

$$\tilde{\phi}(x) := \sup_{w\in W} J^w(x)$$

and find, if possible, an *optimal* impulse control $\tilde{w}$, i.e., find $\tilde{w} \in W$ such that

$$\tilde{\phi}(x) = J^{\tilde{w}}(x).$$

Remark. This is essentially an impulse control problem of the type considered in [BL]. However, in [BL] it is assumed that $-f$ is positive (or lower bounded), and this is not a reasonable assumption in our economic application. Therefore it is not possible to apply their results directly to our situation.

In [BØ] a candidate $\phi_0(x)$ for the solution of Problem 2.1 in the specific application of starting and stopping a resource extraction (see §4) was found by adopting the following dynamic programming argument: Suppose the system initially is in state $x = (t,u,z)$. Then if at a stopping time τ we interfere and start/stop the system, the system gets the impulse $\zeta = 1 - z$ and then the new state becomes $X_\tau = (\tau, U_\tau, Z_\tau)$, where $Z_\tau = \zeta$. The cost of this operation is

$$H(X_{\tau^-},\zeta) = H(\tau, U_\tau, z, 1-z)$$

where H is given by (2.11) above. From then on the maximal profit is $\tilde{\phi}(X_\tau)$. This procedure can of course at most be optimal. We conclude that, for all stopping times τ,

$$\tilde{\phi}(x) \ge E^x\left[\int_s^\tau f(X_t)dt - H(X_{\tau^-}, 1 - Z_{\tau^-}) + \tilde{\phi}(X_\tau)\right]. \tag{2.12}$$

If an optimal impulse control $\tilde{w} = (\tilde{\theta}_1, \tilde{\theta}_2, \cdots)$ exists, then by choosing $\tau = \tilde{\theta}_1$ we get equality in (2.12). Hence $\tilde{\phi}$ must satisfy the equation

$$\tilde{\phi}(x) = \sup_\tau E^x\left[\int_s^\tau f(X_t)dt - H(X_{\tau^-}, 1 - Z_{\tau^-}) + \tilde{\phi}(X_\tau)\right]. \tag{2.13}$$

Using the "high contact principle" or "smooth pasting" a solution $\phi_0(x)$ of equation (2.13) for the resource extraction problem was found in [BØ]. For more information on smooth pasting see e.g., [S]. However, this does not prove that $\phi_0 = \tilde{\phi}$, because there may (a priori) be several solutions of equation (2.13). We will in §4 prove that we indeed have $\phi_0 = \tilde{\phi}$ (under certain conditions). This will be obtained as an application of the more general results we develop in §3.

3. Solution of the optimal switching problem. From now on we put, for fixed $z \in \mathcal{Z}$,

$$Y_t = (t, U_t, z) \qquad (= (t, U_t) \text{ if we suppress } z) \tag{3.1}$$

so that Y_t represents the state of the system corresponding to the "non-interference" impulse control $w_\infty = (\theta_1)$ where $\theta_1 = \infty$. Then Y is a diffusion with generator A given by

$$A = \frac{\partial}{\partial s} + \sum_{i=1}^{n} b_i \frac{\partial}{\partial u_i} + \frac{1}{2} \sum_{i,j=1}^{n} (\sigma\sigma^T)_{ij} \frac{\partial^2}{\partial u_i \partial u_j}. \tag{3.2}$$

In the following we suppress the constant z and regard Y_t as the $(n+1)$-dimensional process (t, U_t).

The following concept is useful:

DEFINITION 3.1. We say that a continuous function $g(x)$ is *stochastically* C^2 in a domain $D \subset \mathbf{R}^{n+1}$ (with respect to Y_t) if all the first partial derivatives of g with respect to t and u and all the second partial derivatives of g with respect to u exist almost everywhere in D with respect to the *Green measures* $G(y, \cdot)$ of Y_t and the following *generalized Dynkin formula* holds:

$$E^y[g(Y_{\theta'})|\mathcal{F}_\theta] = g(Y_\theta) + E^y\left[\int_\theta^{\theta'} Ag(Y_t)dt|\mathcal{F}_\theta\right] \tag{3.3}$$

for all $y \in D$ and for all stopping times $\theta \le \theta' \le \tau_D$, where $\mathcal{F}_\theta$ is the filtration generated by $\{B_{t\wedge\theta}(\cdot)\}_{t\ge 0}$,

$$\tau_D = \inf\{t > 0; Y_t \notin D\} \tag{3.4}$$

and we assume $E^y[\tau_D] < \infty$.

Remark. The *Green measure* $G(y, \cdot)$ is defined by

$$G(y, F) = E^y\left[\int_0^{\tau_D} \chi_F(Y_t)dt\right]; \qquad y \in D, F \quad \text{Borel set in } D.$$

In (3.3) Ag is the operator A applied to g, which makes sense almost everywhere $G(y, \cdot)$ and therefore makes sense in (3.3).

By the classical Dynkin formula all C^2 functions g that satisfy $E^y[\int_0^{\tau_D} |Ag(Y_t)|dt] < \infty$ for all y are stochastically C^2. In [BØ, Lem. 1] conditions are given which imply that a C^1 function that is C^2 outside a "thin" (in a measure sense) regular set is stochastically C^2. This turns out to be sufficient for the application in §4.

Recall that we have assumed that

$$E^x\left[\int_0^\infty |f(X_t^{(w)})|dt\right] < \infty \tag{3.5}$$

for all x and all $w \in W$.

LEMMA 3.2. *Suppose $\phi(s,u,z)$ is a stochastically C^2 function in $D = \mathbf{R}^{n+1}$ with respect to Y_t satisfying the three conditions*

$$\phi(t, U_t, z) \to 0 \quad \text{as } t \to \infty \quad \text{a.s. } P^y \quad \text{for all } z \in \mathcal{Z}; y = (s,u,z), \tag{3.6}$$

the family

$$\{\phi(\tau, U_\tau, z)\}_{\tau \in \mathcal{T}} \tag{3.7}$$

is uniformly integrable with respect to P^y for all $z \in \mathcal{Z}$, where $\mathcal{T}$ is the set of all $\mathcal{F}_t$-stopping times and

$$A\phi + f \leq 0 \quad \text{a.e. with respect to } G(y,\cdot). \tag{3.8}$$

Let $s \leq \theta \leq \theta' \leq \infty$ be two stopping times. Then for all $y = (s,u,z)$ and all z_0 we have

$$\phi(\theta, U_\theta, z_0) \geq E^y\left[\int_\theta^{\theta'} f(Y_t)dt + \phi(\theta', U_{\theta'}, z_0)|\mathcal{F}_\theta\right]. \tag{3.9}$$

(We interpret $\phi(\tau, U_\tau, z_0)$ as 0 if $\tau = \infty$ ($\tau = \theta$ or θ')).

Remark. The requirement that ϕ be stochastically C^2 corresponds to the "high contact" condition in optimal stopping. See [BØ].

Proof. Choose a constant $T < \infty$ and apply the generalized Dynkin formula (3.3) to $g(s,u,z) = \phi(s,u,z)$:

$$E^y[\phi(\theta' \wedge T, U_{\theta' \wedge T}, z)|\mathcal{F}_\theta] = \phi(\theta \wedge T, U_{\theta \wedge T}, z) + E^y\left[\int_{\theta \wedge T}^{\theta' \wedge T} A\phi(t, U_t, z)dt|\mathcal{F}_\theta\right].$$

By (3.8) this gives

$$\phi(\theta \wedge T, U_{\theta \wedge T}, z) \geq E^y\left[\int_{\theta \wedge T}^{\theta' \wedge T} f(Y_t)dt + \phi(\theta' \wedge T, U_{\theta' \wedge T}, z)|\mathcal{F}_\theta\right].$$

Now let $T \to \infty$. Then by continuity

$$\phi(\theta \wedge T, U_{\theta \wedge T}, z) \to \phi(\theta, U_\theta, z)$$

and by (3.5)

$$E^y\left[\int_{\theta \wedge T}^{\theta' \wedge T} f(Y_t)dt\right] \to E^y\left[\int_\theta^{\theta'} f(Y_t)dt\right].$$

Finally, by (3.6) and (3.7)

$$E^y[\phi(\theta' \wedge T, U_{\theta' \wedge T}, z)|\mathcal{F}_\theta] \to E^y[\phi(\theta', U_{\theta'}, z)|\mathcal{F}_\theta].$$

This gives (3.9).

Define the *switching operator* M on the family $\mathcal{H}$ of Borel measurable functions on $\mathbf{R}^{n+1} \times \mathcal{Z}$ by

$$Mh(s,u,z) = \max_{\zeta \in \mathcal{Z} \setminus \{z\}} \{h(s,u,\zeta) - H(x,\zeta)\}; \qquad h \in \mathcal{H}, \qquad x = (s,u,z) \tag{3.10}$$

where H is the switching cost function (see (2.8)–(2.10)).

Note that

$$h(s,u,z) \geq Mh(s,u,z) \Leftrightarrow h(s,u,z) \geq h(s,u,\zeta) - H(x,\zeta) \quad \text{for all } \zeta \neq z. \quad \square \tag{3.11}$$

We are now ready for the first main result of this paper.

THEOREM 3.3. *Let ϕ be a stochastically C^2 function (with respect to Y_t) satisfying* (3.6), (3.7), (3.8), *as well as the condition*

$$\phi \geq M\phi \quad \text{everywhere} \tag{3.12}$$

Then

$$\phi(x) \geq J^w(x) \quad \text{for all} \quad w \in W \quad \text{and all } x = (s,u,z). \tag{3.13}$$

Proof. Let $w = (\theta_1, \theta_2, \cdots)$ with $\theta_1 \geq s$, let $X_t = X_t^{(w)} = (t, U_t, Z_t)$ and put $\theta_0 = s$. Since $\phi \geq M\phi$ we get by (3.11) and Lemma 3.2 applied to $\theta = \theta_k, \theta' = \theta_{k+1}$, and $z_0 = Z_{\theta_k}; k = 0,1,2,\ldots$:

$$\begin{aligned}\phi(\theta_k, U_{\theta_k}, Z_{\theta_k}) &\geq E^y\left[\int_{\theta_k}^{\theta_{k+1}} f(X_t)dt + \phi(\theta_{k+1}, U_{\theta_{k+1}}, Z_{\theta_k})|\mathcal{F}_{\theta_k}\right] \\ &\geq E^y\left[\int_{\theta_k}^{\theta_{k+1}} f(X_t)dt + \phi(\theta_{k+1}, U_{\theta_{k+1}}, \zeta_{k+1}) - H(X_{\theta_{k+1}^-}, \zeta_{k+1})|\mathcal{F}_{\theta_k}\right].\end{aligned} \tag{3.14}$$

Now $\zeta_{k+1} = Z_{\theta_{k+1}}$ by (2.5); so if we take expectation and sum from $k = 0$ to $k = n-1$, we get, with $y = (s,u,z)$,

$$\begin{aligned}&\phi(s,u,z) + \sum_{k=1}^{n-1} E^y[\phi(\theta_k, U_{\theta_k}, Z_{\theta_k})] \\ &\geq E^y\left[\int_s^{\theta_n} f(X_t)dt - \sum_{k=1}^{n} H(X_{\theta_k^-}, Z_{\theta_k}) + \sum_{k=1}^{n} \phi(\theta_k, U_{\theta_k}, Z_{\theta_k})\right].\end{aligned} \tag{3.15}$$

Hence

$$\phi(s,u,z) \geq E^y\left[\int_s^{\theta_n} f(X_t)dt - \sum_{k=1}^{n} H(X_{\theta_k^-}, Z_{\theta_k}) + \phi(\theta_n, U_{\theta_n}, Z_{\theta_n})\right]. \tag{3.16}$$

Now let $n \to \infty$. Then $\theta_n \to \infty$ and by (3.6) and (3.7) we get

$$\phi(s,u,z) \geq E^y\left[\int_s^{\infty} f(X_t)dt + \sum_{k=1}^{\infty} H(X_{\theta_k^-}, Z_{\theta_k})\right],$$

which is (3.13). $\square$

Next we find an optimal impulse control and the corresponding minimal expected cost.

THEOREM 3.4. *Suppose $\hat{\phi} = \hat{\phi}(t,u,z)$ is a stochastically C^2 function (with respect to Y_t) satisfying* (3.6), (3.7), (3.8), (3.12) *and in addition that*

$$A\hat{\phi} + f = 0 \quad \text{on } \{(s,u,z); \hat{\phi}(s,u,z) > M\hat{\phi}(s,u,z)\}. \tag{3.17}$$

Define the impulse control $\hat{w} = (\hat{\theta}_1, \hat{\theta}_2, \ldots; \hat{\zeta}_1, \hat{\zeta}_2, \cdots)$ *as follows: Put*

$$\hat{\theta}_1 = \inf\{t > 0; \hat{\phi}(X_t^{(0)}) = M\hat{\phi}(X_t^{(0)})\}, \quad \text{where } X_t^0 = Y_t, \tag{3.18}$$

and choose $\hat{\zeta}_1$ *such that*

$$M\hat{\phi}(X_{\hat{\theta}_1}^{(0)}) = \hat{\phi}(\hat{\theta}_1, U_{\hat{\theta}_1}, \hat{\zeta}_1) - H(X_{\hat{\theta}_1}^{(0)}, \hat{\zeta}_1). \tag{3.19}$$

Define

$$Z_t^{(1)} = \begin{cases} z & \text{if } 0 \le t < \hat{\theta}_1 \\ \hat{\zeta}_1 & \text{if } \hat{\theta}_1 \le t, \end{cases}$$

and put

$$dX_t^{(1)} = \begin{bmatrix} dt \\ dU_t \\ dZ_t^{(1)} \end{bmatrix},$$

i.e., $X_t^{(1)}$ *is the result of applying the impulse control* $\hat{w}_1 = (\hat{\theta}_1, \infty; \hat{\zeta}_1)$ *to* Y_t.

Inductively, if stopping times $0 \le \hat{\theta}_1 \le \hat{\theta}_2 \le \cdots \le \hat{\theta}_k$ *with corresponding impulses* $\hat{\zeta}_1, \ldots, \hat{\zeta}_k$ *have been constructed, define*

$$\hat{\theta}_{k+1} = \inf\{t > \hat{\theta}_k; \hat{\phi}(X_t^{(k)}) = M\hat{\phi}(X_t^{(k)})\}, \qquad k \ge 0, \tag{3.20}$$

where for $k \ge 1$, $X_t^{(k)}$ *is the result of applying the impulse control* $\hat{w}_k = (\hat{\theta}_1, \cdots, \hat{\theta}_k, \infty)$ *to* Y_t. *Next choose* $\hat{\zeta}_{k+1}$ *such that*

$$M\hat{\phi}(\hat{\theta}_{k+1}, U_{\hat{\theta}_{k+1}}, \hat{\zeta}_k) = \hat{\phi}(\hat{\theta}_{k+1}, U_{\hat{\theta}_{k+1}}, \hat{\zeta}_{k+1}) - H(X_{\hat{\theta}_{k+1}}^{(k)}, \hat{\zeta}_{k+1}). \tag{3.21}$$

Then $\hat{w} \in W$ *and*

$$\hat{\phi}(x) = J^{\hat{w}}(x); \tag{3.22}$$

$$\hat{w} \text{ is optimal for Problem } 2.1, \tag{3.23}$$

and

$$\hat{\phi} = \tilde{\phi}. \tag{3.24}$$

Proof. We repeat the arguments of the proofs of Lemma 3.2 and Theorem 3.3. First note that between $\hat{\theta}_k$ and $\hat{\theta}_{k+1}$ we have $A\hat{\phi} = -f$ so we get equality if we apply Lemma 3.2 to $\theta_k = \hat{\theta}_k, \theta_{k+1} = \hat{\theta}_{k+1}$, and $\phi = \hat{\phi}$. Therefore we get equality in (3.15) so that for all n we have

$$\hat{\phi}(s,u,z) = E^y\left[\int_s^{\hat{\theta}_n} f(X_t)dt - \sum_{k=1}^{n} H(X_{\hat{\theta}_k^-}, Z_{\hat{\theta}_k}) + \hat{\phi}(\hat{\theta}_n, U_{\hat{\theta}_n}, Z_{\hat{\theta}_n})\right]. \tag{3.25}$$

Put $F = \{\omega; \lim_{n\to\infty} \hat{\theta}_n(\omega) < \infty\}$ and $\hat{\theta} = \lim_{n\to\infty} \hat{\theta}_n$. Then letting $n \to \infty$ in (3.25) we get

$$\hat{\phi}(s,u,z) = E^y\left[\int_s^{\hat{\theta}} f(X_t)dt - \sum_{k=1}^{\infty} H(X_{\hat{\theta}_k^-}, Z_{\hat{\theta}_k}) + \hat{\phi}(\hat{\theta}, U_{\hat{\theta}}, Z_{\hat{\theta}}) \cdot \chi_F\right]. \tag{3.26}$$

But if $\omega \in F$, we have that $(\hat{\theta}_k, U_{\hat{\theta}_k}) \to (\hat{\theta}, U_{\hat{\theta}})$ and therefore by (2.9) there exists $a(\omega) > 0$ such that

$$H(X_{\hat{\theta}_k^-}, Z_{\hat{\theta}_k}) \geq a(\omega) \quad \text{for } w \in F \quad \text{for all } k,$$

which gives

$$\sum_{k=1}^{\infty} H(X_{\hat{\theta}_k^-}, Z_{\hat{\theta}_k}) = \infty \quad \text{for } \omega \in F.$$

From (3.26) and (3.5) we can conclude that $P^x(F) = 0$, which shows that $\hat{\theta}_n \to \infty$ almost surely. Therefore $\hat{w} \in W$.

Now we can apply Theorem 3.3 to $\phi = \hat{\phi}, w = \hat{w}$, and by condition (3.17) we get equality in (3.13). So

$$\hat{\phi}(x) = J^{\hat{w}}(x),$$

while from Theorem 3.3

$$\hat{\phi}(x) \geq J^{w}(x) \quad \text{for all } w \in W.$$

It follows that $\hat{w}$ is optimal and that

$$\hat{\phi}(x) = \sup_{w \in W} J^{w}(x) = \tilde{\phi}(x), \text{ as claimed.} \qquad \square$$

Remark. Note that $\hat{\theta}_k$ is the first exit time after $\hat{\theta}_{k-1}$ for $X_t^{(\hat{w})}$ from the set

$$D = \{x; \hat{\phi}(x) > M\hat{\phi}(x)\}. \tag{3.27}$$

Therefore, writing $X_t = X_t^{(\hat{w})}$, we have

$$\begin{aligned} \hat{\phi}(X_{\hat{\theta}_k}) &= M\hat{\phi}(X_{\hat{\theta}_k^-}) + H(X_{\hat{\theta}_k^-}, \hat{\zeta}_k) \\ &= \hat{\phi}(X_{\hat{\theta}_k^-}) + H(X^{(k)}_{\hat{\theta}_k^-}, \hat{\zeta}_k). \end{aligned} \tag{3.28}$$

Suppose we have strict inequality in (2.9), i.e.,

$$H(s,u,z,\zeta_2) < H(s,u,z,\zeta_1) + H(s,u,\zeta_1,\zeta_2) \quad \text{if } z \neq \zeta_1 \neq \zeta_2 \neq z. \tag{3.29}$$

Then if $\hat{\zeta}_{k-1} \neq \zeta \neq \hat{\zeta}_k$, we have

$$\begin{aligned} \hat{\phi}(X_{\hat{\theta}_k}) &\geq \hat{\phi}(\hat{\theta}_k, U_{\hat{\theta}_k}, \zeta) - H(X_{\hat{\theta}_k^-}, \zeta) + H(X_{\hat{\theta}_k^-}, \hat{\zeta}_k) \\ &= \hat{\phi}(\hat{\theta}_k, U_{\hat{\theta}_k}, \zeta) - H(X_{\hat{\theta}_k}, \zeta) \\ &\quad + [H(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_{k-1}, \hat{\zeta}_k) + H(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_k, \zeta) - H(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_{k-1}, \zeta)] \\ &> \hat{\phi}(\hat{\theta}_k, U_{\hat{\theta}_k}, \zeta) - H(X_{\hat{\theta}_k}, \zeta), \quad \text{by (3.29).} \end{aligned} \tag{3.30}$$

Moreover, by (3.28) we have

$$\begin{aligned} \hat{\phi}(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_k) &= \hat{\phi}(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_{k-1}) + H(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_{k-1}, \hat{\zeta}_k) \\ &> \hat{\phi}(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_{k-1}) - H(\hat{\theta}_k, U_{\hat{\theta}_k}, \hat{\zeta}_k, \hat{\zeta}_{k-1}). \end{aligned} \tag{3.31}$$

Combining (3.30) and (3.31) we get

$$\hat{\phi}(X_{\hat{\theta}_k}) > M\hat{\phi}(X_{\hat{\theta}_k}). \tag{3.32}$$

So we see that if (3.29) holds, then the new impulse $\hat{\zeta}_k$ brings the state X_t back into D at the instant $\hat{\theta}_k$ when X_t first hits ∂D (the boundary of D) after $\hat{\theta}_{k-1}$. Thus the optimal strategy can be illustrated as in Fig. 1.

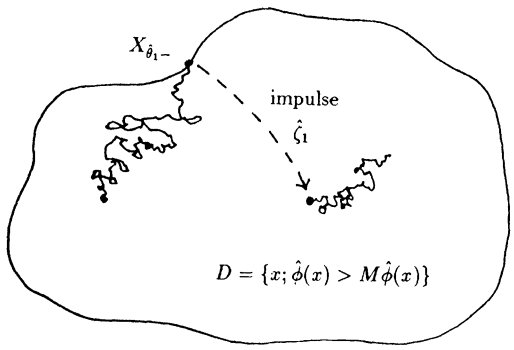


FIG. 1

4. Application to resource extraction. We assume that the price P_t at time t per unit of the resource follows a geometric Brownian motion. This means that P_t is the solution of a stochastic differential equation of the form

$$dP_t = \alpha P_t dt + \beta P_t dB_t, \tag{4.1}$$

where α, β are constants and B_t is a one-dimensional Brownian motion. The solution P_t of (4.1) is

$$P_t = P_0 \exp\left(\left(\alpha - \frac{1}{2}\beta^2\right)t + \beta B_t\right) \quad \text{for } t \geq 0. \tag{4.2}$$

Let Q_t denote the stock of remaining resources in the field. We assume that when the field is open, extraction rate is proportional to the amount of remaining reserves. In other words,

$$dQ_t = -\lambda Z_t Q_t dt, \tag{4.3}$$

where $\lambda > 0$ is a constant and

$$Z_t = \begin{cases} 1 & \text{if the field is open at time } t \\ 0 & \text{if the field is closed at time } t. \end{cases} \tag{4.4}$$

The state X_t of the system at time t is characterized by the four quantities t, P_t, Q_t, Z_t:

$$X_t = \begin{bmatrix} t \\ P_t \\ Q_t \\ Z_t \end{bmatrix}. \tag{4.5}$$

If there is a constant running cost $K > 0$ per time unit, the net discounted profit rate f is given by

$$f(x) = f(s, p, q, z) = (\lambda pq - K)ze^{-\rho s}. \tag{4.6}$$

So in this case Problem 2.1 becomes

$$\tilde{\phi}(x) = \sup_{w \in W}\left\{E^x\left[\int_s^\infty (\lambda P_t Q_t - K)Z_t e^{-\rho t}dt - \sum_j H(X_{\theta_j^-}, 1 - Z_{\theta_j^-})\right]\right\}, \tag{4.7}$$

where H is given by (2.11).

Remark. (1) Note that if we define

$$\Gamma_t = P_t Q_t,$$

then Γ_t satisfies the stochastic differential equation

$$d\Gamma_t = (\alpha - \lambda Z_t)\Gamma_t dt + \beta \Gamma_t dB_t.$$

Furthermore, P_t and Q_t only enter the objective function as $P_t Q_t = \Gamma_t$. Hence the problem could have been reformulated using only Γ_t, and consequently p and q will enter the solution in the form of the product pq.

(2) It is natural to ask if a better performance could be obtained if, instead of either having the field open at full production or entirely closed, we allow the field to be partially open at all times. If we assume that we can avoid opening and closing costs this way, but have the same running cost K, the problem can be formulated as a stochastic control problem as follows:

$$\Phi(x) = \sup_{\mu_t} E^x \left[\int_s^\infty (\lambda \mu_t P_t Q_t - K) e^{-\rho t} dt \right],$$

where $\mu_t = \mu(X_t) \in (0,1)$ represents the degree of production (fraction of full production) we choose at state X_t and where

$$dQ_t = -\lambda \mu_t Q_t dt,$$

while P_t is as before. The Hamilton-Jacobi-Bellman equation for this problem states that (see e.g., [Ø, Ch. 11]):

$$\sup_{m \in (0,1)} \left\{ (\lambda m p q - K) e^{-\rho t} + \frac{\partial \Phi}{\partial t} + \alpha p \frac{\partial \Phi}{\partial p} + \frac{1}{2} \beta^2 p^2 \frac{\partial^2 \Phi}{\partial p^2} - \lambda m q \frac{\partial \Phi}{\partial q} \right\} = 0$$

and that an optimal choice of μ (if it exists) is a value of m for which the supremum is attained. However, in this case the expression is affine in m, so it is clear that no such $m \in (0,1)$ exists. This indicates that the optimal production is "bang-bang": either full production or no production at all. Therefore it is plausible that it suffices to consider the sequential stopping problem (4.7).

Using the high contact principle it is proved in [BØ] that a solution $\phi_0(x) = \phi_0(s,p,q,z)$ of the dynamic programming equation (2.12) corresponding to (4.7) is given by

$$\phi_0(s,p,q,z) = e^{-\rho s} \psi_0(p,q,z), \tag{4.8}$$

where

$$\psi_0(p,q,z) = \begin{cases} u(p,q) - L & \text{if } z = 0 \,\&\, p \geq \xi/q \\ zu(p,q) + (1-z)v(p,q) & \text{if } z = 0 \,\&\, p < \xi/q \\ & \text{or } z = 1 \,\&\, p > \eta/q \\ v(p,q) - C & \text{if } z = 1 \,\&\, p \leq \eta/q. \end{cases} \tag{4.9}$$

Here $u(p,q)$ refers to the open field, and $v(p,q)$ refers to the closed field.

It is proved in [BØ] that $u(p,q)$ and $v(p,q)$ satisfy the following partial differential equations:

$$-ru - \lambda q \frac{\partial u}{\partial q} + \alpha p \frac{\partial u}{\partial p} + \frac{1}{2} \beta^2 p^2 \frac{\partial^2 u}{\partial p^2} = -\lambda p q + K$$

and

$$-rv + \alpha p \frac{\partial v}{\partial p} + \frac{1}{2}\beta^2 p^2 \frac{\partial^2 v}{\partial p^2} = 0.$$

Combined with the boundary values deduced from the high contact principle, this gives the following expressions for u and v:

$$u(p,q) = \frac{pq}{\rho + \lambda - \alpha} - \frac{K}{\rho} + k_1 (pq)^{\nu} \tag{4.10}$$

and

$$v(p,q) = k_2 (pq)^{\gamma} \tag{4.11}$$

with

$$\gamma = \beta^{-2}\left[-\alpha + \frac{1}{2}\beta^2 + \sqrt{\left(\alpha - \frac{1}{2}\beta^2\right)^2 + 2\rho\beta^2}\right] > 1 \quad \text{if } \rho > \alpha, \tag{4.12}$$

$$\nu = \beta^{-2}\left[-\alpha + \lambda + \frac{1}{2}\beta^2 - \sqrt{\left(\alpha - \lambda - \frac{1}{2}\beta^2\right)^2 + 2\rho\beta^2}\right] < 0, \tag{4.13}$$

and $k_1, k_2, \xi > \eta > 0$ are constants that solve the following system of equations:

$$\frac{\xi}{\rho + \lambda - \alpha} + k_1 \xi^{\nu} = k_2 \xi^{\gamma} + \frac{K}{\rho} + L, \tag{4.14}$$

$$\frac{\xi}{\rho + \lambda - \alpha} + \nu k_1 \xi^{\nu} = \gamma k_2 \xi^{\gamma}, \tag{4.15}$$

$$\frac{\eta}{\rho + \lambda - \alpha} + k_1 \eta^{\nu} = k_2 \eta^{\gamma} + \frac{K}{\rho} - C, \tag{4.16}$$

$$\frac{\eta}{\rho + \lambda - \alpha} + \nu k_1 \eta^{\nu} = \gamma k_2 \eta^{\gamma}. \tag{4.17}$$

(Note that (4.14)–(4.17) are the "high contact" equations; (4.14) and (4.15) imply that ψ_0 is C^1 across ξ, and (4.16) and (4.17) imply that ψ_0 is C^1 across η.)

It is proved in [BØ] that if we assume that

$$\rho > \alpha \tag{4.18}$$

and that

$$\text{the system (4.14)–(4.17) has a solution} \quad k_1, k_2, \xi > \eta > 0, \tag{4.19}$$

then (2.13) holds for ϕ_0, i.e.,

$$\phi_0(x) = \sup_{\tau} E^x \left[\int_s^{\tau} (\lambda P_t Q_t - K) Z_t e^{-\rho t} dt - H(X_{\tau^-}, 1 - Z_{\tau^-}) + \phi_0(X_{\tau})\right]. \tag{4.20}$$

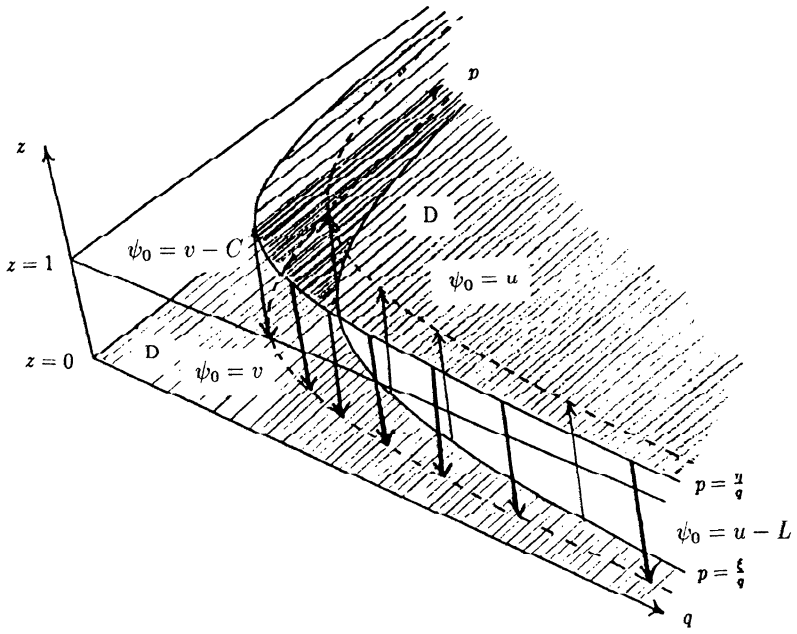


FIG. 2

However, as pointed out earlier this does not imply that $\phi_0 = \tilde{\phi}$, because it is not clear whether the solution of (2.13) is unique.

Remark. We have not been able to derive tractable general conditions for the existence of a solution of the system (4.14)–(4.17). Note, however, that there clearly exist parameter values such that the solution set is not empty. For example, if we choose $\xi > \eta > 0$ and solve (4.15) and (4.17) for k_1, k_2 (a linear system) and substitute these values in (4.14) and (4.16), then we can solve the other equations for, say, L and C if the remaining parameter values are given.

The strategy $w \in W$ corresponding to the candidate ϕ_0 is illustrated in Fig. 2, and can be described as follows (see §3):

(4.21) $$\begin{cases} \text{Jump from the } z = 0 \text{ level (closed field) to the } z = 1 \text{ level} \\ \text{(open field) as soon as } P_tQ_t \geq \xi \text{ and jump from } z = 1 \text{ to } z = 0 \\ \text{as soon as } P_tQ_t \leq \eta. \end{cases}$$

Put

$$H_0(z, \zeta) = \begin{cases} L & \text{if } z = 0, \zeta = 1 \\ C & \text{if } z = 1, \zeta = 0. \end{cases}$$

Then

$$H(s, u, z, \zeta) = e^{-\rho s} H_0(z, \zeta).$$

From (4.9) we see that

$$\psi_0(p, q, 0) = \psi_0(p, q, 1) - H_0(0, 1) \quad \text{if } p \geq \xi/q (\Rightarrow p \geq \eta/q)$$
$$\psi_0(p, q, 1) = \psi_0(p, q, 0) - H_0(1, 0) \quad \text{if } p \leq \eta/q (\Rightarrow p \leq \xi/q)$$

and for other values of (p, q, z) we have

$$\psi_0(p, q, z) > \psi_0(p, q, 1 - z) - H_0(z, 1 - z).$$

We conclude that with M defined as in (3.10),

$$\psi_0(p, q, z) \geq M\psi_0(p, q, z) \quad \text{for all } (p, q, z) \tag{4.22}$$

and if we define *the continuation region*

$$D = \{(p, q, z); \qquad z = 0 \text{ and } p < \xi/q \text{ or } z = 1 \text{ and } p > \eta/q\}, \tag{4.23}$$

then

$$(p, q, z) \in D \Leftrightarrow \psi_0(p, q, z) > M\psi_0(p, q, z). \tag{4.24}$$

It follows from [BØ, Lem. 1] that

$$\phi_0 \text{ is stochastically } C^2 \text{ with respect to } Y_t. \tag{4.25}$$

In this case the generator A of Y_t in (3.2) takes the form

$$Ag(t, p, q) = \frac{\partial g}{\partial t} + \alpha p \frac{\partial g}{\partial p} - \lambda q \frac{\partial g}{\partial q} + \frac{1}{2}\beta^2 p^2 \frac{\partial^2 g}{\partial p^2}. \tag{4.26}$$

In particular, if $g(s, p, q) = e^{-\rho s}h(p, q)$, then

$$Ag = e^{-\rho s}A_0 h, \tag{4.27}$$

where

$$A_0 h(p, q) = -\rho h + \alpha p \frac{\partial h}{\partial p} - \lambda q \frac{\partial h}{\partial q} + \frac{1}{2}\beta^2 p^2 \frac{\partial^2 h}{\partial p^2}. \tag{4.28}$$

We now claim that

$$A_0\psi_0 + f_0 = 0 \quad \text{in } D, \tag{4.29}$$

and interpreting A_0 in the almost everywhere $G(y, \cdot)$ sense (as with A above),

$$A_0\psi_0 + f_0 \leq 0 \quad \text{a.e. in } \mathbf{R}^+ \times \mathbf{R}^+ \times \{0, 1\}, \tag{4.30}$$

where $f_0 = (\lambda pq - K)z$, i.e., $f_0 = e^{\rho t}f$.

Remark. Let $D_1 = \{(p, q); (p, q, 1) \in D\}$ and $D_0 = \{(p, q); (p, q, 0) \in D\}$. Then by (4.24) we have $D_1 \cup D_0 = \mathbf{R}^+ \times \mathbf{R}^+$. So for all (p, q) we have from (4.29) that $A_0\psi_0(p, q) + f_0(p, q, z) = 0$ for some $z \in \{0, 1\}$. The requirement (4.30) can thus be written

$$f_0(p, q, z) \leq f_0(p, q, 1 - z) \quad \text{outside } D,$$

i.e., *if we are switching from state z, we must switch to a state with greater profit rate.*

Proof of (4.29). Recall that in [BØ] (formulas (75) and (76)) it is proved (and it is easily checked) that

$$A_0 v = 0 \quad \text{when } z = 0 \tag{4.31}$$

and

$$A_0 u = -f_0 \quad \text{when } z = 1. \tag{4.32}$$

From (4.31) and (4.9) we conclude that

$$A_0\psi_0 = A_0 v = 0(= -f_0) \quad \text{when } z = 0 \text{ and } p < \frac{\xi}{q}. \tag{4.33}$$

Similarly, from (4.32) and (4.9) we get

$$A_0\psi_0 = A_0 u = -f_0 \quad \text{when } z = 1 \text{ and } p > \frac{\eta}{q}. \tag{4.34}$$

Equation (4.29) follows from (4.33) and (4.34). □

Proof of (4.30). Equation (4.30) is a consequence of (4.20).

The general theory of optimal stopping (see e.g., [Ø]) gives that the right hand side of (4.20)—and hence ϕ_0 itself—is superharmonic with respect to the opertor $g \to A_0 g + f_0$. This implies that $A\phi_0 + f_0 \leq 0$ outside ∂D and hence almost everywhere with respect to $G(y, \cdot)$. □

Next we give a condition which ensures that f satisfies (3.5).

LEMMA 4.1. *Assume that*

$$\rho > \alpha. \tag{4.35}$$

Then

$$E^x\left[\int_0^\infty |f(X_t^{(w)})|dt\right] < \infty \quad \textit{for all } w \in W.$$

Proof. Since $Z_t \leq 1$ and $Q_t \leq Q_0$ for all t and all w it suffices to prove that

$$E^y\left[\int_0^\infty P_t e^{-\rho t} dt\right] < \infty \quad \text{for } y = (0, p, q, 1).$$

Now

$$P_t = p \cdot \exp\left(\left(\alpha - \frac{1}{2}\beta^2\right)t + \beta B_t\right).$$

Hence

$$E^y\left[\int_0^\infty P_t e^{-\rho t} dt\right] = \int_0^\infty \exp((\alpha - \rho)t)dt < \infty$$

since $\alpha - \rho < 0$. □

Finally we observe from (4.9) that the function $\phi_0 = e^{-\rho t}\psi_0$ satisfies (3.6). To verify (3.7) choose $\varepsilon > 0$ and consider

$$R_t = P_t^{1+\varepsilon} e^{-\rho(1+\varepsilon)t} = p^{1+\varepsilon} \cdot \exp\left(\left(\alpha - \frac{1}{2}\beta^2 - \rho\right)(1+\varepsilon)t + \beta(1+\varepsilon)B_t\right).$$

Note that

$$R_t = R_0 + \int_0^t \gamma R_s ds + \int_0^t \sigma R_s dB_s,$$

where

$$\sigma = \beta(1+\varepsilon)$$

and

$$\gamma = \left(\alpha - \frac{1}{2}\beta^2 - \rho\right)(1+\varepsilon) + \frac{1}{2}\beta^2(1+\varepsilon)^2 = (\alpha-\rho) + \varepsilon\left(\alpha + \frac{1}{2}\beta^2 - \rho + \frac{1}{2}\beta^2\varepsilon\right).$$

Choose $\varepsilon > 0$ so small that $\gamma < 0$ and let τ be a stopping time. For all natural numbers N we have

$$E^x[|\phi_0(\tau \wedge N, U_{\tau\wedge N}, z)|^{1+\varepsilon}] \leq q^{1+\varepsilon} E^x[R_{\tau\wedge N}]$$
$$= (pq)^{1+\varepsilon} + E^x\left[\int_0^{\tau\wedge N} \gamma R_s ds\right] \leq (pq)^{1+\varepsilon}.$$

Letting $N \to \infty$ we get

$$E^x[|\phi_0(\tau, U_\tau, z)|^{1+\varepsilon}] \leq (pq)^{1+\varepsilon}$$

for all stopping times τ. This implies (3.7).

Summing up we conclude the following.

THEOREM 4.2. *Assume that* (4.19) *and* (4.35) *hold. Then the function* $\phi_0 = e^{-\rho t}\psi_0$ *given by* (4.9) *solves the starting and stopping problem* (4.7).

The corresponding optimal impulse control $\hat{w}$ is given by (4.21).

Proof. By (4.35) f satisfies condition (3.5). The function $\phi = \phi_0$ satisfies conditions (3.6), (3.7), as well as (3.8), (3.12), and (3.17) in virtue of (4.30), (4.22), and (4.29), respectively. Therefore Theorem 3.4 applies to ϕ_0 and the proof is complete. □

Acknowledgments. We are grateful to Marcus Miller for useful discussions.

REFERENCES

[BL] A. BENSOUSSON AND J.-L. LIONS, *Impulse Control and Quasi-Variational Inequalities*, Gauthier-Villars, Paris, 1984.

[BØ] K. A. BREKKE AND B. ØKSENDAL, *The high contact principle as a sufficiency condition for optimal stopping*, in Stochastic Models and Option Values, D. Lund and B. Øksendal, eds., North-Holland, Amsterdam, 1991, pp. 187–208.

[BS] M. J. BRENNAN AND E. S. SCHWARTZ, *Evaluating natural resource investments*, J. Business, 58 (1985), pp. 135–157.

[D] A. DIXIT, *Entry and exit decisions under uncertainty*, J. Political Economy, 97 (1989), pp. 620–638.

[M] H. MORIMOTO, *Optimal switching for alternating processes*, Appl. Math. Optim., 16 (1987), pp. 1–17.

[MZ] M. MILLER AND L. ZHANG, *Irreversibility and oil production*. working paper, Department of Economics, University of Warwick, Warwick, United Kingdom.

[Ø] B. ØKSENDAL, *Stochastic Differential Equations*, Third ed., Springer-Verlag, New York, 1992.

[S] A. N. SHIRYAYEV, *Optimal Stopping Rules*, Springer-Verlag, New York, 1978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1037–1051, July 1994

009

L^∞-EXACT OBSERVABILITY OF THE HEAT EQUATION WITH SCANNING POINTWISE SENSOR*

ALEXANDER KHAPALOV†

Abstract. The problem of exact observability of the heat equation in an arbitrary space dimension with scanning pointwise sensor is considered in the case when the space for outputs is $L^\infty(\varepsilon,\theta)$, $\varepsilon > 0$. A new method for the construction of observation curves for sensors that are able to ensure $L^\infty(\varepsilon,\theta)$-exact observability at final time is given, based on the maximum principle for the heat equation. An application of the method to the observability problem with discrete-time scanning observations and related approximate controllability results are also discussed.

Key words. observability, controllability, the heat equation, scanning pointwise sensors

AMS subject classifications. primary 35K20, secondary 93B07

1. Introduction, statement of problem. Let Ω be a bounded domain of an n-dimensional Euclidean space R^n with a boundary $\partial\Omega$ of the class C^{2s_0+1}, where $2s_0+1 \geq [n/2]+3$. In Ω we consider the following homogeneous boundary problem:

$$\begin{gathered}\frac{\partial u(x,t)}{\partial t} = \Delta u(x,t),\\ t \in T = (0,\theta), \quad x \in \Omega \subset R^n, \quad Q = \Omega \times T, \quad \Sigma = \partial\Omega \times T,\\ u(x,t)|_\Sigma = 0, \qquad u(x,0) = u_0(x),\end{gathered} \tag{1.1}$$

with an unknown initial condition $u_0(\cdot) \in L^2(\Omega)$.

The main aim of the present paper is to study exact observability of (1.1) in the case when available observations are provided by a scanning pointwise sensor, namely,

$$y(t) = u(\hat{x}(t),t), \qquad t \in T, \tag{1.2}$$

where $y(\cdot)$ is a scalar output and $\hat{x}(t)$, $t \in T$ is an observation curve (measurable, in general) for the sensor, so that

$$\hat{x}(t) \in \bar{\Omega} \quad \text{a.e. in } T,$$

where $\bar{\Omega}$ denotes the closure of Ω. This type of observations requires a corresponding regularity of the solutions of (1.1), which is provided by the above assumption on $\partial\Omega$. Indeed, this assumption implies [16] (see §2 below for details) that all the solutions of system (1.1) are classical for $t > 0$, $x \in \bar{\Omega}$, and, in particular, the following enclosure is verified:

$$u(\hat{x}(\cdot),\cdot) \in L^\infty(T_\varepsilon), \qquad T_\varepsilon = (\varepsilon,\theta)$$

for any $\varepsilon \in (0,\theta)$ and for any solution of (1.1).

System (1.1), (1.2) is said to be *observable* if its initial state can be uniquely determined from the observation $y(\cdot)$ over the time interval $(0,\theta)$. System (1.1), (1.2) is said to be *B-exactly (or continuously) observable at final time* if

$$\exists \gamma > 0 \text{ such that } \| u(\hat{x}(\cdot),\cdot) \|_B \geq \gamma \| u(\cdot,\theta) \|_{L^2(\Omega)} \tag{1.3}$$

* Received by the editors March 11, 1992; accepted for publication (in revised form) December 23, 1992. The work of this author was supported in part by National Science Foundation grant ECS 89-13773.

† International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria. Present address, Department of Electrical and Computer Engineering, Oregon State University, Corvallis, Oregon 97331.

for any solution $u(x,t)$ of system (1.1), where B denotes the space for outputs.

Observability of the parabolic systems with stationary pointwise sensors (when $\hat{x}(t) \equiv \bar{x},\ t \in T$) was previously studied in a large number of publications on the basis of harmonic analysis (see [17], [3], [21], [22], [2], [5], and the literature therein). In [17] Mizel and Seidman established exact observability at time $t = t^*$ of the one-dimensional heat equation with boundary pointwise sensor, assuming that $\Omega = (0,1)$, $u_x(0,t) = u_x(1,t) = 0$ and that t^* is big enough. Dolecki [3] and Sakawa [21] obtained sufficient conditions for exact observability at final (again "big enough") time for the case of internal pointwise sensors. Employing the analytical continuation techniques, Sakawa also derived necessary and sufficient conditions for observability. It is well known now that, in the stationary setting, observability of the system in question depends crucially upon the growth and the multiplicities of the associated eigenvalues. In particular, from Müntz–Szász type theorems [23], [13], [6], it follows that, if $n > 1$, then the stationary scalar sensor is not able to ensure exact observability of system (1.1), (1.2) at final time for $B = C[0,\theta]$, $L^p(T)$, $p \geq 1$. Furthermore, in the known examples [3], [21], [2], [5] of *observable* parabolic systems with stationary observations, the dimension of x does not exceed 2. On the other hand, in physical situations, available outputs (measurement data) are generally finite-dimensional at every moment of time. In the present paper, we show how this difficulty can be overcome by the introduction of moving pointwise sensors.

Remark 1.1. The problem of the choice of trajectories for moving sensors and actuators (as the optimal sensor and actuator allocation problem) arises in a natural way in the context of state estimation and control of distributed parameter systems (see [15], [10], [1], [9], and the literature therein). To our knowledge, observability of distributed-parameter systems with scanning pointwise sensors was previously studied in [3], [14]. In [3], for the sensor moving along the line $\hat{x}(t) = at+b$, Dolecki established (making use of the explicit representation of eigenfunctions) exact observability of the one-dimensional heat equation. In [14] Martin showed how the results of [21] can be reformulated in the form of moving sensors. Both papers reduced, in fact, the problem in question to the traditional stationary setting and did not employ the advantages of the motion itself.

The main result of this paper is a new method for the construction of observation curves for sensors that are able to provide the system with $L^\infty(\varepsilon,\theta)$-exact observability at final time for any $\varepsilon \in (0,\theta)$, given in advance (it is clear that if system (1.1), (1.2) is (exactly) observable at final time on (ε,θ), it is (exactly) observable on any interval (t^*,θ), $t^* \in (0,\varepsilon)$). (The approach developed in this paper was announced in the paper presented by the author at the IV International Conference on Control of Distributed Parameter Systems, held at Vorau, Austria, 1988.) In [11] it was used to analyse the observability problem arising in the framework of the theory of guaranteed estimation of parabolic systems under uncertainty. A more general abstract scheme of the proof of existence results, given in the above-mentioned conference paper and in [11], employed the separability of the space $C(\bar{\Omega} \times [\varepsilon,\theta])$ containing all the solutions of (1.1) taken on the interval $[\varepsilon,\theta]$.

The method proposed in this paper is constructive and makes use of the general representation of the solutions of (1.1) in the form of the generalized Fourier expansion along the eigenfunctions as well as of the separability of the space $C(\bar{\Omega})$. To some extent, it can be treated as an analogue of Galerkin's method coupled with the general a priori estimates techniques when being applied in the framework of the observability theory. From this point of view, the proposed method is rather general. Indeed,

it employes *a priori estimates* of instantaneous type for solutions (in the present paper, we use the maximum principle for the heat equation). The latter allows us to extend the results of the paper to a number of systems (including time-varying), admitting similar a priori estimates, when, instead of the eigenfunctions, an arbitrary appropriate basis—as in Galerkin's scheme—can be used. We state the proposed method in the form of an abstract algorithm. Each iteration of the algorithm can be associated with some part of the observation time-interval and provides $L^\infty(T_\varepsilon)$-exact observability at final time in the corresponding finite-dimensional subspace of $L^2(\Omega)$, spanned by the eigenfunctions of (1.1). Note that we deal with the case when the space for outputs $(L^\infty(T_\varepsilon))$ is not Hilbert. Therefore, we construct a countable net (which can be specified in infinitely many ways) in the pair of the sets of solutions and of outputs. We then establish the linkage between this net and appropriate countable sets of pairs $\{x^k, t_k\}_{k=1}^\infty$ that are to form the "skeletons" for required observation curves. Although a countable number of points in $\bar{\Omega} \times T$ is involved, we recall that this is to recover $u(\cdot, \theta)$, which is an element of the Hilbert space (see also Remarks 4.1 and 4.2(ii) in §4).

The main *existence* result of the paper, associated with the proposed method, can be formulated as follows.

THEOREM 1.1. *Given $\varepsilon \in T$, an arbitrary curve $\hat{x}(\cdot)$ constructed along Procedure* 3.1 (*described in* §3) *ensures $L^\infty(\varepsilon, \theta)$-exact observability of system* (1.1), (1.2) *at final time.*

The paper is organized as follows. Section 2 deals with some preliminary results. In §3 we introduce a scheme for the construction of *continuous* observation curves that are able to ensure required observability at final time in any finite-dimensional subspace (specified in advance), spanned by the eigenfunctions of (1.1). Then Procedure 3.1, describing the general algorithm for the construction of observation curves, is given. Theorem 1.1 is proved in §4. In the same section, we discuss a number of important corollaries and consider the specific case of the one-dimensional heat equation. Section 5 deals with an application of the method to the discrete-time observability problem with scanning sensor that is of practical importance (see also Remark 1.2). For the stationary sensors and infinite time horizon, a similar problem was studied in [8], where the authors used the results of Sakawa [21]. Section 6 deals with approximate controllability results related by duality with the main results of the paper. We restrict ourselves here by the case when $n \leq 3$ and establish approximate controllability of the heat equation with scanning pointwise control in $H^{-1}(\Omega)$.

Remark 1.2. The present stage of technology provides a growing number of examples of observations, where the modelling, based on a scanning sensor approach, might be of practical interest: the sensor (and actuator) allocation problem and an associated estimation (and control) policy for advanced composite materials (so-called "smart or intelligent" structures; see, for example, [19] and the bibliography therein); measurements of surface temperature by optical pyrometers and measurements of vibration and strain in materials using optical registrations; remote sensing of atmospheric species from a ground-, aircraft-, or satellite-based platform [18].

2. Preliminaries. This section deals with several auxiliary results that are employed below. It is well known that the general solution of problem (1.1) may be represented in the form

$$u(x,t) = \sum_{i=1}^{\infty} e^{-\lambda_i t} \langle u_0(\cdot),\, \omega_i(\cdot) \rangle \omega_i(x), \tag{2.1}$$

where

$$\langle u_0(\cdot),\, \omega_i(\cdot)\rangle \;=\; \int_\Omega u_0(x)\omega_i(x)dx$$

and $\lambda_i,\ \omega_i(\cdot)\ (i=1,2,\ldots)$ denote the eigenvalues and the eigenfunctions, orthonormalized in $L_2(\Omega)$, of the spectral problem

$$\Delta\omega_i(x) = -\lambda_i\omega_i(x),$$

so that

$$\lambda_{i+1} \geq \lambda_i > 0, \quad \lambda_i \to +\infty, \quad i \to +\infty.$$

In [16] it was shown that, if $\partial\Omega \in C^{2s}$ and $u_0(\cdot) \in H_D^{2s-1}(\Omega)$, then the generalized solution of (1.1) belongs to $H^{2s,s}(Q)$ and the following estimate is verified:

$$\| u(\cdot,\cdot) \|_{H^{2s,s}(Q)} \leq \text{const}\ \| u_0(\cdot) \|_{H^{2s-1}(\Omega)}, \tag{2.2}$$

where

$$H_D^r(\Omega) = \{\phi \mid \phi \in H^r(\Omega),\ \phi\,|_{\partial\Omega} = \cdots = \Delta^{[(r-1)/2]}\phi\,|_{\partial\Omega} = 0\}.$$

Recall further that the norm

$$\{v,v\}^{1/2} \;=\; \left(\sum_{k=1}^{\infty}\lambda_k^r v_k^2\right)^{1/2}, \quad v_k = \int_\Omega v(x)\,\omega_k(x)\,dx$$

is equivalent to the standard one in $H_D^r(\Omega)$. Due to the smoothing effect, this implies, in particular (as was mentioned in §1), that all the solutions of (1.1) are classical on $\bar\Omega \times [\varepsilon,\theta]$ (see [16] for details).

It is well known [7], [16] that any solution of system (1.1) satisfies the *maximum principle*

$$\max_{x\in\bar\Omega} \mid u(x,t') \mid \geq \max_{x\in\bar\Omega} \mid u(x,t'') \mid, \qquad 0 < t' \leq t'', \tag{2.3}$$

which lies in the basis of the method presented in the next section.

3. $C[\varepsilon,\theta]$-exact observability in finite dimensions. Denote by $L^2_{(k)}(\Omega)$ the finite-dimensional subspace of $L^2(\Omega)$ spanned by the functions

$$\omega_i(\cdot), \qquad i = 1,2,\ldots,k.$$

It is clear that, if $u(x,0)$ belongs to $L^2_{(k)}(\Omega)$, then $u(\cdot,t) \in L^2_{(k)}(\Omega)$, for all $t > 0$.

In this section, we discuss the problem of exact observability with respect to the sequence of the above subspaces. We show that required observation curves can be found among the continuous ones.

Let $\{x^j\}_{j=1}^\infty$ be a sequence of spatial points in $\bar\Omega$ and let $\{t_j\}_{j=1}^\infty$ be a sequence of instants of time in T. We say that an observation curve $\hat x(\cdot)$ has a *skeleton* $\{x^j,t_j\}_{j=1}^J$, where J can be both finite and infinite if it is continuous at all the instants $\{t_j\}_{j=1}^J$ and satisfies the following condition:

$$\hat x(t_j) = x^j, \qquad j = 1,\ldots,J.$$

LEMMA 3.1. *Given $\varepsilon \in (0,\theta)$ and a positive integer k, there exist continuous observation curves that ensure $C[\varepsilon,\theta]$-exact observability of* (1.1), (1.2) *at final time in $L^2_{(k)}(\Omega)$. To specify a required curve, it suffices to determine an appropriate* finite *skeleton.*

Proof of Lemma 3.1. Denote by $Y_{\varepsilon k}$ the set of all the possible outputs (1.2), generated by the initial conditions from $L^2_{(k)}(\Omega)$ and taken on T_ε. Observe that (1.1), (1.2) is $C[\varepsilon,\theta]$-exactly observable in $L^2_{(k)}(\Omega)$ if and only if the mapping

$$\mathbf{P}:\ C[\varepsilon,\theta] \supset Y_{\varepsilon k} \ \to\ L^2_{(k)}(\Omega), \qquad \mathbf{P}u(\hat{x}(\cdot),\cdot) = u(\cdot,\theta)$$

exists and is bounded, so that $Y_{\varepsilon k}$ is a subspace of $C[\varepsilon,\theta]$ and

$$\| \mathbf{P} \| = \sup\{\| \mathbf{P}u(\hat{x}(\cdot),\cdot) \| \mid u(\hat{x}(\cdot),\cdot) \in Y_{\varepsilon k},\ \| u(\hat{x}(\cdot),\cdot) \|_{C([\varepsilon,\theta]} \le 1\} < \infty.$$

Thus, to prove that $\hat{x}(\cdot)$ satisfies Lemma 3.1, it suffices to show that the preimage of the set

$$\{u(\hat{x}(\cdot),\cdot) \mid u(\hat{x}(\cdot),\cdot) \in Y_{\varepsilon k},\ \| u(\hat{x}(\cdot),\cdot) \|_{C[\varepsilon,\theta]} \le 1\}$$

for the mapping $u(\cdot,\theta) \to u(\hat{x}(\cdot),\cdot)$ is bounded in $L^2(\Omega)$.

Select in the interval T_ε an arbitrary monotone sequence of instants

$$\varepsilon = t_0 < t_1 < t_2 < \cdots < t_k < t_{k+1} < \cdots < \theta$$

and denote $\tau_k = (t_{k-1}, t_k)$, $k = 1, \ldots$.

Step 1. Consider first the case when $u(x,t)$ is generated by $u(\cdot,0) \in L^2_{(1)}(\Omega)$. Then, due to (2.1),

$$u(x,t) = e^{-\lambda_1 t} u_{01} \omega_1(x),$$

where

$$u_{01} = \int_\Omega u(x,0)\, \omega_1(x)\, dx.$$

Let $x^1_{(1)}$ be an arbitrary solution of the following optimization problem:

$$| \omega_1(x) | \to \max, \qquad x \in \bar{\Omega},$$

so that

$$| \omega_1(x^1_{(1)}) | = \max_{x \in \bar{\Omega}} | \omega_1(x) | . \tag{3.1}$$

This problem may admit, in general, several solutions. We take any of them.

Select any instant t^1_1 from τ_1 and consider an arbitrary continuous curve $\hat{x}(t)$, $t \in [0,\theta]$ that passes through the point $x^1_{(1)}$ at $t = t^1_1$, so that $\hat{x}(t^1_1) = x^1_{(1)}$. Then, if $u(x,t)$ is such that

$$\| u(\hat{x}(\cdot),\cdot) \|_{C[\varepsilon,\theta]} = \max\{| u(\hat{x}(t),t) | \mid t \in [\varepsilon,\theta]\} \le 1,$$

we obtain the following estimate:

$$| u_{01} | \le e^{+\lambda_1 t^1_1} | \omega_1^{-1}(x^1_{(1)}) | .$$

It is not hard to see that $\mid \omega_1^{-1}(x_{(1)}^1) \mid \neq 0$; otherwise, due to (3.1), $\omega_1(x) \equiv 0$. The last estimate yields

$$\| u(\cdot,\theta) \|_{L^2(\Omega)} \leq e^{+\lambda_1(t_1^1-\theta)}(\text{meas}\,\{\Omega\})^{1/2},$$

since

$$1 = \| \omega_1(\cdot) \|^2_{L^2(\Omega)} \leq \max\{\mid \omega_1(x) \mid^2 \mid x \in \bar{\Omega}\}\ \text{meas}\,\{\Omega\}.$$

Thus, we obtain the conclusion of Lemma 3.1 for system (1.1), (1.2) in the subspace $L^2_{(1)}(\Omega)$ with $\gamma = e^{-\lambda_1(t_1^1-\theta)}(\text{meas}\,\{\Omega\})^{-1/2}$.

Step 2. Let us proceed now with the general case. Denote by Φ_k the set

$$\Phi_k = \{v(\cdot) \mid \| v(\cdot) \|_{L^2(\Omega)} = 1, \quad v(\cdot) \in L^2_{(k)}(\Omega)\}.$$

We note next that Φ_k is also a bounded finite-dimensional subset of $C(\bar{\Omega})$. Therefore, for any positive δ, we can specify in it a finite δ-net

$$\Phi_k^\delta = \{v_k^j(\cdot)\}_{j=1}^{J_k}, \qquad v_k^j(\cdot) \in \Phi_k,$$

where J_k depends upon δ, so that, for any element $v(\cdot) \in \Phi_k$, there exists a positive integer $j = j_* \leq J_k$ such that $\| v(\cdot) - v_k^{j_*}(\cdot) \|_{C(\bar{\Omega})} \leq \delta$. The maximum principle (2.3) (applied for the set Φ_k) allows us to transform Φ_k^δ (due to finite dimension of Φ_k, (2.3) can be extended to $[0,\theta]$) into the δ-net $\Phi_k^\delta(\cdot)$ in the space $C(\bar{\Omega} \times [0,\theta])$,

$$\Phi_k^\delta(\cdot) = \{u_k^j(x,t),\ u_k^j(x,0) = v_k^j(x)\}_{j=1}^{J_k}, \qquad x \in \bar{\Omega}, \quad t \in [0,\theta]$$

for the set of all those solutions $u(x,t)$ that are generated at instant $t = 0$ by initial conditions from Φ_k.

Take any β in the interval $(0,1)$ and select

$$\delta = \delta_k(\beta) = \beta\, \tfrac{1}{2}(\text{meas}\{\Omega\})^{-1/2} e^{-\lambda_k t_k}, \tag{3.2}$$

denoting accordingly $J_k = J_k(\beta)$.

Selecting in τ_k an arbitrary sequence of instants of time t_k^j, $j = 1,2,\ldots,J_k(\beta)$, so that

$$t_{k-1} < t_k^1 < t_k^2 < \cdots < t_k^{J_k(\beta)} < t_k,$$

we introduce the following series of optimization problems for $j = 1,\ldots,J_k(\beta)$.

Problem (k,j). Find $x_{(k)}^j$ in such a way that

$$\max_{x \in \bar{\Omega}} \mid u_k^j(x,t_k^j) \mid = \mid u_k^j(x_{(k)}^j, t_k^j) \mid . \tag{3.3}$$

In general, Problem (k,j) admits nonunique solutions, but, if so, we may take any of them.

Let $\hat{x}(t)$, $t \in [0,\theta]$, be an arbitrary continuous curve in $\bar{\Omega}$ that has the skeleton $\{x_{(k)}^j, t_k^j\}_{j=1}^{J_k(\beta)}$. We show that it satisfies the necessary requirements.

Take any solution of (1.1) such that $u(\cdot,t) \in L^2_{(k)}(\Omega)$, for all $t \in T$,

$$u(x,t) = \sum_{i=1}^{k} e^{-\lambda_i t} < u(\cdot,0),\ \omega_i(\cdot) > \omega_i(x) \tag{3.4}$$

and assume that

$$\| u(\hat{x}(\cdot),\cdot) \|_{C[\varepsilon,\theta]} = \max\{| u(\hat{x}(t),t) | \mid t \in [\varepsilon,\theta]\} \leq 1. \tag{3.5}$$

This yields, in particular,

$$| u(\hat{x}(t_k^j), t_k^j) | \leq 1, \qquad j = 1,\dots,J_k(\beta). \tag{3.6}$$

Denote by α the value of $L^2(\Omega)$-norm of the function $u(\cdot,\cdot)$ taken at $t = 0$,

$$\alpha = \| u(\cdot,0) \|_{L^2(\Omega)}. \tag{3.7}$$

Without loss of generality, we may assume that $\alpha \neq 0$.

Select an element $u_k^{j*}(\cdot,\cdot) \in \Phi_k^{\delta_k(\beta)}(\cdot)$, such that

$$| \alpha^{-1}u(x,t) - u_k^{j*}(x,t) | \leq \delta_k(\beta) \quad \text{for all } x \in \bar{\Omega},\ t \in [0,\theta]. \tag{3.8}$$

Hence, in particular,

$$\| u(\cdot,t_k^{j*}) \|_{C(\bar{\Omega})} \leq \alpha \| u_k^{j*}(\cdot,t_k^{j*}) \|_{C(\bar{\Omega})} + \alpha\delta_k(\beta). \tag{3.9}$$

On the other hand, making use of (3.3) and again (3.8), we obtain

$$\alpha \| u_k^{j*}(\cdot,t_k^{j*}) \|_{C(\bar{\Omega})} = | \alpha u_k^{j*}(x_{(k)}^{j*}, t_k^{j*}) | \leq | u(\hat{x}(t_k^{j*}), t_k^{j*}) | + \alpha\delta_k(\beta). \tag{3.10}$$

Combining (3.5), (3.9), and (3.10) yields

$$\| u(\cdot,t_k^{j*}) \|_{C(\bar{\Omega})} \leq 1 + 2\alpha\delta_k(\beta). \tag{3.11}$$

Observe that, due to (3.4), (3.7),

$$\begin{aligned} \alpha^2 e^{-2\lambda_k t_k^{j*}} &= e^{-2\lambda_k t_k^{j*}} \textstyle\sum_{i=1}^k \langle u(\cdot,0),\, \omega_i(\cdot)\rangle^2 \\ &\leq \sum_{i=1}^k e^{-2\lambda_i t_k^{j*}} \langle u(\cdot,0),\, \omega_i(\cdot)\rangle^2 \\ &= \int_\Omega u^2(x,t_k^{j*})dx \leq \operatorname{meas}\{\Omega\} \| u(\cdot,t_k^{j*}) \|^2_{C(\bar{\Omega})}. \end{aligned} \tag{3.12}$$

Combining (3.11) and (3.12), we arrive at

$$\begin{aligned} &(\operatorname{meas}\{\Omega\})^{-1/2} \left(\sum_{i=1}^k e^{-2\lambda_i t_k^{j*}} \langle u(\cdot,0),\, \omega_i(\cdot)\rangle^2 \right)^{1/2} \\ &- 2\delta_k(\beta) \left(\sum_{i=1}^k \langle u(\cdot,0),\, \omega_i(\cdot)\rangle^2 \right)^{1/2} \leq 1. \end{aligned} \tag{3.13}$$

Taking into account the inequality

$$e^{-\lambda_k t_k} \left(\sum_{i=1}^k \langle u(\cdot,0),\, \omega_i(\cdot)\rangle^2 \right)^{1/2} \leq \left(\sum_{i=1}^k e^{-2\lambda_i t_k^{j*}} \langle u(\cdot,0),\, \omega_i(\cdot)\rangle^2 \right)^{1/2}$$

and (3.13), (3.2), we finally obtain

$$\begin{aligned} \| u(\cdot,\theta) \|_{L^2(\Omega)} &\le \| u(\cdot,t_k) \|_{L^2(\Omega)} \\ &\le \left(\sum_{i=1}^{k} e^{-2\lambda_i t_k^{j*}} \langle u(\cdot,0),\, \omega_i(\cdot) \rangle^2 \right)^{1/2} \\ &\le (\text{meas } \{\Omega\})^{1/2} \frac{1}{1-\beta}. \end{aligned} \tag{3.14}$$

Estimate (3.14), coupled with (3.5), implies the existence and boundedness of $\mathbf{P}$ on $Y_{\varepsilon k} \subset C[\varepsilon,\theta]$. In turn, this provides $C[\varepsilon,\theta]$-exact observability of system (1.1), (1.2) at final time in $L^2_{(k)}(\Omega)$ with the constant

$$\gamma = (\text{meas } \{\Omega\})^{-1/2}(1-\beta) \;\le\; \| \mathbf{P} \|^{-1}, \tag{3.15}$$

so that

$$(\text{meas } \{\Omega\})^{-1/2}(1-\beta) \| u(\cdot,\theta) \|_{L^2(\Omega)} \le \max_{t\in[\varepsilon,\theta]} | u(\hat{x}(t),t) | .$$

This completes the proof of Lemma 3.1.

The following procedure describes the algorithm for the construction of the skeletons for the observation curves ensuring $C[\varepsilon,\theta]$-exact observability of (1.1), (1.2) at final time in all $L^2_{(k)}(\Omega)$.

Procedure 3.1. Let $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$ be given.

1. Select an arbitrary monotone sequence of instants of time $\{t_k\}_{k=1}^{\infty} \subset T_\varepsilon$. It is clear that there exists a limit $\lim_{k\to\infty} t_k = \hat{t} \le \theta$.

2. Determine the values $\delta_k = \delta_k(\beta)$, $k = 1, 2 \ldots$, from (3.2).

3. Given k (we can start from any positive integer k), find a $\delta_k(\beta)$-net $\Phi_k^{\delta_k(\beta)}(\cdot)$. The latter provides the value $J_k = J_k(\beta)$.

4. Selecting an arbitrary monotone sequence $\{t_k^j\}_{j=1}^{J_k(\beta)}$ in the interval $\tau_k = (t_{k-1}, t_k)$, $t_0 = \varepsilon$, find from (3.3) a respective sequence of spatial points $\{x_{(k)}^j\}_{j=1}^{J_k(\beta)}$.

5. Repeat steps 3 and 4 with $k+1$ instead of k.

Let $\varepsilon \in T$ and $\beta \in (0,1)$ be given. We say that an observation curve $\hat{x}(t)$, $t \in [0,\theta]$ is constructed along Procedure 3.1 if it has a skeleton

$$\{x_{(k)}^j, t_k^j\}, \quad k = 1, 2, \ldots, \quad j = 1, \ldots, J_k(\beta),$$

specified according to Procedure 3.1, and $\hat{x}(\cdot)$ is continuous on $[\varepsilon, \hat{t})$ (see step 1 of Procedure 3.1).

From Lemma 3.1, we immediately obtain the following assertion.

COROLLARY 3.1. *Given $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$, let $\hat{x}(\cdot)$ be an arbitrary observation curve constructed along Procedure* 3.1. *Then, for any $k = 1, \ldots,$ to ensure the estimate*

$$(\text{meas } \{\Omega\})^{-1/2}(1-\beta) \| u(\cdot,\theta) \|_{L^2(\Omega)} \le \| u(\hat{x}(\cdot),\cdot) \|_{L^\infty(T_\varepsilon)}$$

for the solutions of (1.1) *with initial data from $L^2_{(k)}(\Omega)$, it suffices to take into account the observations* (1.2) *only over the time interval $\tau_k = (t_{k-1}, t_k)$.*

4. Proof of Theorem 1.1 and discussion of main results. For any given $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$, we show that an arbitrary curve $\hat{x}(\cdot)$ constructed along Procedure 3.1 satisfies the necessary requirements.

We recall first that estimate (3.14) is uniform over $k = 1, \ldots$. Denote by Y_ε the set of all the possible outputs (1.2) taken on T_ε. Observe that (1.1), (1.2) is $L^\infty(T_\varepsilon)$-exact observable if and only if the mapping

$$\mathbf{P}: \; L^\infty(T_\varepsilon) \supset Y_\varepsilon \; \to \; L^2(\Omega), \qquad \mathbf{P}u(\hat{x}(\cdot),\cdot) = u(\cdot,\theta)$$

exists and is bounded (so that the domain of $\mathbf{P}$ may be extended to $\bar{Y}_\varepsilon$),

$$\| \mathbf{P} \| = \sup\{\| \mathbf{P}u(\hat{x}(\cdot),\cdot) \| \mid \; u(\hat{x}(\cdot),\cdot) \in Y_\varepsilon, \; \| u(\hat{x}(\cdot),\cdot) \|_{L^\infty(T_\varepsilon)} \le 1\} \; < \; \infty.$$

To prove Theorem 1.1, it suffices to show that the preimage of the set

$$\{u(\hat{x}(\cdot),\cdot) \mid u(\hat{x}(\cdot),\cdot) \in Y_\varepsilon, \; \| u(\hat{x}(\cdot),\cdot) \|_{L^\infty(T_\varepsilon)} \le 1\}$$

for the mapping $u(\cdot,\theta) \to u(\hat{x}(\cdot),\cdot)$ is bounded in $L^2(\Omega)$.

Take any positive μ. Let $u(x,t)$ be an arbitrary solution of system (1.1) (e.g., (2.1) is fulfilled) such that

$$\| u(\hat{x}(\cdot),\cdot) \|_{L^\infty(T_\varepsilon)} \le 1. \tag{4.1}$$

Split then the sum on the right-hand side of (2.1) into two parts,

$$u(x,t) = u_N(x,t) + v_N(x,t),$$

where

$$u_N(x,t) = \sum_{i=1}^{N} e^{-\lambda_i t} < u(\cdot,0), \, \omega_i(\cdot) > \omega_i(x),$$

$$v_N(x,t) = \sum_{i=N+1}^{\infty} e^{-\lambda_i t} < u(\cdot,0), \, \omega_i(\cdot) > \omega_i(x),$$

in such a way that

$$\| v_N(\cdot,\theta) \|_{L^2(\Omega)} \le \mu, \tag{4.2}$$

and, in addition,

$$\| v_N(\cdot,\cdot) \|_{C(\bar{\Omega}\times[\varepsilon,\theta])} \le \mu.$$

The latter and (4.1) imply that

$$| u_N(\hat{x}(t),t) | \le 1 + \mu \quad \forall t \in \tau_N,$$

where τ_N is defined in step 4 of Procedure 3.1. Applying estimate (3.14) with $k = N$ and with $1 + \mu$ instead of 1 yields

$$\| u_N(\cdot,\theta) \|_{L^2(\Omega)} \le (\text{meas}\,\{\Omega\})^{1/2} \frac{1+\mu}{1-\beta}.$$

Finally, combining (4.2) and the last estimate, we arrive at

$$\| u(\cdot,\theta) \|_{L^2(\Omega)} \leq (\text{meas}\,\{\Omega\})^{1/2}\frac{1+\mu}{1-\beta} + \mu. \tag{4.3}$$

This justifies the existence and boundedness of $\mathbf{P}$ on $Y_\varepsilon \subset L^\infty(T_\varepsilon)$. Recalling that $\mu(>0)$ was selected in arbitrary way, we obtain

$$\| \mathbf{P} \| \leq (\text{meas}\,\{\Omega\})^{1/2}\frac{1+\mu}{1-\beta} + \mu \quad \forall \mu > 0.$$

This allows us to conclude that the needed inequality (1.3) holds with the same γ as in Lemma 3.1.

COROLLARY 4.1. *Given $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$, an arbitrary observation curve $\hat{x}(\cdot)$ constructed along Procedure* 3.1 *makes system* (1.1), (1.2) *be $L^\infty(T_\varepsilon)$-exactly observable at final time with the constant γ from* (3.15).

COROLLARY 4.2. *Given $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$, let $\hat{x}(\cdot)$ be constructed along Procedure* 3.1. *Then*

$$\begin{aligned} &\| u(\hat{x}(\cdot),\cdot) \|_{L^\infty(T_\varepsilon)} \\ &\geq (\text{meas}\,\{\Omega\})^{-1/2}(1-\beta)\,|< u(\cdot,0),\, \omega_i(\cdot) >|\; e^{-\lambda_i\theta}, \qquad i = 1,\dots, \end{aligned} \tag{4.4}$$

for any solution $u(x,t)$ of system (1.1).

Proof of this assertion follows from representation (2.1) and estimate (4.3) (or (1.3)).

Corollary 4.2 implies a "weaker" observability property.

COROLLARY 4.3. *Given $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$, an arbitrary curve $\hat{x}(\cdot)$ constructed along Procedure* 3.1 *provides system* (1.1), (1.2) *with observability.*

Remark 4.1. To construct a curve according to Procedure 3.1, we must determine a countable number of pairs forming its skeleton, although, for its approximation (in the sense of Lemma 3.1), we may restrict ourselves by a finite skeleton. The situation here, to some extent, is similar to the well-known example [3], [21], [2], [5] of the one-dimensional heat equation with stationary pointwise observations when, to obtain observability, we must locate a single sensor at *irrational point* (if $\Omega = (0,1)$). In fact, Procedure 3.1 requires the same *countable* "amount of information" as an irrational point (see also Remark 4.2(ii)), but an appropriate distribution of this information in time and space allows us to solve the problem of $L^\infty(T_\varepsilon)$-exact observability at final time for any space dimension.

Consider a sequence of functions

$$\psi_i(t) = e^{-\lambda_i t}\omega_i(\hat{x}(t)), \qquad t \in T, \quad i = 1,2,\dots,$$

and denote by $L^\infty_{(i)}(T_\varepsilon)$ the subspace of $L^\infty(T_\varepsilon)$ spanned by the functions

$$\psi_j(\cdot), \qquad j = 1,\dots; \quad j \neq i.$$

Set

$$d_i = \inf_{\psi(\cdot)\in L^\infty_{(i)}(T_\varepsilon)} \| \psi_i(\cdot) - \psi(\cdot) \|_{L^\infty(T_\varepsilon)}, \qquad i = 1,2,\dots.$$

Properties of exponentials $\{e^{-\lambda_i t}\}_{i=1}^\infty$ play a crucial role in the study of observability and controllability of the parabolic systems [17], [3], [6], [20]–[22], [2], [5] in the

stationary setting of problem. It is well known [23], [13], [6] that, if the dimension of x is higher than 1 and $\hat{x}(\cdot) \equiv \bar{x}$, then

$$d_i = 0, \qquad i = 1, \dots,$$

not only with respect to $L^\infty(T_\varepsilon)$-norm but also in $C[0,\theta]$, $L^p(T)$, $p \geq 1$. From estimate (4.4), we obtain the following assertion.

COROLLARY 4.4 (minimality in $L^\infty(T_\varepsilon)$). *If $\hat{x}(\cdot)$ is constructed along Procedure 3.1, then*

$$d_i = \inf_{\psi(\cdot) \in L^\infty_{(i)}(T_\varepsilon)} \| \psi_i(\cdot) - \psi(\cdot) \|_{L^\infty(T_\varepsilon)} \geq (\operatorname{meas} \Omega)^{-1/2}(1-\beta)e^{-\lambda_i \theta}, \quad i = 1, 2, \dots, \tag{4.5}$$

regardless of the system's space dimension.

At first, estimates (4.5) appear unexpected but, to explain them, we recall that Corollary 4.2 has been derived on the basis of the maximum principle for the heat equation, which is not affected by the multiplicity of eigenvalues and their growth. From this point of view, the proposed techniques rather employ the eigenfunctions that are always distinct.

Remark 4.2. Procedure 3.1 leaves unanswered the question of the geometry of observation curves providing exact observability. However, a few comments can be made here.

(i) If in Procedure 3.1 we set $\hat{t} = \theta$ (see step 1 of the algorithm), we may construct required curves to be continuous in $[0,\theta)$.

(ii) After a slight modification of Procedure 3.1, namely, if instead of precise solutions in (3.3) we take their approximations (this may affect the value of γ in (3.15)), then all the spatial points forming skeletons may be selected to be with *rational* coordinates.

At the end of this section, we consider the one-dimensional heat equation and give a specific example of observation curves that can solve the observability problem.

Example. Consider the following initial-boundary value problem:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \quad 0 < x < 1, \quad t \in T, \tag{4.6}$$

$$u(t,0) = u(t,1) = 0, \qquad u(x,0) = u_0(x),$$

with stationary pointwise observations

$$y(t) = u(\bar{x}, t), \quad t \in T. \tag{4.7}$$

It is well known [3], [21], [2], [5] that (4.6), (4.7) is observable if and only if the point $\bar{x}$ is irrational, and it is $L^2(T)$-exactly observable if $\bar{x}$ is an irrational number of special type.

Let $\hat{x}(t)$, $t \in [0,\theta]$ be an arbitrary continuous curve connecting the ends of the interval [0,1] in the following way:

$$\exists\, t_1,\ t_2 \in [\varepsilon, \theta] \text{ such that } \hat{x}(t_1) = 0,\ \hat{x}(t_2) = 1. \tag{4.8}$$

Then, applying the maximum principle for the solutions of system (4.6) in the domain

$$D = \{(x,t) \mid 0 \leq x \leq \hat{x}(t),\ t \in [t_1, t_2]\}$$

yields the estimate

$$\| u(\cdot,\theta) \|_{C[0,1]} \leq \| u(\hat{x}(\cdot),\cdot) \|_{C[t_1,t_2]}, \tag{4.9}$$

which ensures exact observability at final time of system (4.6), (4.7) with $B = C[\varepsilon,\theta]$.

The attractive property of scanning sensor in this example (in contrast to stationary one) is that estimate (4.9) is the same for any continuous curve $\hat{x}(\cdot)$ satisfying (4.8). This makes such a class of curves be stable with respect to those perturbations that leave the perturbed curves in the mentioned class.

5. Observability with discrete-time scanning sensors. Consider system (1.1) with the discrete-time observations

$$y_i = u(x^i, t_i), \quad i = 1, 2, \ldots . \tag{5.1}$$

Here $\{y_i)\}_{i=1}^{\infty}$ is measurement data, $\{t_i\}_{i=1}^{\infty} \subset T$ is a monotone sequence of measurement instants; the spatial points $\{x^i\}_{i=1}^{\infty} \subset \bar{\Omega}$ specify the location of scanning sensor at the instants t_i, $i = 1, \ldots$.

We assume that the space for outputs $\mathbf{y} = \{y_1, \ldots, y_i, \ldots\}$ in this section is l^{∞} with the norm

$$\| \mathbf{y} \|_{l^{\infty}} = \sup_{i=1,\ldots} | y_i | .$$

Inequality (1.3) in the definition of B-exact observability at final time, adjusted for system (1.1), (5.1), may be represented as follows:

$$\sup_{i=1,\ldots} | u(x^i, t_i) | \geq \gamma \| u(\cdot,\theta) \|_{L^2(\Omega)} . \tag{5.2}$$

LEMMA 5.1. *Given $\varepsilon \in T$ and $\beta \in (0,1)$, let the sequence of pairs $\{x^i, t_i\}_{i=1}^{\infty}$ in* (5.1) *be selected according to Procedure* 3.1. *Then system* (1.1), (5.1) *is l^{∞}-exactly observable at final time, and estimate* (5.2) *holds with the constant γ from* (3.15).

The proof of Lemma 5.1 immediately follows from the proof of Theorem 1.1.

The following assertion illustrates principal possibilities of the scheme developed in §3, although it may be not of direct practical interest.

THEOREM 5.1. *Let $\varepsilon \in T$ and $\beta \in (0,1)$ be given. There exists a class of skeletons for* (5.1) *such that, for any of its elements, not only the skeleton itself but also any of its restrictions on arbitrary time intervals $(a,b) \subset T_{\varepsilon}$, regardless of the duration, make system* (1.1), (5.1) *l^{∞}-exact observable at final time $t = \theta$ with the same constant γ in* (5.2).

Proof. The idea of the proof is based on the fact that the results of §§3 and 4 employ the observations taken only at a countable set of instants of time: along the skeleton of the observation curve. Furthermore, these instants can be located in an arbitrary way in the interval T_{ε}.

Fix an arbitrary pair $\varepsilon \in T$ and $\beta \in (0,1)$. Let $\{t_i\}_{i=1}^{\infty}$ be an arbitrary set that is dense in T_{ε} and let $\{\delta_j\}_{j=1}^{\infty}$ be an arbitrary sequence of positive numbers such that $\lim_{j\to\infty} \delta_j = 0$. For each interval $(t_i - \delta_j, t_i) \bigcap T_{\varepsilon}$, select next a sequence of pairs according to Procedure 3.1 (with the same given ε and β),

$$\{x_{ij}^k, t_{ij}^k\}_{k=1}^{\infty}, \qquad i,j = 1, \ldots . \tag{5.3}$$

Observe that countability of the set of indices $i,j,k = 1,2 \ldots$ allows us to select all the instants t_{ij}^k to be distinct. Hence, we can renumber the pairs in (5.3) to obtain the sequence of pairs, forming the skeleton satisfying the requirements of Theorem 5.1.

6. From $L^\infty(T_\varepsilon)$-exact observability to approximate controllability. Consider in the domain Ω (satisfying the assumptions of §1) the following mixed problem:

$$\frac{\partial z(x,t)}{\partial t} = \Delta z(x,t) + v(t)\delta(x - x^*(t)), \quad t \in T, \quad x \in \Omega, \tag{6.1}$$
$$z(x,t)|_\Sigma = 0, \qquad z(x,0) = 0,$$

where $v(\cdot) \in L^2(T)$ is control and $x^*(\cdot)$ is a measurable function, $x^*(t) \in \bar{\Omega}$ almost everywhere in T.

System (6.1) is said to be approximately controllable in the Hilbert space H if its attainable set at time $t = \theta$ is dense in H.

To our knowledge, the problem of approximate controllability with internal pointwise controls for the parabolic systems is not well understood. The case of the moving spatially-averaged controls was considered by Martin, who reformulated the results of [20] (dual to [21]) in the form of moving controls. In this section, we apply the observability results, obtained in §§3 and 4, to the study of approximate controllability of (6.1). We restrict ourself by the case when $n \leq 3$, although Corollary 4.3 allows us to consider (in an appropriate space) an arbitrary space dimension.

We define the generalized solution of (6.1) by transposition (see [12, p. 186]) as a unique element of $L^2(Q)$ such that

$$\int_Q z(x,t)(-\frac{\partial \psi}{\partial t} - \Delta\psi)dxdt = \int_0^T \psi(x^*(t),t)v(t)dt \tag{6.2}$$
$$\forall \psi \in H^{2,1}(Q), \quad \psi\,|_\Sigma = 0, \quad \psi\,|_{t=\theta} = 0.$$

Indeed, for $n \leq 3$, any element ψ of $H^{2,1}(Q)$ is of Carathéodory type, and hence $\psi(x^*(\cdot),\cdot) \in L^2(T)$. Furthermore, the argument similar to that in [12, p. 202] gives

$$t \to z(\cdot,t) \text{ is a continuous function of } [0,\theta] \to H^{-1}(\Omega). \tag{6.3}$$

THEOREM 6.1. *Let $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$ be given and $n \leq 3$. Let $x^*(\cdot)$ be an arbitrary measurable curve such that $x^*(t) \equiv \hat{x}(\theta - t)$ for all $t \in (0, \theta - \varepsilon)$, where $\hat{x}(t)$, $t \in T_\varepsilon$ is constructed along Procedure* 3.1. *Then system* (6.1) *is approximately controllable in $H^{-1}(\Omega)$.*

Proof. Take any $\varepsilon \in (0,\theta)$ and $\beta \in (0,1)$. Introduce next the system dual to (6.1) as follows:

$$\frac{\partial u(x,t)}{\partial t} = -\Delta u(x,t), \quad t \in T, \quad x \in \Omega, \tag{6.4}$$
$$u(x,t)|_\Sigma = 0, \quad u(x,\theta) = u_\theta(x), \quad u_\theta(\cdot) \in H_0^1(\Omega),$$

$$y(t) = u(x^*(t),t), \qquad t \in T. \tag{6.5}$$

Observe that system (6.4) is well-posed in backward time and that its solutions belong to $H^{2,1}(Q)$ (see (2.2)). The conclusion of Theorem 6.1 now follows from the duality relations (see, for example, [4]) and Corollary 4.3 (deduced from $L^\infty(T_\varepsilon)$-exact observability), applied to system (6.4), (6.5). Moreover, to prove this, we can consider only controls of the following type:

$$v(\cdot) = \begin{cases} v^*(t) \in L^2(0,\theta-\varepsilon), & t \in (0,\theta-\varepsilon), \\ 0, & t \in (\theta-\varepsilon,\theta). \end{cases}$$

Indeed, in this case, from (6.1) and (6.4), (6.5), we obtain the identity

$$[z(\cdot,\theta),u_\theta(\cdot)] = \int_0^{\theta-\varepsilon} u(x^*(t),t)v^*(t)dt, \tag{6.6}$$

which is valid for $v^*(\cdot) \in L^2(0,\theta-\varepsilon)$, $u_\theta(\cdot) \in H_0^1(\Omega)$, where the symbol $[\cdot,\cdot]$ denotes the duality relation between $H_0^1(\Omega)$ and $(H_0^1(\Omega))' = H^{-1}(\Omega)$. Identity (6.6) implies the conclusion of Theorem 6.1.

7. Concluding remarks. In this paper, the problem of $L^\infty(T_\varepsilon)$-exact observability has been studied for the heat equation with internal scanning pointwise sensor. A new method for the construction of observation curves for sensors providing required observability has been given in the form of an abstract algorithm (Procedure 3.1), based on the classical maximum principle for the heat equation. Each iteration of the algorithm can be associated with some part of the interval (ε,θ) and provides exact observability in the corresponding finite-dimensional subspace of $L^2(\Omega)$, spanned by the eigenfunctions of (1.1). Procedure 3.1 is not affected by the multiplicities and the growth of the eigenvalues of the system in question and can be applied regardless of the system's space dimension. Approximate controllability of the heat equation in the space dimension $n \leq 3$ with scanning pointwise control has been established in $H^{-1}(\Omega)$ for the control curves, provided by the proposed method.

REFERENCES

[1] A. G. Butkovskiy and L. M. Pustyl'nikov, *Mobile Control of Distributed Parameter Systems*, Ellis Horwood Ltd., Chichester, UK, 1987.

[2] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1978.

[3] Sz. Dolecki, *Observation for the one-dimensional heat equation*, Stadia Math., 48 (1973), pp. 291–305.

[4] Sz. Dolecki and D. L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–219.

[5] A. El Jai and A. J. Pritchard, *Sensors and Actuators in the Analysis of Distributed Systems*, John Wiley, New York, 1988.

[6] H. O. Fattorini and D. L. Russell, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., April 1974, pp. 45–69.

[7] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[8] D. S. Gilliam, Z. Li, and C. F. Martin, *Discrete observability of the heat equation on bounded domains*, Internat. J. Control, 48 (1988), pp. 755–780.

[9] A.Yu. Khapalov, *Optimal measurement trajectories for distributed parameter systems*, Systems Control Lett., 18 (1992), pp. 467–477.

[10] C. S. Kubrusly and H. Malebranche, *Sensors and controllers location in distributed systems—A survey*, Automatica, 21 (1985), pp. 117–128.

[11] A. B. Kurzhanski and A. Yu. Khapalov, *An observation theory for distributed-parameter systems*, J. Math. Systems, Estim. Control, 1 (1991), pp. 389–440.

[12] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.

[13] W. A. J. Luxemburg and J. Korevaar, *Entire functions and Müntz–Szász type approximation*, Trans. Amer. Math. Soc., 157 (1971), pp. 23–37.

[14] J-C. E. Martin, *Controllability and observability of parabolic systems—An addendum to two recent papers of Y. Sakawa*, SIAM J. Control Optim., 15 (1977), pp. 363–366.

[15] ———, *Optimal selection of actuators for lumped and distributed parameter systems*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 70–78.

[16] V. P. Mikhailov, *Partial Differential Equations*, Nauka, Moscow, 1976. (English transl., "Mir," Moscow, 1978.)

[17] V. J. MIZEL AND T. I. SEIDMAN, *Observation and prediction for the heat equation*, J. Math. Anal. Appl., 28 (1969), pp. 303–312.

[18] S. OMATU AND J. H. SEINFELD, *Estimation of atmospheric species concentrations from remote sensing data*, IEEE Trans. Geosci. Remote Sensing, GE-20 (1982), pp. 142–153.

[19] C. A. ROGERS, *An introduction to intelligent material systems and structures,* in Intelligent Structures, K. P. Chong, S. C. Liu, and J. C. Li, eds., Elsevier, New York, 1990, pp. 3–41.

[20] Y. SAKAWA, *Controllability for partial differential equations of parabolic type*, SIAM J. Control, 12 (1974), pp. 389–400.

[21] ———, *Observability and related problems for partial differential equations of parabolic type,* SIAM J. Control, 13 (1975), pp. 14–27.

[22] T. I. SEIDMAN, *Observation and prediction for one-dimensional diffusion equations*, J. Math. Anal. Appl., 51 (1975), pp. 165–175.

[23] L. SCHWARTZ, *Étude des sommes d'exponentielles rélees*, Actualités Sci. Indust., No. 959, Hermann, Paris, 1943, MR 7, p. 294.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1052–1074, July 1994

010

BOUNDARY CONTROL OF A ONE-DIMENSIONAL LINEAR THERMOELASTIC ROD*

SCOTT W. HANSEN†

Abstract. Boundary control of a linear partial differential equation that describes the temperature distribution and displacement within a one-dimensional thermoelastic rod is examined. In particular, it is shown that temperature or heat flux control at an endpoint is sufficient to obtain exact null-controllability. This improves earlier results for similar systems in which only partial null-controllability is obtained. Sharp regularity results for the controlled system are also obtained.

Key words. linear thermoelasticity, moment problem, boundary control, regularity

AMS subject classifications. 93B05, 80A20, 70J99

1. Introduction.

Although there is extensive literature on the topic of control and stabilization of elastic systems, relatively little has been published that includes the thermoelastic coupling. This is probably due, in part, to the relatively small effect thermoelastic damping has upon most systems of interest. However, for certain applications such as stabilization of satellite antennas, where large temperature variations are common (e.g., due to moving in and out of shadows), the need to model this coupling becomes critical. Furthermore, the recent work of Gibson, Rosen, and Tao [5] illustrates the importance of modelling even light thermoelastic damping in the design of finite-dimensional compensators.

Some notable literature on stabilization of thermoelastic systems include [13], [14], [16]–[18], and references therein. Very little, however, is known about the controllability structure of thermoelastic systems. In Lagnese and Lions [14], boundary control (e.g., velocity or position control on the boundary) is used to exactly control the mechanical portion of the state space. This type of controllability is called *partial exact controllability.* When this type of control is used, the thermal component of the state is ignored, and, consequently, if the mechanical portion is driven to rest, it will not generally remain there due to the thermal stresses that remain. The main purpose of this paper is to show that, at least for the case of a one-dimensional thermoelastic rod, exact controllability (to zero) of both mechanical and thermal components of the state space is possible by only controlling the thermal (or mechanical) component on the boundary.

A derivation of the equations of one-dimensional nonlinear thermoelasticity can be found in [21]. In the case of a homogeneous rod with uniform cross sections (see [2], [6] for the precise assumptions), the linearization of these equations can be written as

$$\begin{aligned} \frac{\partial\theta}{\partial t}(t,x) &= \frac{\partial^2\theta}{\partial x^2}(t,x) - \frac{\gamma\partial^2 w}{\partial x\partial t}(t,x), \\ \frac{\partial^2 w}{\partial t^2}(t,x) &= c^2\frac{\partial^2 w}{\partial x^2}(t,x) - c^2\gamma\frac{\partial\theta}{\partial x}(t,x), \end{aligned} \tag{1.1}$$

*Received by the editors November 25, 1991; accepted for publication (in revised form) January 15, 1993.

†Department of Mathematics, Iowa State University, Ames, Iowa 50011. A portion of this research was conducted while the author was supported by the Institute for Mathematics and its Applications, 514 Vincent Hall, University of Minnesota, Minneapolis, Minnesota 55455.

which holds on $(t,x) \in (0,\infty) \times \Omega$ ($\Omega = (0,1)$). Here θ represents a relative temperature about the stress-free reference state $\theta = 0$, and w is proportional to the displacement. The constants $\gamma > 0$ and $c > 0$ represent, respectively, the amount of thermal-mechanical coupling and the small-amplitude wave speed about a constant temperature state. (See [6] for a precise definition of γ and c.) In most materials of interest, γ is several orders of magnitude smaller than 1.

The physical quantities relevant to the formulation of boundary conditions for (1.1) are the velocity v, heat flux q, stress σ, and temperature θ, where the first three of these are

$$v(t,x) = \frac{\partial w}{\partial t}(t,x),$$
$$q(t,x) = -\frac{\partial \theta}{\partial x}(t,x),$$
$$\sigma(t,x) = \frac{\partial w}{\partial x}(t,x) - \gamma\theta(t,x).$$

In [6] it was shown that, under any of the boundary conditions,

$$v(t,i) = 0, \quad q(t,i) = 0, \quad i = 0,1; \tag{1.2}$$

$$\sigma(t,i) = 0, \quad \theta(t,i) = 0, \quad i = 0,1; \tag{1.3}$$

$$\sigma(t,0) = \theta(t,0) = v(t,1) = q(t,1) = 0, \tag{1.4}$$

the eigenfunctions associated with (1.1) form a Riesz basis for the space of finite energy states and the corresponding eigenvalues are uniformly shifted into the left half-plane, except for possibly one or two eigenvalues located at the origin. This result is partially restated in Theorem 2.1 and is our starting point in our examination of associated control problems.

In the case of boundary conditions (1.4), there are no eigenvalues at the origin. For this reason, it is notationally convenient to restrict our presentation to control of boundary conditions of the type (1.4), although similar results apply for control of boundary conditions of the type (1.2) or type (1.3).

Let $y(t) = (y_1(t), y_2(t), y_3(t))' = (w_x(t,\cdot), w_t(t,\cdot), \theta(t,\cdot))'$ represent the state of system (1.1) at time t and let $(v_y(x), q_y(x), \sigma_y(x), \theta_y(x))$ represent the velocity, heat flux, and so forth, in terms of the state y. We are mainly concerned with the following boundary control problem associated with (1.1):

$$\frac{dy}{dt} = \tau y \equiv \begin{bmatrix} 0 & D & 0 \\ c^2 D & 0 & -\gamma c^2 D \\ 0 & -\gamma D & D^2 \end{bmatrix} y, \qquad (t,x) \in (0,\infty) \times \Omega, \tag{1.5}$$

$$y(0) = y^0 \quad \text{in } \Omega, \tag{1.6}$$

$$\begin{aligned} \sigma_{y(t)}(0) = 0, \quad \theta_{y(t)}(0) = g(t), \quad t \geq 0, \\ v_{y(t)}(1) = 0, \quad q_{y(t)}(1) = f(t), \quad t \geq 0, \end{aligned} \tag{1.7}$$

where $D = d/dx$, $\gamma > 0$, $c > 0$. Thus, at the left end, the temperature is controlled, and the stress vanishes. At the right end, the heat flux is controlled while the position is fixed.

We must define some function spaces to describe our main results. Let $\mathcal{H} = (L^2(\Omega))^3$ with the energy inner product

$$\langle y, z \rangle = \int_0^1 y_1 \bar{z}_1 + \frac{1}{c^2} y_2 \bar{z}_2 + y_3 \bar{z}_3 \, dx$$

and let

$$\mathcal{D}(A) = \{y \in H^1[0,1] \times H^1[0,1] \times H^2[0,1] \mid \sigma_y(0) = \theta_y(0) = v_y(1) = q_y(1) = 0\}.$$

Now define $A : \mathcal{D}(A) \to \mathcal{H}$ by

$$Ay = \tau y \quad \forall\, y \in \mathcal{D}(A). \tag{1.8}$$

We denote $l^2 = \{(c_k)_{k\in\mathbb{I}} \mid \sum_{k\in\mathbb{I}} |c_k|^2 < \infty\}$, where $\mathbb{I}$ is a countable index set (usually either the integers $\mathbb{Z}$ or positive integers $\mathbb{N}$). For $\alpha \in \mathbb{R}$, define

$$S_\alpha = \left\{ \sum_{k=1}^{\infty} a_k \sin\left(k\pi - \frac{\pi}{2}\right) x \mid (a_k k^\alpha) \in l^2 \right\}, \tag{1.9}$$

$$C_\alpha = \left\{ \sum_{k=1}^{\infty} a_k \cos\left(k\pi - \frac{\pi}{2}\right) x \mid (a_k k^\alpha) \in l^2 \right\}. \tag{1.10}$$

S_α and C_α become Hilbert spaces with, e.g., $\|y\|_{S_\alpha} = \|(a_k k^\alpha)\|_{l^2}$. (When $\alpha < 0$, S_α and C_α are the dual spaces to $S_{-\alpha}$ and $C_{-\alpha}$, respectively.) $C((a,b), M)$ denotes the set of functions that are continuous on the interval (a,b) with values in the space M.

Our main results are the following, together with related results given in §§3–5.

THEOREM 1.1. *Let $y^0 = 0$, $f \in L^2(0,\infty)$, and $g \in L^2(0,\infty)$. Then the solution to (1.5)–(1.7) belongs to $C([0,\infty), S_0 \times C_0 \times S_{-1/2})$. If, additionally, $g \equiv 0$, then the solution belongs to $C([0,\infty), S_1 \times C_1 \times S_{1/2})$. These solution spaces are optimal in the sense that none of the indices $\{0, 1, 1/2, -1/2\}$ may be increased.*

THEOREM 1.2. *Assume that $0 < \gamma \le 1$ in (1.5) and $T > 2/c$.*

(i) *For the boundary control problem (1.5)–(1.7), with $f \equiv 0$, given any $y^0 \in \mathcal{H}$, there exists $g \in L^2[0,T]$ such that $y \in C([0,T], S_0 \times C_0 \times S_{-1/2})$ and $y(T) = 0$.*

(ii) *For the boundary control problem (1.5)–(1.7), with $g \equiv 0$, given any $y^0 \in \mathcal{D}(A)$, there exists $f \in L^2[0,T]$ such that $y \in C([0,T], S_1 \times C_1 \times S_{1/2})$ and $y(T) = 0$. In either case, T cannot in general be reduced to $2/c$.*

Remark 1.3. The following identifications hold:

$$\begin{gathered} S_0 = C_0 = L^2(\Omega), \\ S_1 = \{f \in H^1(\Omega) \mid f(0) = 0\}, \\ C_1 = \{f \in H^1(\Omega) \mid f(1) = 0\}, \\ S_{1/2} = [S_1, S_0]_{1/2} = \{f \in H^{1/2}(\Omega) \mid x^{-1/2} f(x) \in L^2(\Omega)\}, \\ S_{-1/2} = S_{1/2}', \\ \mathcal{H} = S_0 \times C_0 \times S_0, \\ \mathcal{D}(A) = S_1 \times C_1 \times S_2, \end{gathered}$$

with equivalent norms; see [15] and §2. (In the above, H^α denotes the usual Sobolev space of order α, $[F,G]_{1/2}$ denotes the usual interpolation space between spaces F and G, as defined in [15], and $'$ denotes duality with respect to $L^2(\Omega)$.)

In the above theorems, solutions are uniquely defined by continuous extension of the variation of constants formula; see §3 for details.

Proposition 5.1 gives a more general statement of Theorem 1.2, and Remark 5.2 shows that the spaces used in Theorem 1.2 are optimal in a certain sense.

The proof of Theorem 1.1 is given in §3 and involves an application of the Carleson measure criterion of Ho and Russell [10] and Weiss [22], which gives a sharp criterion for wellposedness of control systems. The proof of Theorem 1.2 involves reducing the control problem to a pair of coupled moment problems that are coupled through the control. A general class of such coupled moment problems is examined in §4, where it is shown that there are projections that decouple such moment problems into simpler ones for which known results are applicable. This leads to various controllability results, including Theorem 1.2, which are given in §5.

Results similar to Theorems 1.1 and 1.2 follow in the same way for boundary control systems based on the boundary condition (1.2) or (1.3). Likewise, we could also consider the case where the stress and/or velocity at an end is controlled, and similar results would follow. We mention some of these results in §5.

A short appendix is included which contains the proof of several technical details used throughout the rest of the paper.

2. Preliminaries. Throughout this paper, an *isomorphism* is understood to denote a bounded, invertible operator from one Hilbert space onto another. If X is a separable Hilbert space, a sequence $(\varphi_k)_{k\in\mathbb{N}}$ in X forms a *Riesz basis* for X if $\varphi_k = Be_k$ $(k \in \mathbb{N})$, where (e_k) is an orthonormal basis for X and B is an isomorphism. The following theorem and its corollary were proved in Hansen [6].

THEOREM 2.1. *Let A be defined by* (1.8). *The spectrum of A (and also of A^*) consists of isolated eigenvalues $(\lambda_{kj})_{k\in\mathbb{N},j\in\{1,2,3\}}$ with $\lambda_{kj} = (k\pi - \pi/2)s_{kj}$, where*

$$(s_{kj}^2 + c^2)(s_{kj} + k\pi - \pi/2) + \gamma^2 c^2 s_{kj} = 0. \tag{2.1}$$

The eigenfunctions of A (and also A^*), properly normalized, form a Riesz basis for $\mathcal{H}$.

An analysis of (2.1) in [6] shows that (λ_{kj}) can be decomposed into a real branch $(\mu_k)_{k\in\mathbb{N}}$ and a nonreal branch $(\sigma_k)_{k\in\mathbb{Z}}$ with

$$\begin{aligned} \mu_k &= -\left(k\pi - \frac{\pi}{2}\right)^2 + O(1), \quad k \in \mathbb{N}, \\ \sigma_k &= -\frac{\gamma^2}{2} + ic\left(k\pi - \frac{\pi}{2}\right) + O(k^{-1}), \quad k \in \mathbb{Z}. \end{aligned} \tag{2.2}$$

We let $(\psi_\lambda)_{\lambda\in\sigma(A)}$ denote the normalized eigenvectors of A^* and $(\varphi_\lambda)_{\lambda\in\sigma(A)}$ denote the biorthonormalized eigenvectors of A (each eigenvalue is counted up to its multiplicity), so that $\langle \varphi_{\lambda_k}, \psi_{\lambda_j} \rangle = \delta_{kj}$, where δ_{kj} is the Kronecker delta. There are at most a finite number of eigenvalues of multiplicity greater than 1, and all eigenvalues are simple if $\gamma \leq 1$ (see Lemma A.1) or if $|k|$ is sufficiently large (see [6]). The form of the eigenvectors of A^* is given in the Appendix.

COROLLARY 2.2. *A is the generator of a strongly continuous contraction semigroup $(\mathbb{T}_t)_{t\geq 0}$ on $\mathcal{H}$ for which there exist $M > 1$ and $\beta > 0$ such that*

$$\|\mathbb{T}_t\| \leq Me^{-\beta t} \quad \forall\, t \geq 0.$$

Theorem 2.1 and Corollary 2.2 also hold in the case of boundary condition (1.2) or (1.3), although the energy decay occurs in the orthogonal complement of the nullspace of the generator [6]. In addition, several recent papers [1], [12], [17] have shown

exponential stability to hold for other sets of natural boundary conditions. These exponential stability results are important in that we can infer the existence of optimizing feedbacks for stabilization problems with quadratic cost criteria; see [5].

For any set $S \subset \mathbb{C}$, we can define an associated spectral projection $P(S) \in \mathcal{L}(\mathcal{H})$ by

$$(P(S))\,x = \frac{1}{2\pi i}\int_{\Gamma} R(\lambda; A)x\,d\lambda \quad \forall\, x \in \mathcal{H},$$

where $R(\lambda, A)$ is the resolvent operator of A and where Γ is an appropriate contour that encloses the eigenvalues in S. There is no difficulty in defining Γ, since the spectrum is discrete. In cases where Γ contains infinitely many eigenvalues, convergence for all $x \in \mathcal{H}$ is guaranteed by Theorem 2.1. Let us denote

$$P = P(\mathbb{R}) \quad \text{and} \quad Q = I - P(\mathbb{R}),$$

where I denotes the identity operator on $\mathcal{H}$. Let

$$\Lambda = P\mathcal{H} \quad \text{and} \quad \Sigma = Q\mathcal{H}.$$

Since the projections are continuous, it follows that $\mathcal{H} = \Lambda \oplus \Sigma$.

PROPOSITION 2.3. *Let $\mathbb{T}$ denote the semigroup defined in Corollary 2.2. Then, for $t \geq 0$,*

$$\mathbb{T}_t = \mathbb{S}_t P + \mathbb{G}_t Q, \tag{2.3}$$

where $\mathbb{G}$ extends to a strongly continuous group $(\mathbb{G}_t)_{t\in\mathbb{R}}$ and $\mathbb{S}$ extends to an analytic semigroup $(\mathbb{S}_t)_{\mathrm{Re}\ t>0}$. The infinitesimal generators of $\mathbb{S}$ and $\mathbb{G}$ are given by the restrictions of A, $A|_\Lambda$, and $A|_\Sigma$, respectively.

Proof. The spaces Λ and Σ are closed $\mathbb{T}$-invariant spaces, and hence the restriction of $\mathbb{T}$ to either of these spaces is a C_0 semigroup with respect to the inherited topology. For $t \geq 0$, let $\mathbb{S}_t = \mathbb{T}_t|_\Lambda$ and $\mathbb{G}_t = \mathbb{T}_t|_\Sigma$. It follows that, for $t \geq 0$, $\mathbb{T}_t = \mathbb{T}_t(P+Q) = \mathbb{S}_t P + \mathbb{G}_t Q$; hence (2.3) is valid. For any $x \in \Lambda \cap \mathcal{D}(A)$,

$$Ax = \lim_{t\downarrow 0} \frac{\mathbb{T}_t x - x}{t} = \lim_{t\downarrow 0} \frac{\mathbb{S}_t x - x}{t}.$$

Thus $\mathbb{S}$ is generated by (the densely defined operator) $A|_\Lambda$, and likewise $\mathbb{G}$ is generated by $A|_\Sigma$.

It remains to show that $\mathbb{G}$ extends to a group (by $\mathbb{G}_{-t} = \mathbb{G}_t^{-1}$) and that $\mathbb{S}$ has an analytic extension to $\mathrm{Re}\, t > 0$. Define $F : \mathcal{H} \to l^2$ by

$$\sum_{\lambda_k \in \sigma(A)} c_{\lambda_k}\phi_{\lambda_k} \to (c_{\lambda_k}).$$

Since (ϕ_{λ_k}) forms a Riesz basis for $\mathcal{H}$, F is an isomorphism. Define $\tilde{\mathbb{T}} = (\tilde{\mathbb{T}}_t)_{t\geq 0}$ by $\tilde{\mathbb{T}}_t = F\mathbb{T}_t F^{-1}$, and for $x \in \mathcal{H}$ define $\tilde{x}$ by $\tilde{x} = Fx$. Through this mapping, the pair $(\mathbb{T}, \mathcal{H})$ is isomorphic to $(\tilde{\mathbb{T}}, l^2)$ in the sense that $F\mathbb{T}_t x = \tilde{\mathbb{T}}_t \tilde{x}$ for any $t \geq 0$ and any $x \in \mathcal{H}$. Since any Riesz basis becomes an orthonormal basis under some equivalent inner product (see [25]), it follows that the induced topology $\|\tilde{x}\|_i = \|x\|$ is equivalent to the topology generated by the standard l^2 inner product. Thus, to show that $\mathbb{S}$

extends to an analytic semigroup $(\mathbb{S}_t)_{\mathrm{Re}\, t>0}$, it suffices to show that $\tilde{\mathbb{T}}|_{F\Lambda}$ extends analytically to $\mathrm{Re}\, t > 0$ with respect to the standard l^2 topology. It is easily seen that $\tilde{\mathbb{T}}|_{F\Lambda}$ is a diagonal semigroup on l^2 $(= F\Lambda)$ with (diagonal) generator $(FAF^{-1})|_{F\Lambda}$. If $\mathcal{A}$ is any diagonal generator, it is easy to show that $\|R(\lambda, \mathcal{A})\|$ is inversely proportional to the distance λ is from $\sigma(\mathcal{A})$. Since $\sigma(A|_\Lambda) = \sigma((FAF^{-1})|_{F\Lambda})$ is entirely on the negative real axis, the appropriate resolvent bound [19, p. 62] holds, which shows that $\tilde{\mathbb{T}}|_{F\Lambda}$, and hence also $\mathbb{S}$, extends to an analytic semigroup in $\mathrm{Re}\, t > 0$. Likewise, since $\sigma(FAF^{-1}|_{F\Sigma})$ is contained in a vertical strip of $\mathbb{C}$, it follows from, e.g., Pazy [19, p. 23] that $\tilde{\mathbb{T}}|_{F\Sigma}$ extends to a group. Hence $\mathbb{G}$ also extends to a group. □

It will be useful to introduce notation for certain interpolation spaces. Since $0 \in \rho(A)$ (the resolvent set of A) and $\sigma(-A)$ is in $\{\lambda \in \mathbb{C} \mid \mathrm{Re}\ \lambda > 0\}$, for any $\alpha \in \mathbb{R}$, $(-A)^\alpha$ may be defined as in, e.g., Pazy [19, p. 69]. For $\alpha > 0$, $(-A)^\alpha$ is an isomorphism from $\mathcal{D}((-A)^\alpha)$ (with graph-norm topology) to $\mathcal{H}$. For $\alpha \geq 0$, we define $\mathcal{H}_\alpha$ to be the restriction of $\mathcal{H}$ to $\mathcal{D}((-A)^\alpha)$ with

$$\|x\|_\alpha = \|(-A)^\alpha x\|. \tag{2.4}$$

For $\alpha < 0$, we let $\mathcal{H}_\alpha$ denote the completion of $\mathcal{H}$ with respect to the norm also given by (2.4). The above spaces are explained in more detail in, e.g., [8], [23]. For our problem, we have that $\mathcal{D}(A) = \mathcal{D}(A^*)$. Thus it follows that $\mathcal{H}_1^* = \mathcal{H}_{-1}$ (where the duality pairing is with respect to the completion of $\langle \cdot, \cdot \rangle$). Furthermore, since the eigenfunctions of A form a Riesz basis, it can be shown that, for $\alpha \in [0, 1]$, $\mathcal{H}_\alpha = [\mathcal{H}_1, \mathcal{H}_0]_{1-\alpha}$, where $[\mathcal{H}_1, \mathcal{H}_0]_{1-\alpha}$ is the interpolation space defined in [15]. Using standard properties of interpolation spaces, we can show $\mathcal{H}_\alpha^* = \mathcal{H}_{-\alpha}$ for all $\alpha \in \mathbb{R}$.

We recall a result from Weiss [23], as it applies to our problem.

PROPOSITION 2.4. *For any $\alpha < 0$, A has a unique continuous extension to an operator on $\mathcal{H}_\alpha$, also denoted by A, which is an isomorphism from $\mathcal{H}_{\alpha+1}$ to $\mathcal{H}_\alpha$. Furthermore, if L commutes with A, i.e., if*

$$LAx = ALx \quad \forall x \in \mathcal{H}_1 = \mathcal{D}(A),$$

then the restriction of L to $\mathcal{H}_\alpha$ $(\alpha > 0)$ belongs to $\mathcal{L}(\mathcal{H}_\alpha)$. Furthermore, L has a unique continuous extension to an operator in $\mathcal{L}(\mathcal{H}_\alpha)$ for any $\alpha < 0$.

In particular, the projections P and Q, and semigroups $\mathbb{T}$, $\mathbb{S}$, and $\mathbb{G}$ each have unique continuous extensions to $\mathcal{H}_\alpha$ (for any $\alpha < 0$). Throughout this paper, we make no notational distinction between an operator and its possible extensions. We thus define the spaces Λ_α and Σ_α by

$$\Lambda_\alpha = P\mathcal{H}_\alpha \quad \text{and} \quad \Sigma_\alpha = Q\mathcal{H}_\alpha \quad \forall \alpha \in \mathbb{R}.$$

As a consequence of Proposition 2.4,

$$\mathcal{H}_\alpha = \Lambda_\alpha \oplus \Sigma_\alpha \quad \forall\, \alpha \in \mathbb{R}. \tag{2.5}$$

The spaces Λ_α and Σ_α become Hilbert spaces with the norms $\|\cdot\|_{\Lambda_\alpha}$ and $\|\cdot\|_{\Sigma_\alpha}$ inherited from (2.4).

3. Regularity. In this section, we obtain via the Carleson measure criterion of Ho and Russell [10] and Weiss [22] the spaces of maximal regularity of system (1.5)–(1.7). We begin with a discussion of the Carleson measure criterion.

Consider the control system

$$\dot{x} = \mathcal{A}x + bu(t), \tag{3.1}$$

where $x(t) \in l^2$ is the state, $u \in L^2[0,\infty)$ is the control function, $\mathcal{A}$ is assumed to be diagonal with diagonal elements ν_k, which satisfy

$$\sup_{k\in\mathbb{N}} \operatorname{Re} \nu_k = \omega_0 < 0, \tag{3.2}$$

and $b \in l^2_{-1}$, i.e., is a column vector with components b_k, which satisfy

$$\sum_{k=1}^{\infty} \left| \frac{b_k}{\nu_k} \right|^2 < \infty.$$

Thus $\mathcal{A}$ generates a strongly continuous diagonal semigroup $(T_t)_{t\geq 0}$ on l^2.

For any $h > 0$ and any $\omega \in \mathbb{R}$, let

$$R(h,\omega) = \{z \in \mathbb{C} \mid 0 \leq \operatorname{Re} z \leq h,\ |\operatorname{Im} z - \omega| \leq h\}.$$

DEFINITION 3.1. With $\mathcal{A}$, b, and T as above, b satisfies the *Carleson measure criterion* for the semigroup T if there is some $M \geq 0$ such that, for any $h > 0$ and any $\omega \in \mathbb{R}$,

$$\sum_{-\nu_k \in R(h,\omega)} |b_k|^2 \leq M \cdot h. \tag{3.3}$$

The Carleson measure criterion is used to determine the *admissibility* of the *input element* b in (3.1). The input element b is *admissible* for T if, for some $t > 0$, the sequence $\left(\int_0^t e^{\nu_k(t-s)} b_k v(s) ds\right)_{k\in\mathbb{N}}$ lies in l^2 for all $v \in L^2[0,\infty)$. When b is admissible, for any $\tau > 0$, the operator $\Phi_\tau \colon L^2[0,\infty) \to l^2_{-1}$ defined by

$$\Phi_\tau u = \int_0^\tau T_{\tau-s} b u(s) ds \quad \forall\, u \in L^2[0,\infty) \tag{3.4}$$

maps continuously into l^2. In this case, for any initial condition $x_0 \in l^2$ and any $u \in L^2(0,\infty)$, a unique solution of (3.1) is given by

$$x(t) = T_t x_0 + \Phi_t u, \tag{3.5}$$

with $x \in C([0,\infty), l^2)$. If b is not admissible, then there exists $u \in L^2[0,\infty)$ for which the solution of (3.1) (if it can be defined at all) is not continuous in time.

Remark 3.2. It should be pointed out that the stability restriction (3.2) is unessential; we have defined the Carleson measure criterion as it applies to stable systems. See [10] for the general definition.

Remark 3.3. In Definition 3.1, it is not necessary to verify (3.3) for every possible value of (h,ω). It is enough to consider the pairs (h_n,ω_n) for which $\nu_n = h_n + i\omega_n$. (This follows by a simple geometrical argument.)

THEOREM 3.4 (Ho and Russell, Weiss). *With b, $\mathcal{A}$, and T as above, b is admissible for T if and only if b satisfies the Carleson measure criterion for T.*

The above asserts that the control system (3.1) is well-posed on l^2 (in the above discussed sense) if and only if the sequence (b_k) satisfies (3.3).

For $\alpha \in \mathbb{R}$, we denote $l^2_\alpha = \{(c_k) \mid (|\nu_k|^\alpha c_k) \in l^2\}$.

DEFINITION 3.5. Let $\alpha \in \mathbb{R}$. With b, $\mathcal{A}$, and T as above, the pair (b,T) is *well-posed on l^2_α* if, for some $\tau > 0$, the operator Φ_τ defined in (3.4) maps continuously into l^2_α.

If (b,T) is well-posed on l^2_α, then we may define solutions of (3.1) by (3.5), and these solutions are continuous in time with values in l^2_α. We have the following corollary.

COROLLARY 3.6. *Let $\alpha \in \mathbb{R}$. The pair (b,T) in (3.1) is well-posed on l^2_α if and only if $(b_k|\nu_k|^\alpha)_{k\in\mathbb{N}}$ satisfies the Carleson measure criterion for T.*

Proof. Let $\tau > 0$ and $u \in L^2(0,\infty)$. Let $(\zeta_k) = \Phi_\tau u$ as given by (3.4). If $(b_k|\nu_k|^\alpha)_{k\in\mathbb{N}}$ satisfies the Carleson measure criterion, then $(\zeta_k|\nu_k|^\alpha) \in l^2$ or, equivalently, $(\zeta_k) \in l^2_\alpha$. □

We now return to the control problem (1.5)–(1.7). To apply Theorem 3.4 to our system, the input elements associated with (1.5)–(1.7) must be identified.

Let $G : \mathbb{R}^2 \to \mathcal{H}$ denote the *Green's map* associated with (1.5)–(1.7):

$$\begin{aligned} G(u_1,u_2)' = w; \qquad \tau w = 0 \quad \text{in } \Omega, \\ \sigma_w(0) = 0, \quad v_w(1) = 0, \quad \theta_w(0) = u_1, \quad q_w(1) = u_2. \end{aligned}$$

We obtain $G(u_1,u_2)' = (\gamma(-u_2 x + u_1), 0, -u_2 x + u_1)'$. If $y^0 = 0$ and $f, g \in C_0^\infty(0,\infty)$, then the (classical) solution y to (1.5)–(1.7) at time t coincides with an element of $\mathcal{H}_{-1}$ $(= \mathcal{D}(A^*)^*)$, also denoted by $y(t)$, which is given by (e.g., [3], [24])

$$y(t) = -\int_0^t A\mathbb{T}_{t-\tau} G(g,f)'(\tau)\, d\tau. \tag{3.6}$$

Since the appropriate extensions of A and $\mathbb{T}$ commute on $\mathcal{H}$,

$$\begin{aligned} y(t) &= \int_0^t \mathbb{T}_{t-s}(-AG)(g,f)'(s)\, ds \\ &\equiv \int_0^t \mathbb{T}_{t-s} B(g,f)'(s)\, ds. \end{aligned}$$

The boundary control operator B maps $\mathbb{R}^2$ into $\mathcal{H}_{-1}$ continuously and hence is a sum of two continuous functionals on $\mathcal{H}_1$. From integration by parts,

$$\begin{aligned} \langle B(u_1,u_2)', w\rangle &= \langle -G(u_1,u_2)', A^* w\rangle \\ &= -u_1 \bar{q}_w(0) - u_2 \bar{\theta}_w(1) \quad \forall\, w \in \mathcal{H}_1. \end{aligned}$$

Thus we define b_0, b_1 as elements of $\mathcal{H}_{-1}$ by

$$\begin{aligned} \langle b_0, \bar{w}\rangle &= -q_w(0) \quad \forall\, w \in \mathcal{H}_1, \\ \langle b_1, \bar{w}\rangle &= -\theta_w(1) \quad \forall\, w \in \mathcal{H}_1, \end{aligned} \tag{3.7}$$

so that (3.6) becomes

$$y(t) = \int_0^t \mathbb{T}_{t-s}(b_0 g(s) + b_1 f(s))\, ds \quad \text{on } \mathcal{H}_{-1}. \tag{3.8}$$

The map $(g,f)' \to y$ as given by (3.6) (or (3.8)) is bounded when considered as a map

$$(L^2(0,T))^2 \to C([0,T],\mathcal{H}_{-1}), \qquad (T>0)$$

and thus defines a generalized solution for $(g,f) \in (L^2(0,T))^2$. It follows that

$$\dot{y} = Ay + b_0 g(t) + b_1 f(t), \qquad y(0) = y^0 \in \mathcal{H} \tag{3.9}$$

has a unique continuous solution in $\mathcal{H}_{-1}$ (given by (3.8) if $y^0 = 0$), which satisfies (3.9) on $\mathcal{H}_{-2}$.

By Propositions 2.3 and 2.4, the projections P and Q continuously decompose the solutions in (3.8) by $y(t) = x(t) + z(t)$, where

$$x(t) = \int_0^t \mathbb{S}_{t-s}(Pb_0 g(s) + Pb_1 f(s))\, ds \quad \text{on } \Lambda_{-1}, \tag{3.10}$$

$$z(t) = \int_0^t \mathbb{G}_{t-s}(Qb_0 g(s) + Qb_1 f(s))\, ds \quad \text{on } \Sigma_{-1}. \tag{3.11}$$

Note that all of the results in this section that pertain to diagonal systems apply to system (3.9), since (as in the proof of Proposition 2.3) A, $\mathbb{T}$, $\mathbb{G}$, $\mathbb{S}$, and so forth can be viewed as diagonal operators on l^2 relative to the Riesz basis of eigenfunctions. Likewise, an input element b may be identified with a vector in l^2_{-1} whose components are its respective Fourier coefficients. As such, the Carleson measure criterion can be used to check wellposedness of the pairs $(b,\mathbb{T})$ on $\mathcal{H}_\alpha$.

An analysis of the admissibility of the input elements Pb_0, Pb_1, Qb_0, and Qb_1 provides the smoothest spaces Λ_α and Σ_β in which $x(t)$ and $z(t)$ are time-continuous for all L^2 controls. This then determines the maximal regularity of the solutions $y(t)$ to system (1.5)–(1.7). We have the following result.

PROPOSITION 3.7. *In the above notation,*

(i) $(Pb_0, \mathbb{S})$ *is well-posed on* Λ_α *for all* $\alpha \le -1/4$,
(ii) $(Pb_1, \mathbb{S})$ *is well-posed on* Λ_α *for all* $\alpha \le 1/4$,
(iii) $(Qb_0, \mathbb{G})$ *is well-posed on* Σ_α *for all* $\alpha \le 0$,
(iv) $(Qb_1, \mathbb{G})$ *is well-posed on* Σ_α *for all* $\alpha \le 1$.

Furthermore, the bounds given for α *are sharp.*

Proof. We first prove (i). By (3.7), $b_0 \in \mathcal{H}_{-1}$, and hence $Pb_0 \in \Lambda_{-1}$. Therefore its series

$$\begin{aligned} Pb_0 &= \sum_{k\in\mathbb{N}} \langle Pb_0, \psi_{\mu_k}\rangle \varphi_{\mu_k} \\ &= \sum_{k\in\mathbb{N}} \langle b_0, \psi_{\mu_k}\rangle \varphi_{\mu_k} \equiv \sum_{k\in\mathbb{N}} c_k \varphi_{\mu_k} \end{aligned}$$

converges in Λ_{-1}. The coefficients (c_k) are easily computed from (3.7) and (A.5) (of the Appendix). It follows that there exist positive constants m and M such that

$$mk < |c_k| < Mk \quad \forall\, k \in \mathbb{N}. \tag{3.12}$$

For $k \in \mathbb{N}$, let $b_k = c_k/|\mu_k|^{1/4}$. The semigroup $\mathbb{S}$ can be identified with the diagonal semigroup $\tilde{\mathbb{S}} \equiv \operatorname{diag}(e^{\mu_1 t}, e^{\mu_2 t}, \dots)$ relative to the Riesz basis of eigenfunctions. Thus, by Corollary 3.6, (i) holds if the sequence (b_k) satisfies the Carleson measure criterion for $\tilde{\mathbb{S}}$. Since the eigenvalues (μ_k) grow quadratically (see (2.2)), (3.12) implies that there are constants $m_1 > 0$ and $M_1 > 0$ for which

$$m_1 k < |b_k|^2 < M_1 k \quad \forall k \in \mathbb{N}.$$

It follows that there are positive numbers m_2, m_3, M_2, M_3 for which

$$m_3|\mu_n| < m_2 n^2 < \sum_{k=1}^{n} |b_k|^2 < M_2 n^2 < M_3|\mu_n| \quad \forall n \in \mathbb{N}.$$

Thus, if $N \in \mathbb{N}$ and $h = |\mu_N|$, we have

$$m_3 h \leq \sum_{-\mu_k \in R(h,0)} |b_k|^2 \leq M_3 h. \tag{3.13}$$

Thus (3.3) holds by Remark 3.3, and hence (i) holds. The first inequality in (3.13) shows that $\alpha = -1/4$ cannot be increased.

The proof of (ii) is essentially the same. For (iii) and (iv), the eigenvalues lie in a vertical strip, and their imaginary parts possess a uniform asymptotic separation. From this, it is easy to show that (3.3) holds if and only if the sequence (b_k) in (3.3) is bounded. Estimates (A.5) of the Appendix show that Qb_0 corresponds to a sequence that is bounded and bounded away from zero. Hence (iv) follows, and $\alpha = 0$ is optimal. (A.5) also shows that AQb_1 corresponds to a sequence that is bounded and bounded away from zero. Hence $(AQb_1, \mathbb{G})$ is well-posed on Σ_0. From this, it follows that $(Qb_1, \mathbb{G})$ is well-posed on Σ_1 and that $\alpha = 1$ is optimal. □

The next two lemmas relate the spaces Λ_α and Σ_α to the spaces S_α and C_α.

LEMMA 3.8. *Let S_α and C_α be defined by* (1.9), (1.10) *and assume that $-1 \leq \alpha \leq 1$. Then*

$$\mathcal{H}_\alpha = \Lambda_\alpha \oplus \Sigma_\alpha = S_\alpha \times C_\alpha \times S_{2\alpha} \tag{3.14}$$

with equivalent norms. Furthermore,

$$\Lambda_\alpha \subset S_{2+2\alpha} \times C_{1+2\alpha} \times S_{2\alpha}, \tag{3.15}$$

$$\Sigma_\alpha \subset S_\alpha \times C_\alpha \times S_{1+\alpha}. \tag{3.16}$$

Finally, the mapping $P_{12} : \Sigma_\alpha \to S_\alpha \times C_\alpha$ given by $P_{12}x = (x_1, x_2)$ and the mapping $P_3 : \Lambda_\alpha \to S_{2\alpha}$ defined by $P_3 x = x_3$ are isomorphisms.

The proof relies upon asymptotic properties of the eigenvectors and is given in the Appendix.

LEMMA 3.9. *Let $|\alpha| \leq 1$, $|\beta| \leq 1$, and $|\beta - 2\alpha| < 1$. The following set-equivalences hold:*

$$\Lambda_\alpha + \Sigma_\beta = S_\beta \times C_\beta \times S_{2\alpha}. \tag{3.17}$$

In particular,

$$\Lambda_{-1/4} + \Sigma_0 = S_0 \times C_0 \times S_{-1/2}, \tag{3.18}$$

$$\Lambda_{1/4} + \Sigma_1 = S_1 \times C_1 \times S_{1/2}. \tag{3.19}$$

Proof. It suffices to prove (3.18). The proof of the general case is done in the same way. If $y = x+z$ with $x \in \Lambda_{-1/4}$ and $z \in \Sigma_0$, then, by Lemma 3.8, $y \in S_0 \times C_0 \times S_{-1/2}$. Thus $\Lambda_{-1/4}+\Sigma_0 \subset S_0 \times C_0 \times S_{-1/2}$. Now let $y \equiv (y_1, y_2, y_3)' \in S_0 \times C_0 \times S_{-1/2}$. Let P_3 denote the operator defined in Lemma 3.8 and assume that $\tilde{y} \equiv (\tilde{y}_1, \tilde{y}_2, \tilde{y}_3)' = P_3^{-1} y_3$. Then $P_3 \tilde{y} = \tilde{y}_3 = y_3$, and, by (3.15), $\tilde{y}_1 \in S_{3/2}$ and $\tilde{y}_2 \in C_{1/2}$. Let $x = (y_1 - \tilde{y}_1, y_2 - \tilde{y}_2, 0)'$. Then $y = x + \tilde{y}$ and $x \in S_0 \times C_0 \times S_0 = \mathcal{H}_0 = \Lambda_0 + \Sigma_0$. Also, however, $\tilde{y} \in \Lambda_{-1/4}$; hence $y \in \Lambda_{-1/4} + \Sigma_0$. This proves (3.18). □

Proof of Theorem 1.1. We first examine the case where $f \equiv 0$ in (3.10), (3.11). Let $t \geq 0$. By Proposition 3.7, $x(t) \in \Lambda_{-1/4}$, $z(t) \in \Sigma_0$, and the indices $-1/4$ and 0 are optimal in that they may not be increased. Thus, by Lemma 3.9, $y(t) = x(t) + z(t) \in S_0 \times C_0 \times S_{-1/2}$. Furthermore, (3.17) implies that the index $-1/2$ is optimal, and *not both* of the first two indices (0 and 0) may be increased. A sufficient condition that both the first two indices cannot be increased beyond 0 is that the operator $\mathcal{G}_T : L^2(0,\infty) \to \Sigma_0$ given by

$$\mathcal{G}_T u = \int_0^T \mathbb{G}_{T-s} Q b_0 u(s)\, ds$$

map onto a subspace of finite codimension for sufficiently large T. (Indeed, if this is so and if P_{12} represents the projection operator defined in Lemma 3.8, then, by Lemma 3.8, $P_{12}\mathcal{G}_T$ cannot map into any of the spaces $S_\alpha \times C_\beta$ with $\alpha > 0$ or $\beta > 0$.) In the case where $\gamma \leq 1$, it is shown in the proof of Proposition 5.1 that $\mathcal{G}_T$ is surjective (for large enough T). If $\gamma > 1$, the possibility of multiple eigenvalues arises; however, the same proof shows that $\mathcal{G}_T$ maps onto a subspace of finite codimension. Hence the trajectories $y(t)$ are time-continuous (see §3) with values in $S_0 \times C_0 \times S_{-1/2}$, and (pending the proof of Proposition 5.1) each of the indices are optimal in that none may be increased.

For the case with $g \equiv 0$ in (3.10), (3.11), we have $x(t) \in \Lambda_{1/4}$ and $z(t) \in \Sigma_1$. Thus, by (3.19), $y(t) \in S_1 \times C_1 \times S_{1/2}$, and, as in the previous case, the indices can be shown to be optimal. □

4. A moment problem of mixed parabolic-hyperbolic type. As we see in §5, the problem of controlling (3.9) from an initial state to a terminal state is equivalent to solving an associated moment problem of the following form:

$$c_k = \int_0^T e^{\mu_k s} u(s) ds, \qquad k \in \mathbb{N}, \tag{4.1}$$

$$d_k = \int_0^T e^{\sigma_k s} u(s) ds, \qquad k \in \mathbb{Z}. \tag{4.2}$$

The space of all sequences $(c_k) \cup (d_k)$ for which there exists some $u \in L^2[0,T]$ such that (4.1), (4.2) holds is called the *moment space* of (4.1), (4.2). While the individual

moment spaces of (4.1) and (4.2) are rather well understood, one cannot directly use these results to infer properties of the (joint) moment space of (4.1), (4.2). The main purpose of this section is to show that the moment space of (4.1), (4.2) is the union of of the individual moment spaces for (4.1) and (4.2), provided that T is greater than some nominal value t_c that depends upon the sequence (σ_k).

Because the results of this section pertain to a variety of sequences (σ_k), (μ_k) more general than those defined by (2.1), (2.2), throughout this section, we consider (4.1), (4.2) with the following general assumptions on the exponents (σ_k), (μ_k).

Assumption H0. $\{(\sigma_k)\}_{k\in\mathbb{Z}} \cap \{(\mu_k)\}_{k\in\mathbb{N}} = \emptyset$.

Assumption H1. There exists $\beta \in \mathbb{C}$, $c > 0$, and $(\nu_k)_{k\in\mathbb{Z}} \in l^2$ for which (σ_k) satisfies

(i) $\sigma_k = \beta + ck\pi i + \nu_k$ for all $k \in \mathbb{Z}$,
(ii) $\sigma_k \neq \sigma_j$ unless $j = k$.

Assumption H2. There exist positive ρ, B, δ, ε, and $0 \leq \theta < \pi/2$ for which (μ_k) satisfies

(i) $|\arg(-\mu_k)| \leq \theta$ for all $k \in \mathbb{N}$,
(ii) $|\mu_k - \mu_j| \geq \delta|k^2 - j^2|$ for all $k, j \in \mathbb{N}$,
(iii) $\varepsilon(\rho + Bk^2) \leq |\mu_k| \leq \rho + Bk^2$ for all $k \in \mathbb{N}$.

Assumptions H0, H1, and H2 are considered standing assumptions for all the results of this section.

Eigenvalues associated with one-dimensional hyperbolic systems often satisfy Assumption H1, while those of one-dimensional parabolic (or "abstract parabolic") systems often satisfy Assumption H2. The quadratic growth and separation assumptions in Assumption H2, parts (ii) and (iii), can be replaced by more general growth rates (see [7, Thm. 1.1]); however, we avoid this additional complication here.

It is convenient to introduce a notation for some spaces that we need to use. For $0 \leq a < b$, let

$$W_{[a,b]} = \text{closed span } (e^{\sigma_k t}) \quad \text{in } L^2[a,b],$$
$$E_{[a,b]} = \text{closed span } (e^{-\mu_k t}) \quad \text{in } L^2[a,b].$$

With $\|\cdot\|_{[a,b]} := \|\cdot\|_{L^2[a,b]}$, $W_{[a,b]}$ and $E_{[a,b]}$ are Hilbert spaces.

DEFINITION 4.1. Let H be a Hilbert space with closed subspaces M and N. We say that M and N are *uniformly separated in* H if $M \cap N = \{0\}$ and their sum $M + N$ is H-closed.

Equivalently, the subspaces M and N are uniformly separated in H if and only if there exists $\delta > 0$ (called the minimum gap in Kato [11]) such that, for any $f \in M$ and $g \in N$, each of norm 1, that $\|f - g\| \geq \delta$. See [11] for details.

The following result is the main one of this section and allows us to decouple the moment problem (4.1), (4.2).

THEOREM 4.2. *Assume the standing hypothesis (Assumptions* H0, H1, *and* H2). *For each $T > 2/c$, the spaces $W_{[0,T]}$ and $E_{[0,T]}$ are uniformly separated. This does not hold for $T \leq 2/c$.*

The proof relies upon the several results that follow, and is given later in this section.

Throughout the following, we denote $t_c = 2/c$.

LEMMA 4.3. *For any $a \in \mathbb{R}$, $W_{[a,a+t_c]} = L^2[a, a + t_c]$. Furthermore, for $T \geq t_c$, $(e^{\sigma_k t})_{k\in\mathbb{Z}}$ forms a Riesz basis for each of the spaces $W_{[a,a+T]}$.*

Proof. The sequence $(\sigma_k)_{k\in\mathbb{Z}}$ lies in a vertical strip of $\mathbb{C}$, and $|\operatorname{Im}\sigma_k - ck\pi| \to 0$ as $|k| \to \infty$. This implies (see [21, p.196]) that there exists N such that $(e^{s_k t})_{k\in\mathbb{Z}}$ forms a Riesz basis for $L^2(a, a+t_c)$ for any $a \in \mathbb{R}$, where $s_k = \sigma_k$ if $|k| > N$ and $s_k = \beta + ck\pi i$ if $|k| \le N$. By [21, p.40] and [21, p. 129], a Riesz basis of exponentials for $L^2(a, a+t_c)$ is stable with respect to a change of finitely many exponentials (i.e., for $|k| \le N$, $e^{s_k t} \to e^{\sigma_k t}$). Therefore the first statement of the lemma holds, and the second is true for $T = t_c$. For any $N \in \mathbb{N}$, we can choose a sequence $(e^{\tilde{s}_k t})_{k\in\mathbb{Z}}$ for which $|\operatorname{Im}\tilde{s}_k - ck\pi/N| \to 0$ as $|k \to \infty|$ and (σ_k) is a subsequence of $(\tilde{s}_k)$. As in the proof of the first statement, it follows that $(e^{\tilde{s}_k t})$ forms a Riesz basis for $L^2(a, a+Nt_c)$, for any $a \in \mathbb{R}$. Since a subset of a Riesz basis is necessarily a Riesz basis for the subspace given by its closed span, it follows that $(e^{\sigma_k t})$ forms a Riesz basis for $W_{[a,a+Nt_c]}$, for any $a \in \mathbb{R}$. Thus the second statement of the lemma is true for $T = Nt_c$ for any $N \in \mathbb{N}$. Let $t_c \le T \le Nt_c$. By [21, p.32], $(e^{\sigma_k t})$ forms a Riesz basis for $W_{[a,a+T]}$ if and only if there exist positive numbers m_T and M_T such that, for any $n \in \mathbb{N}$ and arbitrary scalars $c_1, c_2, \ldots, c_n$, we have

$$m_T\|(c_i)\|_{l^2}^2 \le \left\|\sum_{i=1}^{n} c_i e^{\sigma_i t}\right\|_{[a,a+T]}^2 \le M_T\|(c_i)\|_{l^2}^2. \tag{4.3}$$

Let $p_n(t) = \sum_{i=1}^{n} c_i e^{\sigma_i t}$. Since $[a, a+t_c] \subset [a, a+T] \subset [a, a+Nt_c]$, it follows that

$$\|p_n\|_{[a,a+t_c]}^2 \le \|p_n\|_{[a,a+T]}^2 \le \|p_n\|_{[a,a+Nt_c]}^2.$$

Furthermore, (4.3) holds if $T = t_c$ or if $T = Nt_c$. It thus follows that, for arbitrary $T \in (t_c, Nt_c)$, the inequalities in (4.3) hold with $m_T = m_{t_c}$ and $M_T = M_{Nt_c}$. □

The previous lemma implies that, for each $f \in W_{[a,a+T]}$, with $a \in \mathbb{R}$ and $T \ge t_c$, there is a uniquely defined sequence $(c_k) \in l^2$ for which

$$f = \sum_{k\in\mathbb{Z}} c_k e^{\sigma_k t}, \qquad t \in [a, a+T]. \tag{4.4}$$

Thus given any $f \in W_{[a,a+T]}$, we may define an extension $\tilde{f} \in L^2_{\mathrm{loc}}(\mathbb{R})$ by

$$\tilde{f} = \sum_{k\in\mathbb{Z}} c_k e^{\sigma_k t}, \qquad t \in \mathbb{R}. \tag{4.5}$$

LEMMA 4.4. *Let $a, b \in \mathbb{R}$ and assume that $T \ge t_c$. Then the mapping F: $W_{[a,a+t_c]} \to W_{[b,b+T]}$ defined by*

$$Ff = \tilde{f}\Big|_{[b,b+T]}$$

is an isomorphism.

Proof. For $\alpha, \beta \in \mathbb{R}$, with $\beta \ge \alpha + t_c$, let $J_{[\alpha,\beta]}: W_{[\alpha,\beta]} \to l^2$ by

$$J_{[\alpha,\beta]} f = (c_k)_{k\in\mathbb{Z}},$$

where (c_k) is determined by (4.4). Lemma 4.3 implies that $J_{[\alpha,\beta]}$ is an isomorphism. Therefore $F = J_{[b,b+T]}^{-1} J_{[a,a+t_c]}$ is an isomorphism as well. □

For $a, s \in \mathbb{R}$ define $J_a(s) \in \mathcal{L}(L^2[a, a+t_c])$ by

$$(J_a(s)f)(t) = \tilde{f}(t+s)\,\big|_{t\in[a,a+t_c]}, \tag{4.6}$$

where f and $\tilde{f}$ are given by (4.4) and (4.5). (Lemmas 4.3 and 4.4 show that J_a is well-defined.) By Lemma 4.4, for any $a, s \in \mathbb{R}$, $J_a(s)$ is an isomorphism. In fact, it can be seen that $J_a = (J_a(s))_{s\in\mathbb{R}}$ forms a group. The generator of this group was characterized in Russell [20].

PROPOSITION 4.5. *For any $a \in \mathbb{R}$, $J_a = (J_a(s))_{s\in\mathbb{R}}$ is a strongly continuous group of operators on $L^2[a, a+t_c]$. J_a is generated by the derivative operator d/dt on the domain*

$$\mathcal{D}\left(\frac{d}{dt}\right) = \left\{ f \in H^1[a, a+t_c] \,|\, f(a+t_c) - e^{\beta t_c} f(a) = \int_a^{a+t_c} q(\tau) f(\tau) d\tau \right\}, \tag{4.7}$$

where $q \in L^2(a, a+t_c)$ is uniquely determined by a and (σ_k) and satisfies $\|q\| \leq M_a \|(\nu_k)\|_{l^2}$ for some $M_a > 0$.

An explicit formula for q can be found in [20]. This, however, does not concern us.

The next result concerns properties of the spaces $E_{[0,T]}$.

PROPOSITION 4.6. *Let $0 < \alpha < \pi/2 - \theta$ (θ is defined in Assumption* H2*) and assume that T and ν are positive. Each $f \in E_{[0,T]}$ has an analytic extension $\hat{f}$ to the the region $\Delta_\nu = \{\lambda \in \mathbb{C} \,|\, |\arg\lambda| < \alpha, |\lambda| > \nu\}$. Furthermore, there exist positive constants M, ω such that, for any $\lambda \in \Delta_\nu$,*

$$|\hat{f}(\lambda)| \leq M e^{-\omega\rho|\lambda|} \|f\|_{[0,T]} \quad \forall f \in E_{[0,T]}, \tag{4.8}$$

where M and ω depend only upon B, δ, ε, and θ (of (H2)).

This result was proved in more generality in [7].

A key point in the above proposition is that M and ω are independent of ρ (in Assumption H2) and the particular sequence $(\mu_k)_{k\in\mathbb{N}}$. Thus, by selectively removing a finite number of μ_k (those with the biggest real parts) from (μ_k), we can increase the ρ defined in Assumption H2 without affecting the constants M and ω and hence obtain any desired decay rate in (4.8) for the functions generated by such a subsequence of exponentials.

We restate this key point as follows.

COROLLARY 4.7. *Let r and T be positive numbers. The space $E_{[0,T]}$ can be decomposed into the direct sum $F \oplus R$, where F is finite-dimensional and all functions $f \in R$ have an analytic continuation $\hat{f}(z)$, which satisfies*

$$|\hat{f}(z)| < M e^{-r|z|} \|f\|_{[0,T]} \quad \forall z \in \Delta_\nu,\ f \in R. \tag{4.9}$$

Proof of Theorem 4.2. Let $\epsilon > 0$. We wish to show that the spaces $W_{[0,t_c+\varepsilon]}$ and $E_{[0,t_c+\varepsilon]}$ are uniformly separated. Let $J_{\varepsilon/2} = \left(J_{\varepsilon/2}(s)\right)_{s\in\mathbb{R}}$ denote the group defined by (4.6) and Proposition 4.5. Since groups are obviously invertible, there exist $m > 0$ and $r_0 > 0$ for which

$$\|J_{\varepsilon/2}(s) f\| > m e^{-r_0 s} \|f\| \quad \forall\, s > 0,\ f \in L^2(\varepsilon/2, t_c + \varepsilon/2).$$

It follows from the above and Lemma 4.4 that there exists $\tilde{m} > 0$ for which

$$\|\tilde{f}\|_{[s,s+t_c]} > \tilde{m} e^{-r_0 s} \|f\| \quad \forall\, s > 0,\ f \in W_{[\varepsilon/2, t_c+\varepsilon]}, \tag{4.10}$$

where $\tilde{f}$ is defined by (4.5).

Let $r > r_0$. We may without loss of generality assume that all functions in $E_{[0,t_c+\varepsilon]}$ have analytic continuations that satisfy (4.9). (This follows, since $E_{[0,t_c+\varepsilon]}$ could be decomposed as in Corollary 4.7, and, since $(\mu_k) \cup (\sigma_k)$ are distinct, F is necessarily uniformly separated from $W_{[0,t_c+\varepsilon]}$.)

Now assume, to the contrary, that the two spaces are not uniformly separated. Then there exists $(f_n)_{n\in\mathbb{N}} \in E_{[0,t_c+\varepsilon]}$ and $(g_n)_{n\in\mathbb{N}} \in W_{[0,t_c+\varepsilon]}$, each of norm 1 for which

$$\|f_n - g_n\|_{[0,t_c+\varepsilon]} \to 0 \quad \text{as } n \to \infty.$$

Since each f_n has an analytic continuation $\hat{f}_n$ that satisfies the bound (4.9) (use $T = t_c + \varepsilon$ and the M determined by $\nu = \varepsilon/2$), it follows that $(\hat{f}_n)$ forms a normal family on compact subsets of $\Delta_0 = \{z \in \mathbb{C} \,|\, z \neq 0,\ |\arg z| < \alpha\}$. It follows that there exists a subsequence, still denoted $(\hat{f}_n)$, which converges uniformly on the interval $I = [\varepsilon/2,\, t_c + \varepsilon]$ to $f(t)$. Since obviously $f \in L^2(I)$, we know that $\|g_n - f\|_I \to 0$ as $n \to \infty$. Thus $f \in W_I$, and, by Vitali's convergence theorem [9], f has an analytic continuation $\hat{f}$ to Δ_0 for which

$$|\hat{f}(t)| \leq Me^{-rt} \quad \forall t \geq \epsilon/2. \tag{4.11}$$

Assume for the moment that $\|f\|_I = 0$. It then follows that $\|g_n\|_I \to 0$, and consequently $\|g_n\|_{[0,\varepsilon/2]} \to 1$ as $n \to \infty$. This, however, is impossible by Lemma 4.4. Thus $\|f\|_I > 0$.

Let $F(s,t) = \hat{f}(s+t)$. Since $\hat{f}$ is differentiable, it follows that $\partial F/\partial s = \partial F/\partial t$ for $s + t > \varepsilon/2$. Furthermore, $\hat{f}\,|_I = \tilde{f}\,|_I \in W_I$. Thus, for all $s \in (0, \varepsilon/2)$,

$$F(s, t_c + \varepsilon/2) - e^{\beta t_c} F(s, \varepsilon/2) = \int_{\varepsilon/2}^{t_c+\varepsilon/2} q(\tau) F(s,\tau)\, d\tau. \tag{4.12}$$

Morera's theorem can be used to show that the the right-hand side of (4.12) is analytic in Δ_0. Since the left-hand side of (4.12) is also analytic in this region, we conclude that (4.12) holds for all $s \in (\varepsilon/2, \infty)$. Hence right-translations of $\hat{f}$ are given by $J_{\varepsilon/2}$ (in Proposition 4.5) acting upon f, as are those of $\tilde{f}$; i.e., for $s > 0$,

$$\hat{f}\,|_{[s+\varepsilon/2,\, s+t_c+\varepsilon/2]} = J_{\varepsilon/2}(s)\left(f\,|_{[\varepsilon/2,\, \varepsilon/2+t_c]}\right) = \tilde{f}\,|_{[s+\varepsilon/2,\, s+t_c+\varepsilon/2]}.$$

It thus follows that $\tilde{f}(t) = \hat{f}(t)$ for $t > \varepsilon/2$, but this is in conflict with (4.10) and (4.11). □

The following results relate Theorem 4.2 to the moment problem (4.1), (4.2).

PROPOSITION 4.8. *Let $(d_k)_{k\in\mathbb{Z}} \in l^2$. Then, for any $T \geq t_c$, there is a unique $u \in W_{[0,T]}$, which solves the moment problem* (4.2). *Any $f \in L^2[0,T]$ given by $f = u + v$ with $v \in W^{\perp}_{[0,T]}$ also solves* (4.2).

Proof. This follows easily from Lemma 4.3. □

PROPOSITION 4.9. *Assume that, for any $p > 0$, $(c_k)_{k\in\mathbb{N}}$ satisfies*

$$|c_k| e^{pk} \to 0 \quad \text{as } k \to \infty. \tag{4.13}$$

Then, given any $\tau > 0$, there exists a unique $u \in E_{[0,\tau]}$, which solves the moment problem (4.1). *Any $f \in L^2[0,\tau]$ given by $f = u + v$ with $v \in E_{[0,\tau]}^{\perp}$ also solves* (4.1).

Proof. From [7, Thm. 1.1], there is a $p_0 > 0$ for which the biorthonormal functions $(q_k(t))$ to $(e^{\mu_k t})$ in $E_{[0,\tau]}$ satisfy

$$\|q_k\|_{[0,\tau]} \leq M e^{p_0 k},$$

for some $M > 0$. We define $u = \sum_{k=1}^{\infty} c_k q_k$. It is easily checked that $u \in E_{[0,\tau]}$, and both u and f solve (4.1). □

Remark 4.10. Condition (4.13) can be weakened to

$$|c_k| \cdot e^{p_0 k} \to 0 \quad \text{as } k \to \infty,$$

where a suitable p_0 can be found from [7, Thm. 1.1].

THEOREM 4.11. *Under the standing hypothesis* (*Assumptions* H0, H1, H2), *given any sequence $(c_k)_{k\in\mathbb{N}}$ that satisfies* (4.13) *and any $(d_k)_{k\in\mathbb{Z}} \in l^2$, for any time $T > t_c$, there exists $u \in L^2(0,T)$, which simultaneously solves the moment problems* (4.1) *and* (4.2). *This does not hold for $T \leq t_c$.*

Proof. If $T = t_c$, the solution to (4.2) is unique (this follows from Lemma 4.3), and hence it is not in general possible to simultaneously solve (4.1) and (4.2). If $T < t_c$, then (4.2) does not necessarily have a solution, and hence it is necessary that $T \geq t_c$. Thus assume that $T > t_c$. By Theorem 4.2, the spaces $E \equiv E_{[0,T]}$ and $W \equiv W_{[0,T]}$ are uniformly separated. Thus $V := E + W$ is closed, and hence a Hilbert space with $\|\cdot\|_V = \|\cdot\|_{[0,T]}$. (So $V = E \oplus W$.) Let $E^{\perp}$, $W^{\perp}$ denote the orthogonal complements of E, W in V. Let P_E denote the orthogonal projection from V onto E. By a theorem in Kato [11, Chap. 4, §4], $E^{\perp}$ and $W^{\perp}$ are also uniformly separated, and hence $V = E^{\perp} \oplus W^{\perp}$. From this, it is easy to show that (the restriction) $P_E|_{W^{\perp}}$ is an isomorphism. Likewise, we may define an orthogonal projection P_W for which $P_W|_{E^{\perp}}$ is an isomorphism. By Propositions 4.8 and 4.9, there exist $g \in W_{[0,T]}$, which solves (4.2), and $f \in E_{[0,T]}$, which solves (4.1). Let

$$u = (P_E|_{W^{\perp}})^{-1} f + (P_W|_{E^{\perp}})^{-1} g.$$

We easily see that u solves both (4.1) and (4.2), and, since $P_E|_{W^{\perp}}$ and $P_W|_{E^{\perp}}$ are isomorphisms, $u \in L^2[0,T]$. □

Remark 4.12. If, in addition to the hypothesis of Theorem 4.11, it is known that $(d_k\sigma_k)_{k\in\mathbb{Z}} \in l^2$, then the solution u of (4.1), (4.2) may be assumed to satisfy $u(0) = 0$ and have a (distributional) derivative in L^2. This can be proved by a modification of a result in [4]. However, without any preconditions on (d_k), it can be shown that there do not in general exist smooth solutions to (4.1), (4.2) regardless of how large T is.

5. Controllability.

Consider

$$\dot{y}(t) = Ay(t) + bu(T-t), \quad 0 < t < T; \qquad y(0) = y^0, \tag{5.1}$$

where A is defined in (1.8), $u \in L^2[0,T]$, b represents b_0 or b_1 in (3.7), and y^0 belongs to an appropriate space that we specify later. If we wish to control the state to some

terminal state y^T in time T, the variation of parameters formula must hold (on an appropriate space) as follows:

$$y^T - \mathbb{T}_T y^0 = \int_0^T \mathbb{T}_s b v(s) ds.$$

Using the same decomposition as in (3.10), (3.11), we must have

$$x^T - \mathbb{S}_T x^0 = \int_0^T \mathbb{S}_\tau P b u(\tau) d\tau, \tag{5.2}$$

$$z^T - \mathbb{G}_T z^0 = \int_0^T \mathbb{G}_\tau Q b u(\tau) d\tau, \tag{5.3}$$

where $x = Py$, $z = Qy$, and likewise for x^T and z^T. So that the solution to (5.1) exists pointwise in time, we require that (5.2) and (5.3) hold on the respective spaces in which $(\mathbb{S}, Pb)$ and $(\mathbb{G}, Qb)$ are well-posed. Thus, if b represents b_0 (respectively, b_1), then (5.2) should hold on $\Lambda_{-1/4}$ (respectively, $\Lambda_{1/4}$), and (5.3) should hold on Σ_0 (respectively, Σ_1).

When (5.2) and (5.3) are integrated against the eigenfunctions of A^*, we arrive at the pair of coupled moment problems (4.1), (4.2), where (σ_k) and (μ_k) are defined by (2.1) and

$$c_k = \frac{\langle x^T, \psi_{\mu_k}\rangle - e^{\mu_k T}\langle x^0, \psi_{\mu_k}\rangle}{\langle b, \psi_{\mu_k}\rangle}, \qquad d_k = \frac{\langle z^T, \psi_{\sigma_k}\rangle - e^{\sigma_k T}\langle z^0, \psi_{\sigma_k}\rangle}{\langle b, \psi_{\sigma_k}\rangle}. \tag{5.4}$$

The sequences $(\langle b, \psi_{\mu_k}\rangle)_{k\in\mathbb{N}}$ and $(\langle b, \psi_{\sigma_k}\rangle)_{k\in\mathbb{Z}}$ each consist of only nonzero terms, and their asymptotic properties are given in (A.5) in the Appendix.

We easily see from (2.2) that Assumptions H0, H1, and H2 of the previous section are satisfied, provided that there are no multiple eigenvalues.

To describe the controllability of (5.1), we consider separately the problems of null-controllability and reachability. We say that a $\mathbb{T}$-invariant space M_0 is *b-null-controllable in time T* if, given any $y^0 \in M_0$, there exists $u \in L^2(0,T)$ for which (5.2), (5.3) hold (on the proper spaces) with $x^T = z^T = 0$. Likewise, we say that a $\mathbb{T}$-invariant space M_T is *b-reachable in time T* if, given any $y^T \in M_T$, (5.2) and (5.3) hold with $x^0 = z^0 = 0$.

We have the following result.

PROPOSITION 5.1. *Let $T > 2/c$, $\alpha \in \mathbb{R}$, and $0 < \gamma \le 1$.*

(i) *The space $\Sigma_0 + \Lambda_\alpha$ is b_0-null-controllable in time T, and $\Sigma_1 + \Lambda_\alpha$ is b_1-null-controllable in time T.*

(ii) *Let $V = \{\sum_{k\in\mathbb{N}} c_k \phi_{\mu_k} \mid (c_k)$ satisfies (4.13)$\}$. The space $\Sigma_0 + V$ is b_0-reachable in time T, and $\Sigma_1 + V$ is b_1-reachable in time T.*

In either case, the result does not remain true for $T \le 2/c$ or if Σ_0 or Σ_1 are replaced by larger $\mathbb{G}$-invariant spaces.

Proof. Let us first prove that $\Sigma_0 + \Lambda_\alpha$ is b_0-null-controllable in time T. (This is the case where the temperature is controlled at the left end of the rod.) Let $y^0 = x^0 + z^0$ with $x^0 \in \Lambda_\alpha$ and $z^0 \in \Sigma_0$. Since

$$z^0 = -\mathbb{G}_T^{-1} \int_0^T \mathbb{G}_\tau Q b v(\tau)\, d\tau,$$

it is necessary (since $\mathbb{G}_T$ is an isomorphism on Σ_0) that $z^0 \in \Sigma_0$ for (5.3) to hold on Σ_0. Thus Σ_0 cannot be replaced by a larger $\mathbb{G}$-invariant space. With $x^0 \in \Lambda_\alpha$, it follows from the analyticity of $\mathbb{S}$ that $\mathbb{S}_T x^0 \in \Lambda_0$. Hence, if the moment problem determined by (5.2), (5.3) has a solution, then (5.2) and (5.3) hold on the appropriate spaces $\Lambda_{-1/4}$ and Σ_0, respectively. From Lemma A.1 (in the Appendix) and (2.2), we can easily see that the eigenvalues (σ_k), (μ_k) satisfy Assumptions (H0), (H1), and (H2) of §4. To compute (c_k) and (d_k) in (4.1), (4.2), we use (5.4) and (A.5) (in the Appendix) and find that there are positive numbers m and M for which

$$m|\langle z^0, \psi_{\sigma_k}\rangle| \le |d_k| \le M|\langle z^0, \psi_{\sigma_k}\rangle| \quad \forall k \in \mathbb{Z}, \tag{5.5}$$

$$|c_k| \le Mk^{2\alpha} e^{\mu_k T} \|x^0\|_{\Lambda_\alpha}. \tag{5.6}$$

Thus $(d_k) \in l^2$ and (c_k) satisfies (4.13). Hence, by Theorem 4.11, the moment problem has a solution for $T > 2/c$ (but not in general for $T \le 2/c$). The proof that $\Lambda_\alpha + \Sigma_1$ is b_1-null-controllable is essentially the same. Thus (i) holds.

For the problem of reachability, first note that, if $y^T = x^T + z^T$ with x^T and z^T as in the hypothesis, then (5.2) and (5.3) hold on the proper spaces, provided that the moment problem has a solution. The moment problem that corresponds to (5.3) is easily seen to satisfy (5.5) and hence is solvable for any $(d_k) \in l^2$. Similarily, with $x^T \in V$, it is easily seen that the coefficients (c_k) satisfy (4.13). Thus (ii) holds by Theorem 4.11. □

More general statements can be made about the reachable space for the parabolic component (see [4]).

The proof of Theorem 1.2 now easily follows.

Proof of Theorem 1.2. Let $T > 2/c$. First, consider (1.5), (1.6) with $f(t) \equiv 0$. (This is equivalent to (5.1) with $b = b_0$.) Since $y^0 \in \mathcal{H}$, certainly we have that $y^0 \in \Sigma_0 + \Lambda_{-1/4}$. By Proposition 5.1, there exists $u \in L^2(0, 2/c + \epsilon)$ for which (5.2) and (5.3) hold with $x^T = z^T = 0$. Since $x \in C([0,T], \Lambda_{-1/4})$ and $z \in C([0,T], \Sigma_0)$, it follows from Lemma 3.9 that $y = x + z \in C([0,T], S_0 \times C_0 \times S_{-1/2})$. Part (ii) of Theorem 1.2 is proved likewise.

Remark 5.2. Theorem 1.2 is optimal in a couple of respects. In part (i), by Proposition 5.1 and Lemma 3.9, the space $\mathcal{H} = S_0 \times C_0 \times S_0$ is the largest null-controllable space of the form $S_\alpha \times C_\alpha \times S_0$. Likewise, in part (ii), the space $\mathcal{D}(A) = S_1 \times C_1 \times S_2$ is the largest null-controllable space of the form $S_\alpha \times C_\alpha \times S_2$. Similar statements can be made regarding reachability. Furthermore, by Theorem 1.1, the spatial regularity of the solutions given in Theorem 1.2 is optimal in the sense described in Theorem 1.1. (This means that, for general L^2-controls, no improvement in spatial regularity is possible. Of course, the spatial regularity can be improved if the controls are known to be smooth. However, by Remark 4.12, there does not in general exist smooth controls unless the initial/terminal spaces are restricted.)

Remark 5.3. Proposition 5.1 implies a certain *partial exact controllability* result. Namely, for the case of temperature control ($b = b_0$), given any $y^0 \in \mathcal{H}$ and any $z^T \in L^2(\Omega) \times L^2(\Omega)$, for any $\varepsilon > 0$, it is possible to find a control $g \in L^2(0, 2/c + \varepsilon)$ that transfers y^0 to a state y^T, which has z^T for its first two components. (The third component is not controlled.) More loosely stated, the mechanical components are exactly controllable on $(L^2(\Omega))^2$ in time $T = 2/c + \varepsilon$. Likewise for the case of heat flux control, the mechanical components are exactly controllable on $S_1 \times C_1$ in time T.

The above asserts that it is possible to exactly control the mechanics (position, velocity) of the rod with temperature (or heat flux) control alone. Furthermore, Theorem 1.2 shows that null-controllability of the whole state space (position, velocity, temperature) is possible. We could ask whether it is possible to drive an initial state y^0 to a terminal state of the form $y^T = (y_1^T, y_2^T, 0)'$. As the following shows, this is not generally possible without some severe restrictions.

NEGATIVE RESULT 5.4. *Let b denote b_0 or b_1. For any $n > 0$, there exists $y_n \in S_n \times C_n$ for which the state $y^T = (y_n, 0)$ is not b-reachable in any time $T > 0$.*

Sketch of proof. For $n > 0$, let $P_{12} : \mathcal{H}_n \to S_n \times C_n$ by $(y_1, y_2, y_3) \to (y_1, y_2)$ and define $M_n = P_{12}\Lambda_n$. If the space $M_n \times \{0\}$ were b_1-reachable, then the corresponding moment problems must necessarily have solutions. Hence the set $\mathcal{C}$ of sequences (c_k) corresponding to PM_n should be in the moment space of (4.1). (P is the projection in (5.2).) From (5.4) and estimates in the Appendix, it follows that there exists N (which depends upon n) such that, if

$$\sum_{k\in\mathbb{N}} |c_k| k^N < \infty, \tag{5.7}$$

then $(c_k) \in \mathcal{C}$. Let $(q_k)_{k\in\mathbb{N}}$ denote the biorthonormal sequence to $(\exp(\mu_j t))_{k\in\mathbb{N}}$ in $E_{[0,T]}$. ($E_{[0,T]}$ was defined in §4.) It is known [7] that $\|q_k\| \geq m_1 e^{m_0 k}$ for some $m_0 > 0$, $m_1 > 0$. The solution to (4.1) is given by

$$u = \sum_{k\in\mathbb{N}} c_k q_k$$

and must converge for all (c_k) in the moment space. However, there are clearly many sequences (c_k) satisfying (5.7) for which $\|c_k q_k\| \to \infty$ as $k \to \infty$. □

As mentioned in the Introduction, results similar to Theorems 1.1 and 1.2 apply if the stress or velocity are controlled instead of the temperature or heat flux at an endpoint. For example, consider the boundary control problem (1.5), (1.6), with the boundary conditions

$$\begin{aligned} \sigma_{y(t)}(0) &= g(t), \quad \theta_{y(t)}(0) = 0, \quad t \geq 0, \\ v_{y(t)}(1) &= f(t), \quad q_{y(t)}(1) = 0, \quad t \geq 0. \end{aligned} \tag{5.8}$$

This system can be shown to be equivalent to

$$\frac{dy}{dt} = Ay(t) + b_\sigma f(t) + b_v g(t), \qquad y(0) = y^0, \tag{5.9}$$

where A is defined by (1.8) and where the input elements b_σ and b_v are defined by

$$\begin{aligned} \langle b_\sigma, \bar{z}\rangle &= -v_z(0) = -z_2(0) \quad \forall\, z = (z_1, z_2, z_3)' \in \mathcal{H}_1, \\ \langle b_v, \bar{z}\rangle &= \sigma_z(1) = z_1(1) - \gamma z_3(1) \quad \forall\, z = (z_1, z_2, z_3)' \in \mathcal{H}_1. \end{aligned}$$

Hence (5.8), (5.9) can be analyzed in the same manner as (3.7), (3.9). In this way, we can obtain the following results, which we state without proof.

PROPOSITION 5.5. *Let $y^0 = 0$, $f \in L^2(0,\infty)$, and $g \in L^2(0,\infty)$. Then the solution to* (1.5), (1.6), (5.8) *belongs to $C([0,\infty), S_0 \times C_0 \times S_{1/2})$. If, additionally, $g \equiv 0$, then the solution belongs to $C([0,\infty), S_0 \times C_0 \times S_1)$. These solution spaces are optimal in the sense that none of the indices $\{0, 1/2, 1\}$ may be increased.*

PROPOSITION 5.6. *Assume that* $0 < \gamma \leq 1$ *in* (1.5) *and* $T > 2/c$.

(i) *For the boundary control problem* (1.5), (1.6), (5.8), *with* $f \equiv 0$, *any* $y^0 \in \mathcal{H}$ *can be controlled to zero by some* $g \in L^2[0,T]$. *The resulting solution is in* $C((0,T], S_0 \times C_0 \times S_{1/2}) \cap C([0,T], \mathcal{H})$.

(ii) *For the boundary control problem* (1.5), (1.6), (5.8), *with* $g \equiv 0$, *any* $y^0 \in \mathcal{H}$ *can be controlled to zero by some* $f \in L^2[0,T]$. *The resulting solution is in* $C((0,T], S_0 \times C_0 \times S_1) \cap C([0,T], \mathcal{H})$.

In Propositions 5.5 and 5.6, the control time and all the spaces involved can be shown to be optimal in the same sense as those of Theorems 1.1 and 1.2.

Appendix. As described in §2, the eigenfunctions of A^* consist of a real branch $(\mu_k)_{k\in\mathbb{N}}$ and a nonreal branch $(\sigma_k)_{k\in\mathbb{Z}}$, which are determined by the characteristic equations (2.1) and satisfy the asymptotic estimates (2.2). The nonreal branch consists of complex conjugate pairs for which

$$\bar{\sigma}_k = \sigma_{-k+1} \quad \forall k \in \mathbb{N}. \tag{A.1}$$

Let $r_k = k\pi - \pi/2$ for $k \in \mathbb{N}$. For $k \in \mathbb{N}$, the associated eigenfunctions are given by (see [6])

$$\psi_{\sigma_k} = \begin{pmatrix} \sin r_k x \\ \dfrac{\sigma_k}{r_k} \cos r_k x \\ \dfrac{-\gamma\sigma_k}{\sigma_k + r_k^2} \sin r_k x \end{pmatrix}, \qquad \psi_{\mu_k} = \begin{pmatrix} \dfrac{c^2\gamma}{(\mu_k/r_k)^2 + c^2} \sin r_k x \\ \dfrac{(\mu_k/r_k)c^2\gamma}{(\mu_k/r_k)^2 + c^2} \cos r_k x \\ \sin r_k x \end{pmatrix}. \tag{A.2}$$

For $k \leq 0$, ψ_{σ_k} are given by conjugation, as in (A.1). The above eigenfunctions are not normalized (as was assumed in §2), but they are *almost normalized*; that is, their norms are bounded and bounded away from zero. Since all the estimates we derive here concern only the asymptotic order, the estimates that we obtain here remain valid for the normalized eigenfunctions of §2.

For a sequence $(c_k)_{k\in\mathbb{N}}$, let us say that $c_k = \mathcal{O}(k^\alpha)$ if there are positive numbers m and M for which $mk^\alpha \leq |c_k| \leq Mk^\alpha$. It can be seen from (2.1) and (2.2) that

$$\psi_{\sigma_k} = \begin{pmatrix} \mathcal{O}(1) \cdot \sin r_k x \\ \mathcal{O}(1) \cdot \cos r_k x \\ \mathcal{O}(k^{-1}) \cdot \sin r_k x \end{pmatrix}, \qquad \psi_{\mu_k} = \begin{pmatrix} \mathcal{O}(k^{-2}) \cdot \sin r_k x \\ \mathcal{O}(k^{-1}) \cdot \cos r_k x \\ \mathcal{O}(1) \cdot \sin r_k x \end{pmatrix}. \tag{A.3}$$

By [6, Rem. 3.3], the eigenfunctions of A likewise satisfy

$$\phi_{\sigma_k} = \begin{pmatrix} \mathcal{O}(1) \cdot \sin r_k x \\ \mathcal{O}(1) \cdot \cos r_k x \\ \mathcal{O}(k^{-1}) \cdot \sin r_k x \end{pmatrix}, \qquad \phi_{\mu_k} = \begin{pmatrix} \mathcal{O}(k^{-2}) \cdot \sin r_k x \\ \mathcal{O}(k^{-1}) \cdot \cos r_k x \\ \mathcal{O}(1) \cdot \sin r_k x \end{pmatrix}. \tag{A.4}$$

Let b_0 and b_1 denote the input elements defined by (3.7). From (A.2) and (A.3), we have

$$\begin{aligned} \langle b_0, \psi_{\sigma_k} \rangle &= \mathcal{O}(1), & \langle b_0, \psi_{\mu_k} \rangle &= \mathcal{O}(k), \\ \langle b_1, \psi_{\sigma_k} \rangle &= \mathcal{O}(k^{-1}), & \langle b_1, \psi_{\mu_k} \rangle &= \mathcal{O}(1). \end{aligned} \tag{A.5}$$

Proof of Lemma 3.8. The first equality in (3.14) is just (2.5). For the second, note that $\mathcal{H} = S_0 \times C_0 \times S_0$ and $\mathcal{H}_1 = S_1 \times C_1 \times S_2$. It follows from standard properties of interpolation spaces (e.g., [15]) that, for $\alpha \in [0,1]$,

$$\begin{aligned}\mathcal{H}_\alpha &= [\mathcal{H}_1, \mathcal{H}]_{1-\alpha} \\ &= [S_1 \times C_1 \times S_2, S_0 \times C_0 \times S_0]_{1-\alpha} \\ &= S_\alpha \times C_\alpha \times S_{2\alpha}.\end{aligned}$$

The above also holds for $\alpha \in [-1,0]$ by duality.

To prove (3.15), we first note from the eigenvalue estimates (2.1), for any $\alpha \in \mathbb{R}$,

$$\Lambda_\alpha = \left\{ \sum_{k=1}^{\infty} c_k \varphi_{\mu_k} \mid (c_k k^{2\alpha}) \in l^2 \right\}. \tag{A.6}$$

Thus, if $x = (x_1, x_2, x_3)' = \sum_{k=1}^{\infty} c_k \varphi_{\mu_k} \in \Lambda_\alpha$, then by (A.4) and (A.6)

$$\begin{aligned} x_1 &= \sum_{k=1}^{\infty} c_k \cdot \mathcal{O}(k^{-2}) \cdot \sin r_k x \ \in S_{2+2\alpha}, \\ x_2 &= \sum_{k=1}^{\infty} c_k \cdot \mathcal{O}(k^{-1}) \cdot \cos r_k x \ \in C_{1+2\alpha}, \\ x_3 &= \sum_{k=1}^{\infty} c_k \cdot \mathcal{O}(1) \cdot \sin r_k x \ \in S_{2\alpha}. \end{aligned}$$

Hence $\Lambda_\alpha \subset S_{2+2\alpha} \times C_{1+2\alpha} \times S_{2\alpha}$. Theorem 2.1 and (A.6) imply that $(k^{-2\alpha}\phi_{\mu_k})$ forms a Riesz basis for Λ_α. Hence an equivalent norm $|\cdot|$ on Λ_α is given by $|x| = \|(c_k k^{2\alpha})\|_{l^2} = \|P_3 x\|_{S_\alpha}$. Thus P_3 is an isomorphism from Λ_α to S_α. Similar arguments show that (3.16) holds and that $P_{12}\big|_{\Sigma_\alpha}$ is an isomorphism. □

LEMMA A.1. *Let A be defined by* (1.8). *For $0 \le \gamma \le 1$, the spectrum of A consists entirely of simple eigenvalues.*

Proof. For $k \in \mathbb{N}$, let $r_k = k\pi - \pi/2$. First, assume that for some $k \in \mathbb{N}$, the characteristic equation (2.1) has a double root; i.e., $p_k(x) \equiv (x^2 + c^2)(x + r_k) + \gamma^2 c^2 x$ can be written as

$$p_k(x) = (x + a)(x + b)^2,$$

where a and b are positive. (Any double root is clearly real, and the roots must be negative since A is dissipative.) Equating coefficients of the two polynomials leads to

$$1 + \gamma^2 = \frac{(a + 2b)(2a + b)}{ab},$$

which is impossible with $\gamma^2 \le 8$. Thus, if λ is a double eigenvalue, there exists distinct positive integers j, k such that $p_k(\lambda/r_k) = 0$ and $p_j(\lambda/r_j) = 0$. This can be written as

$$\lambda^3 + \lambda^2 r_k^2 + \lambda c^2(1 + \gamma^2) r_k^2 + c^2 r_k^4 = 0, \tag{A.7}$$

$$\lambda^3 + \lambda^2 r_j^2 + \lambda c^2(1 + \gamma^2) r_j^2 + c^2 r_j^4 = 0. \tag{A.8}$$

Let $G = 1 + \gamma^2$, $S = r_k^2 + r_j^2$, and $P = r_k^2 r_j^2$. By eliminating the λ^3 term and, respectively, the constant term in (A.7), (A.8), we find (using $\lambda \neq 0$) that

$$\lambda^2 + \lambda c^2 G + c^2 S = 0, \qquad \lambda^2 S + \lambda P + c^2 G P = 0.$$

All the coefficients are positive. We again eliminate the highest-order terms and, respectively, the constant terms to obtain

$$\lambda(c^2 GS - P) + c^2(S^2 - GP) = 0, \tag{A.9}$$

$$\lambda(GP - S^2) + P(c^2 G^2 - S) = 0. \tag{A.10}$$

It is easy to show that, if any of the coefficients in (A.9) or (A.10) are zero, then they all are. In this case, we have $GP = S^2$, but this is impossible for $\gamma^2 < 3$, since

$$S^2 = (r_k^2 + r_j^2)^2 \geq 4(r_k^2 r_j^2) = 4P. \tag{A.11}$$

We may thus assume that none of the coefficients in (A.9) or (A.10) are zero. Next, we eliminate λ from (A.9), (A.10) and find that

$$c^4 G^2 + c^2 S(S^2/PG - 3) + P/G = 0.$$

Since c^2 is positive, the coefficient of c^2 must be negative, and the discriminant (of the quadratic polynomial in c^2) must be positive. This leads to

$$\left(\frac{S^2}{PG} - 3\right)^2 > 4/3,$$

which is impossible by (A.11) for $\gamma^2 \leq 1$. □

Remark A.2. It is worth noting that double eigenvalues are possible if γ is larger: if $c^2 = 81\pi^2/4000$ and $\gamma^2 = 91/9$, then $\lambda = (9/40)\pi^2 \exp(i2\pi/3)$ is a double eigenvalue (corresponding to $k = 1$ and $j = 2$ in the notation of (A.7), (A.8)). This shows that null-controllability does not hold for all $\gamma > 0$. Our restriction: $0 < \gamma \leq 1$ is only sufficient to ensure that no double eigenvalues occur.

REFERENCES

[1] J. A. BURNS, Z. LIU AND S. ZHENG, *On the energy decay of a linear thermoelastic bar*, J. Math. Anal. Appl., to appear.

[2] A. DAY, *Heat Conduction within Linear Thermoelasticity*, Springer-Verlag, New York, 1985.

[3] H. O. FATTORINI, *Boundary control systems*, SIAM J. Control Optim., 6 (1968), pp. 349–385.

[4] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic problems in one space dimension*, Arch. Rational Mech. Anal., 4 (1971), pp. 272–292.

[5] J. S. GIBSON, I. G. ROSEN, AND G. TAO, *Approximation in control of thermoelastic systems*, SIAM J. Control Optim., 30 (1992), pp. 1163–1189.

[6] S. W. HANSEN, *Exponential energy decay in a linear thermoelastic rod*, J. Math. Anal. Appl., 167 (1992), pp. 429-442.

[7] ———, *Bounds on functions biorthogonal to sets of complex exponentials; Control of damped elastic systems*, J. Math. Anal. Appl., 158 (1991), pp. 487–508.

[8] S. HANSEN AND G. WEISS, *The operator Carleson measure criterion for admissibility of control operators for diagonal semigroups on l^2*, Systems Control Lett., 16 (1991), pp. 219–227.

[9] E. HILLE, *Analytic Function Theory*, Vol. II, Ginn & Co., New York, 1962.
[10] L. F. HO AND D. L. RUSSELL, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–640.
[11] T. KATO, *Perturbation Theory for Linear Operators*, Grundlehren der mathematischen Wissenschaften, 2nd ed., Vol. 132, Springer-Verlag, New York, 1976.
[12] J. V. KIM, *On the energy decay of a linear thermoelastic bar and plate*, SIAM J. Math. Anal., 23 (1992), pp. 889–899.
[13] J. E. LAGNESE, *Boundary Stabilization of Thin Plates*, SIAM Studies in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.
[14] J. E. LAGNESE AND J.-L. LIONS, *Modelling Analysis and Control of Thin Plates*, Springer-Verlag, collection Recherches en Mathématiques Appliquées, New York, 1989.
[15] J.-L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, New York, 1972.
[16] Z. Y. LIU, *Approximation and Control of a Thermoviscoelastic System*, PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1989.
[17] Z. LIU AND S. ZHENG, *Exponential stability of semigroup associated with a thermoelastic system*, Quart. Appl. Math., LI (1993), pp. 535–545.
[18] K. NARUKAWA, *Boundary value control of thermoelastic systems*, Hiroshima Math. J., 13 (1983), pp. 227–272.
[19] A. PAZY, *Semigroups of Linear Operators and Applications and Partial Differential Equations*, Springer-Verlag, New York, 1983.
[20] D. L. RUSSELL, *Canonical forms and spectral determination for a class of distributed parameter control systems*, J. Math. Anal. Appl., 52 (1978), pp. 186–225.
[21] M. SLEMROD, *Global existence, uniqueness, and asymptotic stability of classical smooth solutions in one-dimensional nonlinear thermoelasticity*, Arch. Rat. Mech. Anal., 76 (1981), pp. 97–133.
[22] G. WEISS, *Admissibility of input elements for diagonal semigroups on l^2*, Systems Control Lett., 10 (1988), pp. 79–82.
[23] ———, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[24] D. WASHBURN, *A bound on the boundary input map for parabolic equations with application to time optimal control*, SIAM J. Control Optim., 17 (1979), pp. 652–671.
[25] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1075–1097, July 1994

011

CONTROL OF INFINITE BEHAVIOR OF FINITE AUTOMATA*

J. G. THISTLE† AND W. M. WONHAM‡

Abstract. A problem in the control of automata on infinite strings is defined and analyzed. The key to the investigation is the development of a fixpoint characterization of the "controllability subset" of a deterministic Rabin automaton, the set of states from which the automaton can be controlled to the satisfaction of its own acceptance condition. The fixpoint representation allows straightforward computation of the controllability subset and the construction of a suitable state-feedback control for the automaton. The results have applications to control synthesis, automaton synthesis, and decision procedures for logical satisfiability; in particular, they represent a direct, efficient and natural solution to Church's problem, the construction of winning strategies for two-player zero-sum ω-regular games of perfect information, and the emptiness problem for automata on infinite trees.

Key words. discrete-event systems, synthesis problems, Church's problem, tree automaton emptiness, ω-languages, ω-automata, ω-regular games

AMS subject classifications. 93B50, 03D05, 68Q68, 03B70, 03B25, 03B45, 90D05, 68Q60, 03C30, 03C50, 03B35

1. Introduction.

This paper and the companion article [42] outline a theory of the control of infinite behavior of discrete-event systems. Based on [43], the articles extend some of the fundamental results of the finitary *supervisory control theory* of Ramadge and Wonham (see [35]). The present paper focuses on a key computational problem that arises not only in connection with control but also in several other contexts within the study of infinite behavior of discrete-event systems. As formulated here, the problem concerns the control of finite automata on infinite strings, commonly termed finite *ω-automata*.

A survey of the theory of ω-automata is given in [44]. Like their more familiar counterparts on finite strings, ω-automata consist of transition structures and accompanying acceptance conditions. However, whereas acceptance conditions for finite strings depend on the state *last* entered by the automaton upon reading a symbol string, the criterion for infinite strings instead involves the set of states entered *infinitely often* during the processing of the string. This article is specifically concerned with *Rabin* acceptance conditions (see §3).

Though relatively new to control theory, ω-automata are well-established tools in the theory of discrete-event systems. First introduced by Büchi as a means of deciding satisfiability of certain logical formulas [3] and by Muller as a natural means of describing infinite behavior of asynchronous switching circuits [27], they have since been applied to numerous studies of digital hardware, computer software and the associated logics (see, for example, [44], [12], [22], [23], [48], [46], [29], [1]), and, to a more limited extent, discrete-event control systems (see [2], [17], [36], [18], [40]).

* Received by the editors July 19, 1991; accepted for publication (in revised form) January 20, 1993.

† Département de génie électrique et de génie informatique, Ecole Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Canada H3C 3A7. The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant OGP0007399 and a Postdoctoral Fellowship.

‡ Systems Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Canada M5S 1A4. The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant OGP0007399.

Automata are traditionally interpreted as information-processing devices, "reading" strings of symbols (or inputs of other forms), executing a state transition in response to each symbol, and either accepting or rejecting strings according to their acceptance conditions. From the perspective of control, however, they are more naturally viewed as dynamic systems that spontaneously execute sequences of state transitions, "generating" strings of symbols as they do so, the acceptance condition distinguishing those strings that in some sense represent desirable behavior [34]. This is the view that we adopt here. Moreover, it is assumed here that, at any point in the system's operation, it is possible to restrict the set of symbols that may be generated to any one of a given family of subsets of the complete alphabet; this represents a means of control over the state transitions of the automaton. The article examines the computation of the "controllability subset" of a deterministic Rabin automaton (the set of states from which the automaton can be controlled to the satisfaction of its own acceptance condition) and the construction of corresponding control strategies.

This problem arises naturally in the study of the supervisory control of infinite behavior of discrete-event systems; indeed, the results of the current article are applied to effective supervisor synthesis in [42]. The same problem, however, also occurs in several other branches of the theory of discrete-event systems. In its earliest formulation, it represents a basic paradigm for program synthesis: *Church's problem* is that of constructing an automaton whose infinite input-output behavior satisfies a given logical formula [6], [7]. As many of the relevant logical languages are equivalent in expressiveness to finite ω-automata, Church's problem is often formulated in purely automata-theoretic terms, which make it formally equivalent to the problem studied here [24], [4], [33], [29], [28]. The first comprehensive solution to Church's problem exploited its equivalence to the problem of constructing winning strategies for two-person, zero-sum, and ω-regular games of perfect information [24], [4]. Simpler solutions have since been obtained through reduction to a further equivalent problem, the so-called *emptiness problem for automata on infinite trees* [44], [33], [29], [1], [49], which represents an important means of deciding satisfiability for a variety of logics, including many propositional logics of programs [12], [31], [32], [38], [47].

The present control-theoretic formulation admits a particularly direct and natural solution. First, the controllability subset is defined "operationally" in terms of the infinite strings generated by the automaton under suitable control. It is then shown that this operational definition can be replaced by a "denotational" one, a more direct characterization of the controllability subset as a certain fixpoint of an "inverse dynamics operator," which depends only on the one-step dynamics of the controlled automaton. The computation of this fixpoint is straightforward, and intermediate results of the calculation can be used to compute a suitable control in the form of a state feedback map. This approach is essentially optimal in computational complexity.

Section 2 presents some preliminaries concerning extremal fixpoints of monotone operators. In §3 the controllability subset is formally defined. Section 4 discusses some operations on automata that allow for structural induction in later definitions and proofs. The "inverse dynamics" and "p-reachability" operators of a deterministic Rabin automaton are then defined in §5. In §6 the controllability subset is characterized as a certain fixpoint of these operators. The computational complexity of deciding membership of a state in the controllability subset is examined in §7, and it is shown that straightforward computation of the fixpoint is essentially optimal in this respect. The results of the article are illustrated with an example in §8 and compared with earlier work in §9.

One of the advantages of the methods of the article is its extensibility to control under liveness assumptions. The case of liveness assumptions represented by Büchi acceptance conditions is outlined in [41], where the results of the present article are also summarized. More general forms of liveness assumptions will be treated in future reports.

2. Preliminaries: Fixpoints of monotone operators. The methods of this paper require special notation for extremal fixpoints of monotone operators. This section presents a suitable calculus based on the use of "fixpoint quantifiers." Originally applied to the denotational semantics of recursion (see [8]), such quantifiers have also been used to extend the power of various logics of programs [10], [30], [21], [45].

Our definitions of monotonicity and continuity and our presentation of the fundamental results of Tarski and Knaster (Theorem 2.1) are adapted from those of [16].

2.1. Monotone and continuous operators. A k-ary *operator* on a power set 2^X is a map $f : (2^X)^k \to 2^X$.

An operator is *monotone* if it preserves inclusion, that is,

$$X_i \subseteq X_i' \Longrightarrow f(X_1, \ldots, X_i, \ldots X_k) \subseteq f(X_1, \ldots, X_i', \ldots X_k), \qquad 1 \leq i \leq k.$$

An operator is $\cup$-*continuous* if, for any i, $1 \leq i \leq k$, and any nondecreasing sequence $X_i^0 \subseteq X_i^1 \subseteq X_i^2 \cdots$,

$$\bigcup_{j=0}^{\infty} f(X_1, \ldots, X_i^j, \ldots, X_k) = f\left(X_1, \ldots, \bigcup_{j=0}^{\infty} X_i^j, \ldots, X_k\right).$$

An operator is $\cap$-*continuous* if, for any i, $1 \leq i \leq k$ and any nonincreasing sequence $X_i^0 \supseteq X_i^1 \supseteq X_i^2 \cdots$,

$$\bigcap_{j=0}^{\infty} f(X_1, \ldots, X_i^j, \ldots, X_k) = f\left(X_1, \ldots, \bigcap_{j=0}^{\infty} X_i^j, \ldots, X_k\right).$$

Both $\cup$-continuity and $\cap$-continuity imply monotonicity; for operators on finite power sets, the reverse implications hold.

Monotonicity is important for our purposes because it implies the existence of extremal fixpoints (in the sense of set inclusion). The continuity properties provide a convenient means of computing such fixpoints (see Theorem 2.1 below).

2.2. A fixpoint calculus. The expressions of our fixpoint notation consist of monotone operators applied to subsets and the "fixpoint quantifiers" μ and ν that quantify over subsets. Expressions of the form

$$\mu Y.\ \phi(Y) \qquad \text{(respectively, } \nu Y.\ \phi(Y)\text{)}$$

represent the least (respectively, greatest) $Y \subseteq X$ such that $Y = \phi(Y)$, in other words, the least (respectively, greatest) fixpoints of the operator that maps every $Y \subseteq X$ to $\phi(Y)$. The question of the existence of such fixpoints is dealt with below.

- For any $Y' \subseteq X$,

$$\mu Y.\ (Y' \cup Y) = Y'.$$

The dual result is

$$\nu Y.\ (Y' \cap Y) = Y'.$$

• Adding fixpoint quantifiers to the expressions in the previous example, we find that

$$\nu Y'.\ \mu Y.\ (Y' \cup Y) = \nu Y'.\ Y' = X$$

and, dually,

$$\mu Y'.\ \nu Y.\ (Y' \cap Y) = \mu Y'.\ Y' = \emptyset.$$

2.3. Fixpoint lemmata. The basic results on extremal fixpoints of monotone operators follow from more general theorems of Tarski and Knaster.[1]

THEOREM 2.1 (Tarski–Knaster). *Let $f : 2^X \to 2^X$ be a monotone operator on X. Then f has least and greatest fixpoints; in fact,*

(i) $\mu Y.\ f(Y) = \bigcap\{Y' \subseteq X :\ Y' = f(Y')\} = \bigcap\{Y' \subseteq X :\ Y' \supseteq f(Y')\}$,

(i′) $\nu Y.\ f(Y) = \bigcup\{Y' \subseteq X :\ Y' = f(Y')\} = \bigcup\{Y' \subseteq X :\ Y' \subseteq f(Y')\}$,

(ii) *If f is $\cup$-continuous then* $\mu Y.\ f(Y) = \bigcup_{i=0}^{\infty} f^i(\emptyset)$,

(ii′) *If f is $\cap$-continuous then* $\nu Y.\ f(Y) = \bigcap_{i=0}^{\infty} f^i(X)$

(*where f^i denotes the i-fold composition of f with itself*).

LEMMA 2.2. *Let $f_1,\ f_2 : 2^X \to 2^X$ be monotone operators on X. If*

$$f_1(Y) \subseteq f_2(Y) \quad \forall Y \subseteq X,$$

then (a) $\mu Y.\ f_1(Y) \subseteq \mu Y.\ f_2(Y)$ *and* (b) $\nu Y.\ f_1(Y) \subseteq \nu Y.\ f_2(Y)$.

Theorem 2.1 guarantees that operators have extremal fixpoints, provided that they are monotone. Monotonicity is clearly preserved under composition of operators, and Lemma 2.2 shows that it is preserved under fixpoint quantification.[2] Together, these results imply that the semantics of our calculus are well defined.

3. Controllability subsets of Rabin automata. Before defining Rabin automata and their controllability subsets, we establish some notation for formal languages.

Σ^* denotes the set of all finite strings (including the *empty string*, denoted by 1, having length 0) over some finite symbol alphabet Σ; Σ^ω denotes the set of (countably) infinite strings over Σ; Σ^∞ denotes $\Sigma^* \cup \Sigma^\omega$. A *language* is any set of strings over Σ; in particular, an *ω-language* is a subset of Σ^ω.

A finite string $k \in \Sigma^*$ is a *prefix* of $v \in \Sigma^\infty$ if k is an initial substring of v; in this case, we write $k \leq v$; if k is not identical to v (i.e., if k is a *proper* prefix), we may write $k < v$. For any string $v \in \Sigma^\infty$, we let $\mathrm{pre}(v)$ denote its set of (finite) prefixes, that is, $\mathrm{pre}(v) := \{k \in \Sigma^* : k \leq v\}$.

A *Rabin automaton* [26], [33] is a 5-tuple

$$\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$$

[1] See [39]. For more information on the history of these results, see [25]. The authors are grateful to Karen Rudie for pointing out this second reference.

[2] That is, if $f : (2^X)^k \longrightarrow 2^X$ is monotone, then so is the operator $f_1^\mu : (2^X)^{k-1} \longrightarrow 2^X$ defined by

$$f_1^\mu : (Y_2, \ldots, Y_k) \mapsto \mu Y_1.\ f(Y_1, Y_2, \ldots, Y_k),$$

and similarly for greatest fixpoints.

consisting of: an *alphabet* Σ, a finite *state set* X, a *transition function* $\delta : \Sigma \times X \longrightarrow 2^X$, an *initial state* x_0, and a *family of accepting pairs* $(R_p, I_p) \in 2^X \times 2^X$ with index set P.

It is convenient to extend δ to a map $\delta : \Sigma^* \longrightarrow 2^X$ according to

$$(1, x) \overset{\delta}{\mapsto} x,$$

$$(k\sigma, x) \overset{\delta}{\mapsto} \bigcup\{\delta(\sigma, x_k) : x_k \in \delta(k, x)\} \quad \forall\, k \in \Sigma^*,\ \sigma \in \Sigma.$$

A *path* on $\mathcal{A}$ of a string $v \in \Sigma^\infty$ is a total function $\pi : \mathrm{pre}(v) \longrightarrow X$ such that

$$\pi(1) = x_0 \quad \forall\, k \in \mathrm{pre}(v),\ \sigma \in \Sigma\ :\ k\sigma \in \mathrm{pre}(v) \Rightarrow \pi(k\sigma) \in \delta(\sigma, \pi(k)).$$

Thus a path determines a state sequence consistent with the form of the string and the transition structure of the automaton.

Often, we wish to discuss "paths" that do not begin with the initial state (i.e., for which $\pi(1) \neq x_0$). For this, we define $\mathcal{A}_x$ to be the automaton obtained from $\mathcal{A}$ by designating $x \in X$ the initial state.

In keeping with our interpretation of automata as generators, we say that a string $v \in \Sigma^\infty$ is *generated* by $\mathcal{A}$ if v has a path on $\mathcal{A}$.

The *recurrence set* of a path $\pi : \mathrm{pre}(s) \longrightarrow X$ is

$$\Omega_\pi := \{x \in X : |\pi^{-1}(x)| = \omega\}.$$

In other words, the recurrence set is the set of states entered infinitely often on a given path. The recurrence set Ω_π is nonempty if and only if the string s is infinite.

For our purposes, it is convenient to adopt a slight modification of the standard definition of acceptance. A path π is *accepted* if

$$\exists p \in P : \Omega_\pi \cap R_p \neq \emptyset,\ \Omega_\pi \subseteq I_p.$$

Such paths represent infinite state sequences along which, for some $p \in P$, R_p is "continually recurrent" (is occupied infinitely often) and I_p is "eventually invariant" (is occupied almost always). (It is more customary to specify in place of I_p its complement, say $\overline{I_p}$, and require the equivalent condition that $\Omega_\pi \cap \overline{I_p} = \emptyset$, in other words, that $\overline{I_p}$ be "finitely recurrent" (occupied almost never).)

The ω-language *accepted* by $\mathcal{A}$ is the set of all (infinite) strings over Σ that have accepted paths on $\mathcal{A}$.

We say that $\mathcal{A}$ is *deterministic* if $|\delta(\sigma, x)| \leq 1$ for all $\sigma \in \Sigma, x \in X$. In this case, we represent the transition function as a partial function $\delta : \Sigma^* \times X \longrightarrow X$, writing $\delta(k, x)!$ to signify that the function is defined for the particular argument (k, x). If $\mathcal{A}$ is deterministic, then every string has at most one path π on $\mathcal{A}$.

According to our definition, a string is accepted by $\mathcal{A}$ if it has a path on $\mathcal{A}$ along which, for some $p \in P$, R_p is continually recurrent and I_p is eventually invariant. While this notion of acceptance may appear arbitrary, it is quite general in the sense that all of the languages that are accepted by finite ω-automata, the so-called ω-regular languages, are accepted by deterministic Rabin automata [26], [5], [9]. Indeed, the Rabin acceptance condition arises naturally in the "determinization" of nondeterministic ω-automata [26], [37], [14].

To introduce a control feature, we assume that a family $\mathbf{C} \subseteq 2^\Sigma$ of *control patterns* is given, representing subsets of the alphabet to which we can restrict, at any point

in the operation of the automaton, the set of symbols that it may generate. Control strategies can be represented as "feedback maps" $f : \Sigma^* \longrightarrow \mathbf{C}$ (pfn), which can be interpreted as associating with the sequence of all past symbols generated by the automaton a corresponding control action. Consistent with this interpretation, we say that $v \in \Sigma^\infty$ is generated by $\mathcal{A}$ *under* f if v is generated by $\mathcal{A}$, and, for all $k\sigma \in \text{pre}(v)$, $f(k)$ is defined and $\sigma \in f(k)$. We say that f is complete with respect to $\mathcal{A}$ if for all $k \in \Sigma^*$ generated by $\mathcal{A}$ under f, $f(k)$ is defined [34], [42].

We can now define the set from which the infinite behavior of an automaton can be controlled to the satisfaction of its acceptance condition. For any Rabin automaton $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$, define $P^{\mathcal{A}} \subseteq X$ as the set of all $x \in X$ for which there exists a complete feedback map $f : \Sigma^* \longrightarrow \mathbf{C}$ such that

1) Every $s \in \Sigma^\omega$ generated by $\mathcal{A}_x$ under f is accepted by $\mathcal{A}_x$, and

2) For any $k \in \Sigma^*$ generated by $\mathcal{A}_x$ under f, there exists $\sigma \in \Sigma$ such that $k\sigma$ is generated by $\mathcal{A}_x$ under f.

Clause 1) captures the notion that f should control $\mathcal{A}$ to the satisfaction of its acceptance condition. Clause 2) eliminates trivial solutions by requiring that every finite string generated by $\mathcal{A}_x$ under f have proper extensions that are also generated by $\mathcal{A}_x$ under f; for deterministic $\mathcal{A}$, this means that the control strategy represented by f must avoid system deadlocks.

We call $P^{\mathcal{A}}$ the *controllability subset* of $\mathcal{A}$. The above definition can be considered "operational" in the sense that it is stated in terms of the infinite strings generated by $\mathcal{A}$ under suitable control. While straightforward and intelligible from an intuitive standpoint, this description is of limited mathematical usefulness. The main results of the article establish an alternative representation that can be described as "denotational," in the sense that it is mathematically much more direct; it characterizes the controllability subset as a certain fixpoint of an operator that depends in a simple manner on the transition structure of the automaton and the family of control patterns. The new definition allows both efficient computation of the controllability subset and effective synthesis of appropriate controls.

4. Automaton structure. The recursive form of our fixpoint characterization of the controllability subset and the inductive nature of some of our proofs necessitate a precise notion of the structural complexity of automata and methods of converting complex automata to simpler ones. One useful measure of structural complexity is the number of pairs in the acceptance condition; another is the number of "live" states as defined by Rabin [33].

Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a Rabin automaton. The set of *live* states of $\mathcal{A}$ is given by[3]

$$L(\mathcal{A}) := \{x \in X : (\exists \sigma \in \Sigma)\ \delta(\sigma, x) \neq x\}.$$

In other words, a state is live if other states can be reached from it. An approximate opposite to liveness is "degeneracy." A state $x \in X$ is *degenerate* if there are transitions leaving x, but all of them simply lead back to x; more precisely, $x \in X$ is degenerate if

$$\exists \sigma \in \Sigma :\ \delta(\sigma, x)! \quad \forall \sigma \in \Sigma : \delta(\sigma, x) = x.$$

The set of degenerate states of $\mathcal{A}$ is denoted by $D(\mathcal{A})$. The subsets $L(\mathcal{A})$ and $D(\mathcal{A})$ are, of course, disjoint, but $L(\mathcal{A}) \cup D(\mathcal{A})$ may be a proper subset of X; indeed, $L(\mathcal{A}) \cup D(\mathcal{A}) = \{x \in X : (\text{there exists } \sigma \in \Sigma)\delta(\sigma, x)!\}$.

[3] Rabin excludes the initial state from the set of live states.

For any $x \in X$, $X' \subseteq X$, and $p \in P$, the following operations[4] on the Rabin automaton $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ potentially reduce its structural complexity as measured by $|L(\mathcal{A})|$ and $|P|$:

self-looping of a subset: $\mathcal{A}(\hookrightarrow X') := (\Sigma, X, \delta', x_0, \{(R'_p, I'_p) : p \in P\})$, where

$$\delta'(\sigma, x') = \begin{cases} x' & \text{if } x' \in X', \\ \delta(\sigma, x') & \text{otherwise,} \end{cases}$$
$$R'_p = R_p \cup X', \quad I'_p = I_p \cup X' \quad \forall p \in P;$$

restriction to a subset: $\mathcal{A} \restriction X' := (\Sigma, X, \delta', x_0, \{(R'_p, I'_p) : p \in P\})$, where

$$\delta'(\sigma, x') = \begin{cases} \delta(\sigma, x') & \text{if } x' \in X', \\ x' & \text{otherwise,} \end{cases}$$
$$R'_p = R_p \cap X', \quad I'_p = I_p \cap X' \quad \forall p \in P;$$

exclusion of a pair: Let $\mathcal{A} \restriction (I_p \cup D(\mathcal{A})) = (\Sigma, X, \delta', x_0, \{(R'_q, I'_q) : q \in P\})$. Then

$$\mathcal{A} \downarrow p := \begin{cases} (\Sigma, X, \delta', x_0, \{(R'_q, I'_q) : q \in P\}) & \text{if } |P| = 1, \\ (\Sigma, X, \delta', x_0, \{(R'_q, I'_q) : q \in P \setminus \{p\}\}) & \text{if } |P| > 1. \end{cases}$$

Self-looping of a subset $X' \subseteq X$ turns every $x \in X'$ into a degenerate state and ensures that the singleton $\{x\}$ satisfies the acceptance criterion. On the other hand, restriction to a subset $X' \subseteq X$ turns all other states into degenerate states that do *not* satisfy the acceptance condition. Finally, exclusion of a pair indexed by $p \in P$ restricts the automaton to the subset $I_p \cup D(\mathcal{A})$ and, provided that $|P| > 1$, eliminates the pair (R_p, I_p). All three of these operations potentially reduce the number of live states, while the third potentially reduces the number of pairs in the acceptance condition.

The effects of the operations on the subset $P^{\mathcal{A}}$ can be described as follows.

PROPOSITION 4.1. *Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a deterministic Rabin automaton and suppose that $x \in X$, $X' \subseteq X$ and $p \in P$. Then*

(a) $P^{\mathcal{A}} \cap D(\mathcal{A}) = \bigcup_{p \in P} (R_p \cap I_p) \cap D(\mathcal{A})$,

(b) $P^{\mathcal{A}} \cup X' \subseteq P^{\mathcal{A}(\hookrightarrow X')}$,

(c) $P^{\mathcal{A} \restriction X'} \subseteq P^{\mathcal{A}} \cap X'$,

(d) $P^{\mathcal{A} \downarrow p} \subseteq P^{\mathcal{A}} \cap (I_p \cup D(\mathcal{A}))$.

Proof. The proof follows by definition. □

Part (a) means intuitively that an automaton can be controlled to satisfy its acceptance condition from a degenerate state if and only if that degenerate state itself satisfies the acceptance criterion (since no other states can be reached from it). Part (b) states that self-looping makes it easier to force the automaton to satisfy its acceptance criterion by creating degenerate states that satisfy the criterion, while (c) and (d) state, respectively, that restriction and exclusion make it harder by creating degenerate states that do not satisfy the acceptance condition and by strengthening the acceptance condition.

5. The inverse dynamics operator and p-reachability operators. We characterize $P^{\mathcal{A}} \subseteq X$ as a certain fixpoint of the following monotone operator. Let

[4] The operations of "self-looping" and "restriction" are based on similar operations defined in [33].

$\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a deterministic Rabin automaton. Its *inverse dynamics operator* is given by

$$\theta^{\mathcal{A}} : 2^X \longrightarrow 2^X,$$
$$X' \mapsto \{x \in X : (\exists \Gamma \in \mathbf{C})[(\forall \sigma \in \Gamma)\delta(\sigma, x) \in X',\ (\exists \sigma \in \Gamma)\delta(\sigma, x)!]\}.$$

For any $X' \subseteq X$, $\theta^{\mathcal{A}}(X')$ is the set of all states in which the automaton can be controlled so that its next state belongs to X'.

The subset $P^{\mathcal{A}}$ is indeed one of the fixpoints of $\theta^{\mathcal{A}}$, as shown in the following result.

PROPOSITION 5.1. *Let $\mathcal{A}$ be a deterministic Rabin automaton. Then*

$$P^{\mathcal{A}} = \theta^{\mathcal{A}}(P^{\mathcal{A}}).$$

Proof. The proof follows by definition. □

This result does not determine $P^{\mathcal{A}}$ uniquely, since $\theta^{\mathcal{A}}$ may have many fixpoints, but we show that, with the use of the inverse dynamics operator, $P^{\mathcal{A}}$ can be uniquely represented by an expression in our fixpoint calculus. Other significant state subsets can be represented in similar fashion; for example, if $X_1 \subseteq X$, then

$$\nu X_2.\ [\theta^{\mathcal{A}}(X_2) \cap X_1]$$

denotes the supremal "control-invariant" subset of X_1 (see Theorem 2.1 (i′)); the fixpoint

$$\mu X_2.\ \theta^{\mathcal{A}}(X_1 \cup X_2)$$

is the "reachability subset" of X_1, the set of states from which the automaton can be controlled to reach X_1 (see Theorem 2.1(ii)); for $I_p \subseteq X$, the subset

$$\mu X_2.\ [\theta^{\mathcal{A}}(X_1 \cup X_2) \cap I_p]$$

is the set of states from which $\mathcal{A}$ can be controlled to reach $X_1 \subseteq X$ by way of a path that lies within[5] I_p.

To write a succinct expression for $P^{\mathcal{A}}$, it is convenient to generalize this second notion of reachability. Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a Rabin automaton. For any $p \in P$, the *p-reachability operator* of $\mathcal{A}$ is given by

$$\rho_p^{\mathcal{A}} : (2^X)^2 \longrightarrow 2^X,$$
$$(X_1, X_2) \mapsto \mu X_3.\ [\theta^{\mathcal{A}}(X_1) \cup [\theta^{\mathcal{A}}(X_1 \cup X_2 \cup X_3) \cap I_p]].$$

Thus $\rho_p^{\mathcal{A}}(\emptyset, X_2) \subseteq X$ is simply the set of states from which $\mathcal{A}$ can be controlled to reach X_2 by way of a path that lies within I_p. In general, $\rho_p^{\mathcal{A}}(X_1, X_2)$ is the set of states from which $\mathcal{A}$ can be controlled to reach X_2 by way of a path that lies within $I_p \subseteq X$ or, failing that, to reach X_1 (and to do so in at most one state transition after leaving I_p).

It is shown in the next section that the subset $P^{\mathcal{A}}$ can be described in terms of extremal fixpoints of the p-reachability operators. As preliminaries, we present the following two results, which respectively relate the inverse dynamics operator and the p-reachability operators to the structure of automata.

[5] With the possible exception of the final state in the path, which may belong to $X_1 \setminus I_p$.

PROPOSITION 5.2. *Let $\mathcal{A} = (X, \Sigma, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a Rabin automaton and suppose $x \in X$, $p \in P$ and $X', X'' \subseteq X$. Then*

(a) $\theta^{\mathcal{A}}(X') \cap D(\mathcal{A}) = X' \cap D(\mathcal{A})$,

(b) $\theta^{\mathcal{A}(\hookrightarrow X'')}(X') \setminus X'' = \theta^{\mathcal{A}}(X') \setminus X''$,

(c) $\theta^{\mathcal{A}\restriction X''}(X') \cap X'' = \theta^{\mathcal{A}}(X') \cap X''$,

(d) $\theta^{\mathcal{A}\restriction p}(X') \cap [I_p \cup D(A)] = \theta^{\mathcal{A}}(X') \cap [I_p \cup D(A)]$.

Part (a) simply means that all transitions leaving a degenerate state lead back to that state. Part (b) says that the self-looping of a subset X'' affects only transitions from states belonging to X''. On the other hand, part (c) says that restriction to $X'' \subseteq X$ affects only transitions from states *not* belonging to X''; part (d) is similar.

PROPOSITION 5.3. *Let $\mathcal{A} = (X, \Sigma, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a Rabin automaton and suppose that $x \in X$, $X_1, X_2 \subseteq X$, and $p \in P$. Then*

(a) $\rho_p^{\mathcal{A}}(X_1, X_2) \cap D(\mathcal{A}) = [X_1 \cup (X_2 \cap I_p)] \cap D(\mathcal{A})$,

(b) *If $X' \subseteq X_1 \subseteq X$, $\rho_p^{\mathcal{A}(\hookrightarrow X')}(X_1, X_2) \setminus X' = \rho_p^{\mathcal{A}}(X_1, X_2) \setminus X'$,*

(c) *If $X_1 \subseteq X' \subseteq X$, $\rho_p^{\mathcal{A}\restriction X'}(X_1, X_2) \subseteq \rho_p^{\mathcal{A}}(X_1, X_2) \cap X'$,*

(d) *If $X_1, \rho_p^{\mathcal{A}}(X_1, X_2) \subseteq X'$ and $\rho_p^{\mathcal{A}\restriction X'}(X_1, X_2) = \rho_p^{\mathcal{A}}(X_1, X_2)$.*

Proof. See [43]. □

Part (a) reflects the nature of degeneracy: degenerate states lead only to themselves. Part (b) states that the self-looping of a subset $X' \subseteq X_1$ does not affect the p-reachability of (X_1, X_2) from other states: intuitively, if the automaton ever reaches the subset X', then it will also have reached X_1, so further transitions (namely, those affected by the self-looping operation) will be irrelevant. Part (c) means that restriction to a subset $X' \supseteq X_1$ limits p-reachability of a pair (X_1, X_2) (by eliminating paths to X_1 that lie partly outside X' and by shrinking the set I_p). On the other hand, if the pair (X_1, X_2) is not p-reachable from X', then the restriction makes no difference; this is reflected in part (d).

6. Fixpoint characterization of the controllability subset. The p-reachability operators admit a concise representation of the controllability subset. Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a deterministic Rabin automaton. Then $C^{\mathcal{A}} \subseteq X$ is given by[6]

$$C^{\mathcal{A}} := \begin{cases} \mu X_1.\ \nu X_2.\ \rho_p^{\mathcal{A}}(X_1, X_2 \cap R_p), & \text{if } |P| = 1, \\ \mu X_1.\ [\bigcup\limits_{p \in P} \nu X_2.\ \rho_p^{\mathcal{A}}(X_1, C^{\mathcal{A}(\hookrightarrow X_1 \cup (X_2 \cap R_p))\restriction p})] & \text{if } |P| > 1. \end{cases}$$

Consider the case where $|P| = 1$. By Theorem 2.1(ii), $C^{\mathcal{A}}$ is the least upper bound of the nondecreasing sequence of subsets $C_0 \subseteq C_1 \subseteq C_2 \subseteq \cdots$, given by

$$C_0 := \emptyset,$$
$$C_{i+1} := \nu X_2.\ \rho_p^{\mathcal{A}}(C_i, X_2 \cap R).$$

Intuitively, $C_1 \subseteq X$ is the largest $X_2 \subseteq X$ from which the automaton can be controlled to reach $X_2 \cap R_p$ by way of a path that lies within I_p (see Theorem 2.1(i′)); in other words, C_1 is the set of states from which the automaton can be controlled to remain forever within I_p and to enter R_p infinitely often. By induction, C_i represents the subset from which the automaton can be controlled so that it enters R_p infinitely often and enters $X \setminus I_p$ fewer than i times. Thus $C^{\mathcal{A}}$ is indeed the set of states from

[6] Existence of $C^{\mathcal{A}}$ follows by induction on $|P|$ from Proposition 6.2(b).

which $\mathcal{A}$ can be controlled to enter R_p infinitely often and eventually to enter I_p and remain there.

If $|P| > 1$, $C^{\mathcal{A}}$ is the least upper bound of the nondecreasing sequence $C_0 \subseteq C_1 \subseteq C_2 \subseteq \cdots$, given by

$$C_0 := \emptyset,$$
$$C_{i+1} := \bigcup_{p \in P} \nu X_2.\ \rho_p^{\mathcal{A}}(C_i, C^{\mathcal{A}(\hookrightarrow C_i \cup (X_2 \cap R_p))|p}).$$

Thus $C_1 \subseteq X$ is the set of states from which, for some $p \in P$, $\mathcal{A}$ can be controlled to remain forever within I_p and either enter R_p infinitely often or satisfy the acceptance condition obtained by excluding the pair (R_p, I_p), in other words, to remain forever within I_p and satisfy the original acceptance condition. By induction, C_i is the subset in which a sequence $p_i, p_{i-1}, p_{i-2}, \ldots, p_1$ of i elements of P can be inductively chosen in such a way that, if p is the latest element to have been selected, then $\mathcal{A}$ can be controlled so that it either remains within I_p forever and satisfies its acceptance condition or eventually reaches a state in which a new element in the sequence can be chosen. (When the last element p_1 is chosen, $\mathcal{A}$ is in a state from which it can be controlled to remain forever within I_{p_1} and satisfy its acceptance condition, that is, $\mathcal{A}$ is in C_1.) It follows that $C^{\mathcal{A}}$ is the subset from which $\mathcal{A}$ can be controlled to satisfy its acceptance condition.

As this interpretation suggests, $C^{\mathcal{A}}$ coincides with $P^{\mathcal{A}}$, as is shown in the next result.

PROPOSITION 6.1. *For any deterministic Rabin automaton $\mathcal{A}$, $P^{\mathcal{A}} = C^{\mathcal{A}}$.*

Proof. The proof follows by Proposition 6.3 and Theorem 6.4, below. □

Before proving the results that lead to Proposition 6.1, we establish some basic properties of $C^{\mathcal{A}}$ in Proposition 6.2.

PROPOSITION 6.2. *Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a deterministic Rabin automaton. Suppose that $x \in X$, $X' \subseteq X$, and $p \in P$. Then*

(a) $C^{\mathcal{A}} \cap D(\mathcal{A}) = \bigcup_{p \in P}(R_p \cap I_p) \cap D(\mathcal{A})$,

(b) $C^{\mathcal{A}(\hookrightarrow X')} \supseteq C^{\mathcal{A}} \cup X'$,

(c) $C^{\mathcal{A}(\hookrightarrow X')} = C^{\mathcal{A}} \iff X' \subseteq C^{\mathcal{A}}$,

(d) $C^{\mathcal{A}|X'} \subseteq C^{\mathcal{A}} \cap X'$,

(e) $C^{\mathcal{A}|p} \subseteq C^{\mathcal{A}} \cap [I_p \cup D(\mathcal{A})]$,

(f) $C^{\mathcal{A}} = \theta^{\mathcal{A}}(C^{\mathcal{A}})$,

(g) *if $L(\mathcal{A}) \subseteq I_p$ and $X' \supseteq C^{\mathcal{A}(\hookrightarrow X' \cap R_p \cap I_p)} \cap R_p \cap I_p$ then $C^{\mathcal{A}(\hookrightarrow X' \cap R_p \cap I_p)|p} = C^{\mathcal{A}(\hookrightarrow X' \cap R_p \cap I_p)}$,*

(h) *if $L(\mathcal{A}) \subseteq I_p$ then $C^{\mathcal{A}} = \nu X_2.\ \theta^{\mathcal{A}}(C^{\mathcal{A}(\hookrightarrow X_2 \cap R_p \cap I_p)})$.*

Proof. See [43]. □

Part (a) means that the automaton can be controlled to satisfy its acceptance condition from a degenerate state if and only if that degenerate state itself satisfies the acceptance criterion. This is natural, since degenerate states lead only to themselves.

Part (b) describes the way the controllability subset generally expands when a subset is self-looped, owing to the creation of degenerate states that satisfy the acceptance criterion. Part (c) states that the controllability subset is unaffected by this operation if and only if the states that are self-looped already belong to the controllability subset.

Parts (d) and (e) reflect the fact that restriction to a subset and exclusion of a pair shrink the controllability subset by creating degenerate states that do not satisfy the acceptance criterion and by strengthening the acceptance criterion.

Part (f) means simply that the controllability subset is a fixpoint of the inverse dynamics operator.

Part (g) describes a situation in which exclusion of a pair (R_p, I_p) does not affect the controllability subset, namely, that in which all live states belong to I_p (so that the restriction to $I_p \cup D(\mathcal{A})$ is of no consequence) and a sufficiently large subset of R_p has been self-looped (so that the elimination of the pair (R_p, I_p) does not strengthen the acceptance condition).

Finally, part (h) states that, when all live states belong to some I_p, the controllability subset is the set of all states from which the automaton can be controlled to satisfy its acceptance condition or simply to enter $R_p \cap I_p$ infinitely often (the invariance of I_p following automatically).

These preliminary results allow us to prove that $P^{\mathcal{A}}$ and $C^{\mathcal{A}}$ coincide (Proposition 6.1). We first establish that $P^{\mathcal{A}} \subseteq C^{\mathcal{A}}$; in other words, if $\mathcal{A}$ can be controlled to satisfy its acceptance condition from state x, then $x \in C^{\mathcal{A}}$:

PROPOSITION 6.3. *For any deterministic Rabin automaton $\mathcal{A}$, $P^{\mathcal{A}} \subseteq C^{\mathcal{A}}$.*

Proof. We proceed by induction on the number of live states of $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$, in the manner of [33]. Note that

$$\begin{aligned} & P^{\mathcal{A}} \cap D(\mathcal{A}) \\ &= \bigcup_{p \in P} (R_p \cap I_p) \cap D(\mathcal{A}) \qquad \text{(Prop. 4.1(a))} \\ &= C^{\mathcal{A}} \cap D(\mathcal{A}) \qquad \text{(Prop. 6.2(a))}. \end{aligned}$$

By Proposition 5.1, it thus suffices to show that $P^{\mathcal{A}} \cap L(\mathcal{A}) \subseteq C^{\mathcal{A}}$.

If $\mathcal{A}$ contains no live states, then the result holds vacuously. For the induction step, suppose that $x \in P^{\mathcal{A}} \cap L(\mathcal{A})$ and assume that the result holds for all Rabin automata with fewer live states than $\mathcal{A}$. We prove that $x \in C^{\mathcal{A}}$.

By assumption, there exists some complete feedback map $f : \Sigma^* \longrightarrow \mathbf{C}$ satisfying both clauses of the definition of $P^{\mathcal{A}}$. The central issue in the proof is the nature of the relationship between the strings generated by $\mathcal{A}_x$ under f and the live states of $\mathcal{A} =: (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$. The following three cases exhaust the possibilities:

(a) There exists a live state $x' \in X$ such that, for all $k' \in \Sigma^*$ generated by $\mathcal{A}_x$ under f, $\delta(k', x) \neq x'$;

(b) For some pair of live states $x', x'' \in X$, there exists $k' \in \Sigma^*$ generated by $\mathcal{A}_x$ under f such that $\delta(k', x) = x'$ and for all $k'' \in \Sigma^*$ generated by $\mathcal{A}_x$ under f such that $k' < k''$, $\delta(k'', x) \neq x''$;

(c) For all pairs of live states $x', x'' \in X$, and every $k' \in \Sigma^*$ generated by $\mathcal{A}_x$ under f such that $\delta(k', x) = x'$, there exists $k'' \in \Sigma^*$ generated by $\mathcal{A}_x$ under f such that $k' < k''$ and $\delta(k'', x) = x''$.

In case (a), we have

$$\begin{aligned} x &\in P^{\mathcal{A}\restriction(X \setminus \{x'\})} \\ &\subseteq C^{\mathcal{A}\restriction(X \setminus \{x'\})} \qquad \text{(by inductive hypothesis)} \\ &\subseteq C^{\mathcal{A}} \quad \text{(Prop. 6.2(d))}. \end{aligned}$$

Similarly, for case (b), we have

$$\begin{aligned} x' = \delta(k', x) &\in P^{\mathcal{A}\restriction(X \setminus \{x''\})} \\ &\subseteq C^{\mathcal{A}\restriction(X \setminus \{x''\})} \qquad \text{(by inductive hypothesis)} \\ &\subseteq C^{\mathcal{A}} \qquad \text{(Prop. 6.2(d))}. \end{aligned}$$

Thus

$$\begin{aligned} x &\in P^{\mathcal{A}} \\ &\subseteq P^{\mathcal{A}(\hookrightarrow x')} \qquad \text{(Prop. 4.1(b))} \\ &\subseteq C^{\mathcal{A}(\hookrightarrow x')} \qquad \text{(by inductive hypothesis)} \\ &= C^{\mathcal{A}} \qquad \text{(Prop. 6.2(c)).} \end{aligned}$$

In case (c), there exists a string s generated by $\mathcal{A}_x$ under f having a path π on $\mathcal{A}$ such that $\Omega_\pi = L(\mathcal{A})$. It follows that, for some $p \in P$,

$$L(\mathcal{A}) \subseteq I_p \quad \text{and} \quad L(\mathcal{A}) \cap R_p \neq \emptyset.$$

Furthermore, for any $x''' \in L(\mathcal{A})$, we have

$$\begin{aligned} x''' &\in P^{\mathcal{A}} \\ &= \theta^{\mathcal{A}}(P^{\mathcal{A}}) \qquad \text{(Prop. 5.1)} \\ &\subseteq \theta^{\mathcal{A}}(P^{\mathcal{A}(\hookrightarrow L(\mathcal{A})\cap R_p)}) \qquad \text{(Prop. 4.1 (b))} \\ &= \theta^{\mathcal{A}}(C^{\mathcal{A}(\hookrightarrow L(\mathcal{A})\cap R_p)}) \qquad \text{(by inductive hypothesis).} \end{aligned}$$

Thus

$$\begin{aligned} L(\mathcal{A}) &\subseteq \nu X_2.\ \theta^{\mathcal{A}}(C^{\mathcal{A}(\hookrightarrow X_2 \cap R_p \cap I_p)}) \qquad \text{(Thm. 2.1(i'))} \\ &= C^{\mathcal{A}} \qquad \text{(Prop. 6.2(h))} \end{aligned}$$

This completes the induction. □

The next result establishes the converse of Proposition 6.3, namely, it shows that, if the deterministic Rabin automaton $\mathcal{A}$ is in state $x \in C^{\mathcal{A}}$, then it can be controlled to satisfy its acceptance condition. Moreover, it states that "state feedback" is sufficient. Together, Proposition 6.3 and Theorem 6.4 thus imply that a Rabin automaton can be controlled to satisfy its acceptance condition if and only if it can be so controlled by means of state feedback.

THEOREM 6.4 (state feedback). *Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ be a deterministic Rabin automaton. Then there exists a total map $\phi^{\mathcal{A}} : C^{\mathcal{A}} \longrightarrow \mathbf{C}$ such that*

1) *For any $s \in \Sigma^\omega$ and any path $\pi : \mathrm{pre}(s) \longrightarrow X$ of s on $\mathcal{A}$, the condition*

$$\pi(1) \in C^{\mathcal{A}} \quad \text{and} \quad \forall k\sigma \in \mathrm{pre}(s) \ :\ \sigma \in \phi^{\mathcal{A}}(\pi(k))$$

implies that

$$\exists p \in P \ :\ \Omega_\pi \cap R_p \neq \emptyset \quad \text{and} \quad \Omega_\pi \subseteq I_p;$$

2) *For any $x \in C^{\mathcal{A}}$,*

$$\begin{gathered} \exists \sigma \in \phi^{\mathcal{A}}(x) \ :\ \delta(\sigma, x)!, \\ \forall \sigma \in \phi^{\mathcal{A}}(x) \ :\ \delta(\sigma, x)! \implies \delta(\sigma, x) \in C^{\mathcal{A}}. \end{gathered}$$

Proof. The details of the proof are intricate, but the methods are simple. We consider every least fixpoint as the limit of a nondecreasing sequence of state subsets, by Theorem 2.1. Membership in these subsets is used to define a well-founded partial ordering or "ranking" of states: the earlier a state occurs in the sequence of subsets, the lower its rank. The theorem is then proved through arguments concerning

the effect of a suitable state feedback control on various ranks of the system state; these typically show that, under appropriate conditions, a particular rank is either nonincreasing or strictly decreasing.

We construct a suitable map by induction on $|P|$. We give only the induction step; the base of the induction (dealing with the case where $|P| = 1$) is similar; see [43] for details.

Suppose then that $|P| > 1$ and assume that the result holds for all automata having fewer pairs. Then

$$C^{\mathcal{A}} = \mu X_1.\ \left[\bigcup_{p \in P} \nu X_2.\ \rho_p^{\mathcal{A}}(X_1, C^{\mathcal{A}(\hookrightarrow X_1 \cup (X_2 \cap R_p))\downarrow p}) \right].$$

By Theorem 2.1, this fixpoint is the least upper bound of the nondecreasing sequence $C_0^{\mathcal{A}} \subseteq C_1^{\mathcal{A}} \subseteq C_2^{\mathcal{A}} \subseteq \cdots$, given by

$$\begin{aligned} C_0^{\mathcal{A}} &:= \emptyset, \\ C_{i+1}^{\mathcal{A}} &:= \bigcup_{p \in P} \nu X_2.\ \rho_p^{\mathcal{A}}(C_i^{\mathcal{A}}, C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (X_2 \cap R_p))\downarrow p}) \\ &= \bigcup_{p \in P} C_{i+1,p}^{\mathcal{A}}, \end{aligned}$$

where

$$\begin{aligned} C_{i+1,p}^{\mathcal{A}} &:= \nu X_2.\ \rho_p^{\mathcal{A}}(C_i^{\mathcal{A}}, C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (X_2 \cap R_p))\downarrow p}) \\ &= \rho_p^{\mathcal{A}}(C_i^{\mathcal{A}}, C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (C_{i+1,p}^{\mathcal{A}} \cap R_p))\downarrow p}) \\ &= \mu X_3.\ \ [\theta^{\mathcal{A}}(C_i^{\mathcal{A}}) \\ &\qquad \cup [\theta^{\mathcal{A}}(C_i^{\mathcal{A}} \cup C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (C_{i+1,p}^{\mathcal{A}} \cap R_p))\downarrow p} \cup X_3) \cap I_p]]. \end{aligned}$$

Each $C_{i+1,p}^{\mathcal{A}}$ is the least upper bound of the nondecreasing sequence, given by

$$\begin{aligned} C_{i+1,p,0}^{\mathcal{A}} &:= \emptyset, \\ C_{i+1,p,j+1}^{\mathcal{A}} &:= [\theta^{\mathcal{A}}(C_i^{\mathcal{A}}) \\ &\qquad \cup [\theta^{\mathcal{A}}(C_i^{\mathcal{A}} \cup C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (C_{i+1,p}^{\mathcal{A}} \cap R_p))\downarrow p} \cup C_{i+1,p,j}^{\mathcal{A}}) \cap I_p]] \\ &= C_{i+1,p,j+1,0}^{\mathcal{A}} \cup C_{i+1,p,j+1,1}^{\mathcal{A}}, \end{aligned}$$

where

$$\begin{aligned} C_{i+1,p,j+1,0}^{\mathcal{A}} &:= \theta^{\mathcal{A}}(C_i^{\mathcal{A}}), \\ C_{i+1,p,j+1,1}^{\mathcal{A}} &:= \theta^{\mathcal{A}}(C_i^{\mathcal{A}} \cup C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (C_{i+1,p}^{\mathcal{A}} \cap R_p))\downarrow p} \cup C_{i+1,p,j}^{\mathcal{A}}) \cap I_p. \end{aligned}$$

Thus

$$C^{\mathcal{A}} = \bigcup_{i=1}^{|X|} \bigcup_{p \in P} \bigcup_{j=1}^{|X|} \bigcup_{k=0,1} C_{i,p,j,k}^{\mathcal{A}}.$$

Define total feedback maps $\phi_{i,p,j,k}^{\mathcal{A}} : C_{i,p,j,k}^{\mathcal{A}} \longrightarrow \mathbf{C}$ for each of the components $C_{i,p,j,k}^{\mathcal{A}}$ as follows:

- If $k = 0$, define $\phi_{i+1,p,j+1,k}^{\mathcal{A}} : C_{i+1,p,j+1,k}^{\mathcal{A}} \longrightarrow \mathbf{C}$ so that, for all $x \in C_{i+1,p,j+1,0}^{\mathcal{A}}$,

$$\forall \sigma \in \phi_{i+1,p,j+1,k}^{\mathcal{A}}(x)\ :\ \delta(\sigma, x) \in C_i^{\mathcal{A}};$$

• If $k = 1$, define $\phi^{\mathcal{A}}_{i+1,p,j+1,k} : C^{\mathcal{A}}_{i+1,p,j+1,k} \longrightarrow \mathbf{C}$ so that, for all $x \in C^{\mathcal{A}}_{i+1,p,j+1,1} \subseteq I_p$,

$$\forall \sigma \in \phi^{\mathcal{A}}_{i+1,p,j+1,k}(x) : \\ \delta(\sigma, x) \in C^{\mathcal{A}}_i \cup C^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p} \cup C^{\mathcal{A}}_{i+1,p,j} \ .$$

Choose a total ordering of P and order the 4-tuples (i,p,j,k) lexicographically. Define a total map $\phi^{\mathcal{A}} : C^{\mathcal{A}} \longrightarrow \mathbf{C}$ so that

$$\phi^{\mathcal{A}} : x \mapsto \begin{cases} \phi^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}(x) & \text{if } x \in C^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p} \setminus C^{\mathcal{A}}_i \cup R_p, \\ \phi^{\mathcal{A}}_{i+1,p,j+1,k}(x) & \text{otherwise,} \end{cases}$$

where (i,p,j,k) is the least 4-tuple in the lexicographic ordering such that $x \in C^{\mathcal{A}}_{i+1,p,j+1,k}$.

Clause 2) of the theorem follows from the definition of $\phi^{\mathcal{A}}$ (by the inductive hypothesis). For clause 1), suppose that $s \in \Sigma^\omega$ and there exists a path $\pi : \mathrm{pre}(s) \longrightarrow X$ of s on $\mathcal{A}$ such that

$$\pi(1) = x \quad \text{and} \quad (\forall k\sigma \in \mathrm{pre}(s))[\sigma \in \phi^{\mathcal{A}}(\pi(k))].$$

We must show that there exists $p \in P \ : \ \Omega_\pi \cap R_p \neq \emptyset$ and $\Omega_\pi \subseteq I_p$.

Define

$$I\text{−rank} : C^{\mathcal{A}} \longrightarrow \mathbb{N} \times P, \\ x \mapsto (i,p),$$

where (i,p) is the least pair (in the lexicographic ordering) such that $x \in C^{\mathcal{A}}_{i,p}$.

Note that

$$\begin{aligned} & C^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p} \setminus (C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p) \cup D(\mathcal{A})) \\ & = \theta^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}(C^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}) \setminus (C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p) \cup D(\mathcal{A})) \\ & \qquad \text{(Prop. 6.2 (f))} \\ & = [\theta^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}(C^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}) \cap I_p] \setminus (C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p) \\ & \qquad \cup D(\mathcal{A})) \text{ (Prop. 6.2 (e))} \\ & = [\theta^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))}(C^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}) \cap I_p] \setminus (C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p) \\ & \qquad \cup D(\mathcal{A})) \text{ (Prop. 5.2 (d))} \\ & = [\theta^{\mathcal{A}}(C^{\mathcal{A}(\hookrightarrow C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}) \cap I_p] \setminus (C^{\mathcal{A}}_i \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p) \cup D(\mathcal{A})) \\ & \qquad \text{(Prop. 5.2 (b))} \\ & \subseteq C^{\mathcal{A}}_{i+1,p}. \end{aligned}$$

It follows, by the definition of $\phi^{\mathcal{A}}$ and the inductive hypothesis, that, for all $x \in C^{\mathcal{A}}, \sigma \in \phi^{\mathcal{A}}(x)$, if $\delta(\sigma,x) \notin D(\mathcal{A})$, then

$$I\text{−rank}(\delta(\sigma,x)) \leq I\text{−rank}(x),$$

and, if I−rank$(x) = (i,p)$ and $x \notin I_p \cup D(\mathcal{A})$, then, by Proposition 6.2(e),

$$I\text{−rank}(\delta(\sigma,x)) < I\text{−rank}(x) \quad \forall \sigma \in \phi^{\mathcal{A}}(x).$$

If (i,p) is the least pair for which $\Omega_\pi \cap C^{\mathcal{A}}_{i+1,p} \neq \emptyset$, then it follows that $\Omega_\pi \subseteq C^{\mathcal{A}}_{i+1,p} \cap (I_p \cup D(\mathcal{A}))$. Note that, if $\Omega_\pi \cap C^{\mathcal{A}}_{i+1,p} \cap D(\mathcal{A}) \neq \emptyset$, then there exists $q \in P \ : \ \Omega_\pi \cap R_q \neq \emptyset$

and $\Omega_\pi \subseteq I_q$ holds (by Proposition 6.2(a)). We may therefore assume instead that $\Omega_\pi \subseteq I_p$.

Define

$$R\text{-rank} : C^{\mathcal{A}} \longrightarrow \mathbb{N} \times P \times \mathbb{N},$$
$$x \mapsto (i, p, j)$$

where (i, p, j) is the least triple such that $x \in C^{\mathcal{A}}_{i,p,j}$.

By definition of $\phi^{\mathcal{A}}$, if $x \in C^{\mathcal{A}}$, $\sigma \in \phi^{\mathcal{A}}(x)$, $I\text{-rank}(x) = (i+1, p)$, and $\delta(\sigma, x) \notin C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p}$, then

$$R\text{-rank}(\delta(\sigma, x)) < R\text{-rank}(x).$$

If (i, p) is the least pair such that $\Omega_\pi \cap C^{\mathcal{A}}_{i+1,p} \neq \emptyset$, then it follows that $\Omega_\pi \cap C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p} \neq \emptyset$. Then, however, by the inductive assumption, we have either $\Omega_\pi \cap R_p \neq \emptyset$ or $\Omega_\pi \subseteq C^{\mathcal{A}(\hookrightarrow C_i^{\mathcal{A}} \cup (C^{\mathcal{A}}_{i+1,p} \cap R_p))\downarrow p} \setminus (C_i^{\mathcal{A}} \cup R_p)$. The result follows by another application of the inductive assumption. □

7. The complexity of computing controllability subsets. In this section, we demonstrate that the computation of controllability subsets by straightforward calculation of the appropriate fixpoint is essentially optimal. We first establish that the problem is NP-complete and then show that calculation of the fixpoint is singly exponential in the number of accepting pairs and polynomial in the size of the state set. These results match the strongest known results for the polynomially equivalent emptiness problem for Rabin automata on infinite trees [13], [29].

THEOREM 7.1. *The problem of deciding membership in the controllability subset of a deterministic Rabin automaton is* NP*-complete.*

Proof. NP*-hardness.* The proof follows by reduction from the emptiness problem for Rabin automata on infinite trees over a one-symbol alphabet. This was shown to be NP-hard in [13].

Membership in NP. If a Rabin automaton $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ is deterministic, then it follows from Proposition 6.3 and Theorem 6.4 that $x \in C^{\mathcal{A}}$ if and only if there exists a "state feedback" map $\phi : X \longrightarrow \mathbf{C}$ (pfn) such that $x \in \text{dom}(\phi)$ and if the following conditions hold:

1) For any $s \in \Sigma^\omega$ and any path $\pi : \text{pre}(s) \longrightarrow X$ of s on $\mathcal{A}$, the condition

$$\pi(1) \in \text{dom}(\phi) \quad \text{and} \quad \forall k\sigma \in \text{pre}(s) \ : \ \sigma \in \phi^{\mathcal{A}}(\pi(k))$$

implies that

$$\exists p \in P \ : \ \Omega_\pi \cap R_p \neq \emptyset \quad \text{and} \quad \Omega_\pi \subseteq I_p;$$

2) For any $x' \in \text{dom}(\phi)$,

$$\exists \sigma \in \phi(x') \ : \ \delta(\sigma, x')! \text{ and}$$
$$\forall \sigma \in \phi(x') \ : \ \delta(\sigma, x')! \implies \delta(\sigma, x') \in \text{dom}(\phi).$$

However, any map $\phi : X \longrightarrow \mathbf{C}$ can be constructed in polynomial time, and the conditions 1) and 2) can be checked in polynomial time, using the results of [15]. NP-hardness follows. See [43] for details. □

THEOREM 7.2. *The controllability subset of a deterministic Rabin automaton $\mathcal{A}$ can be computed in time $\bigcirc(kl(mn)^{3m})$, where k is the number of control patterns,*

l is the size of the alphabet, m is the number of state subset pairs in the acceptance condition, and n is the number of states.

Proof. The proof follows by straightforward analysis of the fixpoint computation. See [43] for details. □

8. Example. Consider the Rabin automaton $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ pictured in Fig. 8.1; the index set P is $\{\alpha, \beta\}$ and the pairs $(R_\alpha, I_\alpha) = (\{4\}, \{1, 2, 3, 4\})$ and $(R_\beta, I_\beta) = (\{-4\}, \{-1, -2, -3, -4\})$ are represented by the pairs of dotted and dashed boxes.

For simplicity, we take the alphabet Σ to be X^2; the transition function is represented by the arcs of the diagram, the symbol associated with a transition from state i to j being (i, j). The automaton is thus deterministic.

The family $\mathbf{C}$ of control patterns is the set of all subsets of Σ that contain the following symbols:

$$(0, \pm 1),\ (\pm 1, \mp 1),\ (\pm 3, \pm 4),\ (\pm 4, \pm 3),\ (\pm 2, \pm 3),\ (\pm 2, \pm 4).$$

This means that the state transitions corresponding to these symbols cannot be prevented, while all others can.

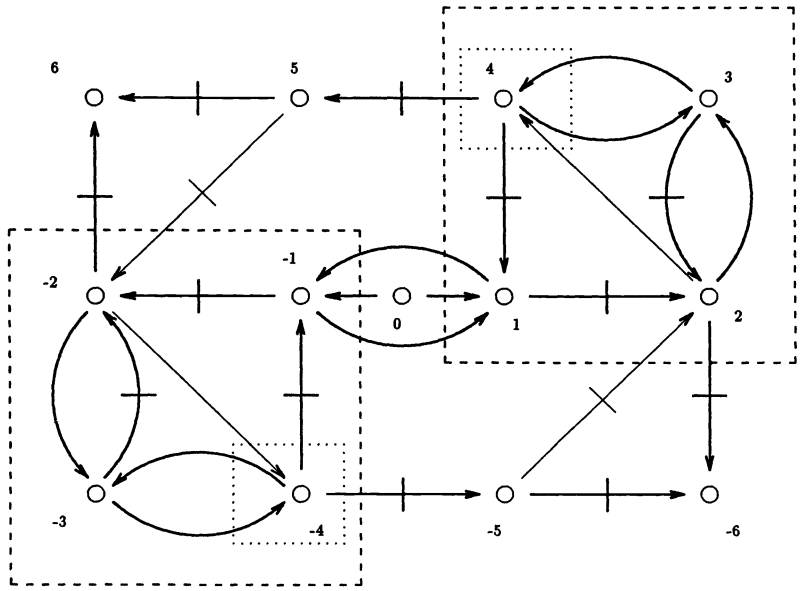


FIG. 8.1. *Rabin automaton $\mathcal{A}$.*

The controllability subset is computed by recursive application of (ii) and (ii′) of Theorem 2.1. The calculation is displayed in Table 8.1 using the notation of [43].

At the beginning of the procedure and after completing the first column of the table, it is necessary to construct an automaton of the form $\mathcal{A}(\hookrightarrow X_1 \cup (X_2 \cap R_p)) \downarrow p$ for every $p \in P$ and to compute its controllability subset. In our example, we begin the computation by constructing the automaton

$$\mathcal{A}(\hookrightarrow \emptyset \cup (X \cap R_\alpha)) \downarrow \alpha = \mathcal{A}(\hookrightarrow R_\alpha) \downarrow \alpha,$$

shown in Fig. 8.2, and calculating $C^{\mathcal{A}(\hookrightarrow R_\alpha)\downarrow\alpha}$, as shown in Table 8.2. (The corresponding results for $\beta \in P$ are obtained by replacing the states with their negatives.)

TABLE 8.1
Computation of $C^{\mathcal{A}}$.

	$C^{\mathcal{A}}_{1,\alpha,1,1}$	$C^{\mathcal{A}}_{2,\alpha,1,0}$	$\phi^{\mathcal{A}}(x)$
6			
5		α, β	$\Sigma \setminus \{(5,6)\}$
4	α	α, β	$\Sigma \setminus \{(4,1),(4,5)\}$
3	α	α, β	$[\Sigma \setminus \{(3,2)\}]$
2	α	α, β	$[\Sigma \setminus \{(2,-6)\}]$
1			
0			
-1			
-2	β	α, β	$[\Sigma \setminus \{(-2,6)\}]$
-3	β	α, β	$[\Sigma \setminus \{(-3,-2)\}]$
-4	β	α, β	$\Sigma \setminus \{(-4,-1),(-4,-5)\}$
-5		α, β	$\Sigma \setminus \{(-5,-6)\}$
-6			

$C^{\mathcal{A}}_{1,\beta,1,1} = C^{\mathcal{A}}_{2,\beta,1,0}$

Once the first column of Table 8.1 is computed, we construct

$$\mathcal{A}(\hookrightarrow X_1 \cup (X \cap R_\alpha)) \downarrow \alpha = \mathcal{A}(\hookrightarrow X_1) \downarrow \alpha$$

(where X_1 is the set of states marked by either α or β in the first column of Table 8.1) and calculate its controllability subset. The results are displayed in Fig. 8.3 and Table 8.3. (The corresponding results for β are again obtained by symmetry.)

Note that the states ± 6 are excluded from $C^{\mathcal{A}}$; they are "dead ends" from which no other state can be reached. The states $-1, 0, 1$ are excluded because of the cycle formed by 1 and -1.

A state feedback controller is given by the last column of Table 8.1. It is computed by identifying the subsets defined in the proof of Theorem 6.4 with the appropriate entries in the table and assigning control patterns that satisfy the rules set out in the proof. Those control patterns that appear in square brackets are taken from the feedback maps for $\mathcal{A}(\hookrightarrow R_\alpha) \downarrow \alpha$ and $\mathcal{A}(\hookrightarrow R_\beta) \downarrow \beta$, in accordance with the rules. The transition structure of Fig. 8.4 represents the set of strings generated by the automaton under the state feedback map. It can be plainly seen that every infinite string generated is accepted by $\mathcal{A}$ and that every finite string generated has an extension that is also generated under the state feedback control.

9. Discussion. We have presented a procedure for the computation of the *controllability subset* of a deterministic Rabin automaton, namely, the set of states from which the automaton can be controlled to the satisfaction of its acceptance condition. The key to the method is the representation of the controllability subset as a fixpoint of an *inverse dynamics operator*, which depends only on the one-step dynamics of the controlled system. Straightforward computation of this fixpoint matches tight upper bounds on the complexity of the problem. Moreover, intermediate results of the calculation allow the construction of a state feedback map that provides suitable control action.

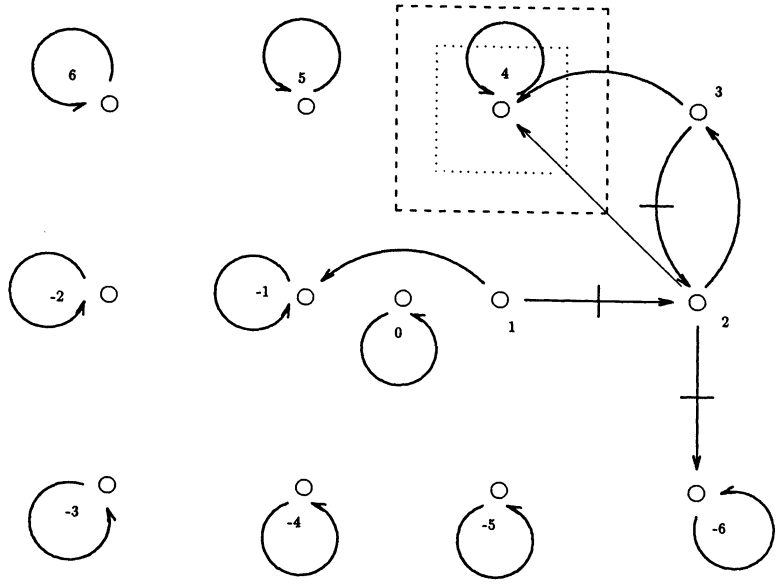


FIG. 8.2. *Simplified automaton* $\mathcal{A}(\hookrightarrow R_\alpha) \downharpoonright \alpha$.

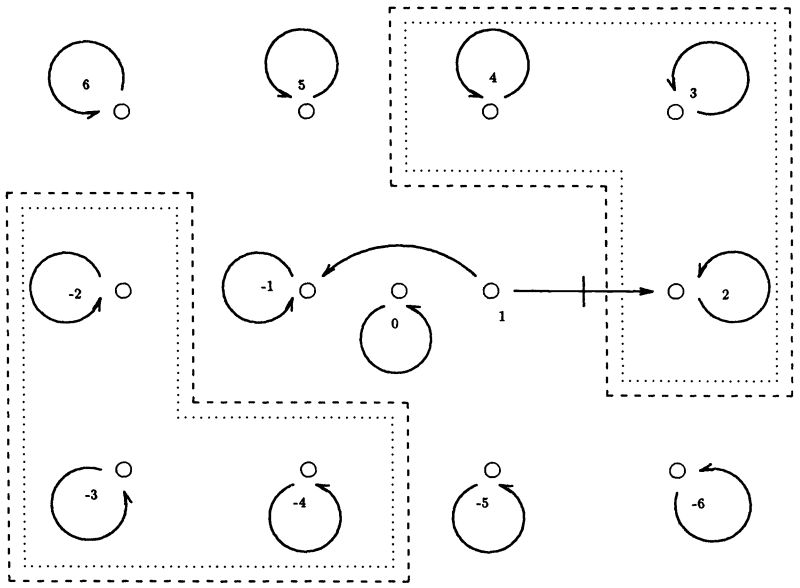


FIG. 8.3. *Simplified automaton* $\mathcal{A}(\hookrightarrow X_1) \downharpoonright \alpha$, *where* $X_1 = \{-4, -3, -2, 2, 3, 4\}$. (*The dotted and dashed boxes represent a single accepting pair.*)

The problem studied in the article was, in effect, first solved by Büchi and Landweber [24], [4], who applied game-theoretic techniques to the study of Church's problem [7]. Simpler solutions were later obtained through the equivalent emptiness problem for automata on infinite trees [33], [29]. The earliest approaches to the emptiness problem were developed by Rabin [33] and Hossley and Rackoff [20], employing, respectively, the structural induction method used in the present article and a reduction to the emptiness problem for automata on finite trees.

TABLE 8.2
Computation of $C^{\mathcal{A}(\hookrightarrow R_\alpha)\restriction\alpha}$.

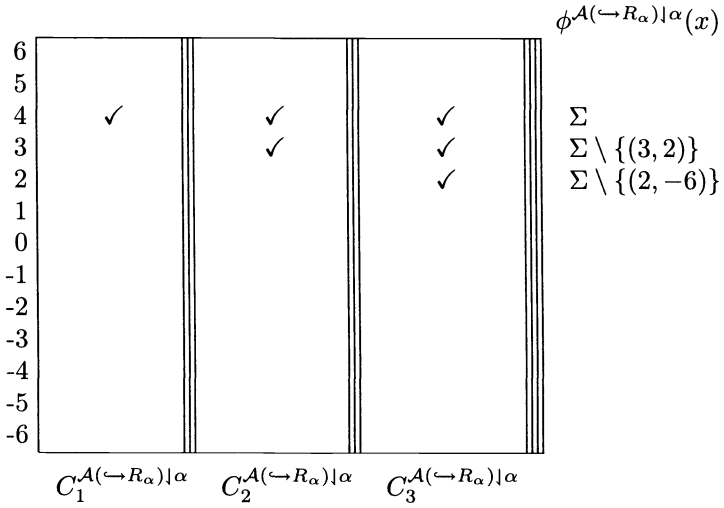
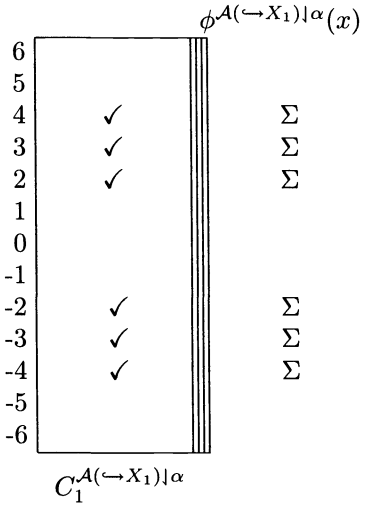


TABLE 8.3
Computation of $C^{\mathcal{A}(\hookrightarrow X_1)\restriction\alpha}$, *where* $X_1 = \{-4, -3, -2, 2, 3, 4\}$.



A key result of both of these approaches is the so-called "finite model theorem," which states roughly that a given tree automaton accepts some infinite tree if and only if it accepts an infinite tree that has a finite description; moreover, such a tree can be effectively constructed. This result was strengthened by Emerson to a "small model theorem," which states that an automaton accepts some infinite tree if and only if it accepts an infinite tree obtained by "unwinding" some finite graph embedded in its own transition structure [11]. Emerson's proof is not directly constructive.[7] The state

[7] The small model theorem does not hold for more general *Muller* automata, but Gurevich and Harrington have established a positive result stating roughly that a Muller automaton can be controlled to satisfy its own acceptance condition if and only if it can be so controlled by means of

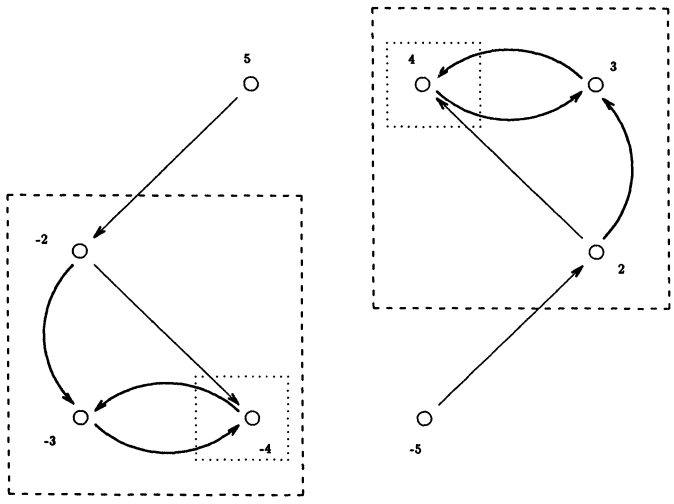


FIG. 8.4. *Transition structure representing strings generated by $\mathcal{A}$ under state feedback.*

feedback result of the current article (Theorem 6.4) is a new, constructive version of the small model theorem.

Two recent approaches to the emptiness problem are those of Pnueli and Rosner [29] and Emerson and Jutla [13]. These have established the tight upper bound on the computational complexity of the problem given in §7. Pnueli and Rosner's method was to refine the technique of Hossley and Rackoff, while Emerson and Jutla applied results on logical model-checking, expressing the acceptance condition in temporal logic, translating the resulting formula into a logical fixpoint calculus, and then checking for the existence of a model of the resulting formula embedded in the automaton's transition structure. The efficacy of this approach follows from the small model theorem.

In technical terms, the method of this paper is essentially a synthesis of those of Rabin and Emerson and Jutla, employing Rabin's method of induction on the number of "live" states and Emerson and Jutla's use of a fixpoint calculus and induction on the number of accepting pairs. The new solution is computationally simpler than Rabin's and mathematically more direct than Emerson and Jutla's: unlike those of [13], the results of this article do not depend on the small model theorem; in fact, the small model theorem is essentially a corollary of the main results of this report.

In addition to these technical advantages, the current approach has a system-theoretic flavor, which, in our opinion, renders it more transparent than the combinatorial, automata-theoretic techniques of [33], [20], [29] or the model-theoretic methods of [13]. This, together with its mathematical directness, suggests that the new technique more readily admits useful extensions. Indeed, in [41] the authors outline a generalization of the present results that allows for liveness assumptions represented by Büchi acceptance conditions. This solves an instance of a problem posed, but not constructively solved, in [1]. The methods of [20], [29] do not appear to admit such an extension.[8]

---

feedback of its own state and that of a buffer that stores the sequence of the most recent visits to the respective accepting subsets [19], [50].

[8] The treatment of liveness in Wong-Toi and Dill [49] (i.e., as a qualification of the specification) is inappropriate for our setting.

To conclude, this paper presents a direct, efficient, and natural solution to a basic problem in discrete-event system theory having applications to control synthesis, program synthesis, and logical decidability. The results illustrate fruitful interchange between control and computer science.

Acknowledgments. The reading of a preliminary version of these results by the first author's thesis examiners, particularly Professors Eric Hehner, Raymond Kwong, and Amir Pnueli, is gratefully acknowledged.

REFERENCES

[1] M. ABADI, L. LAMPORT, AND P. WOLPER, *Realizable and unrealizable specifications of reactive systems*, in Automata, Languages and Programming, 16th Internat. Colloquium, Stresa, Italy, July 1989, Proceedings (Lecture Notes in Computer Science, No. 372), Springer-Verlag, Berlin, New York, 1989, pp. 1–17.

[2] A. ARNOLD AND M. NIVAT, *Controlling behaviours of systems: Some basic concepts and some applications*, in Mathematical Foundations of Computer Science 1980 (Lecture Notes in Computer Science, No. 88), 1980, Springer-Verlag, New York, pp. 113–122.

[3] J. R. BÜCHI, *On a decision method in restricted second order arithmetic*, in Logic, Methodology and Philosophy of Science, Proc. 1960 Internat. Congress, Stanford, CA, 1962, Stanford University Press, pp. 1–11.

[4] J. R. BÜCHI AND L. H. LANDWEBER, *Solving sequential conditions by finite-state strategies*, Trans. Amer. Math. Soc., 138 (1969), pp. 295–311.

[5] Y. CHOUEKA, *Theories of automata on ω-tapes: A simplified approach*, J. Comput. System Sci., 8 (1974), pp. 117–141.

[6] A. CHURCH, *Application of logic to the problem of circuit synthesis*, in Summer Institute for Symbolic Logic, Cornell University, Ithaca, NY, 1957, pp. 3–50.

[7] ———, *Logic, arithmetic and automata*, in Proc. Internat. Congress of Mathematicians, August 15–22, 1962, Djursholm, Sweden, 1963, Institut Mittag-Leffler, pp. 23–35.

[8] J. DE BAKKER, *The fixed point approach in semantics: Theory and applications*, in Foundations of Computer Science, J. de Bakker, ed., Mathematical Centre Tracts, Amsterdam, 1975, pp. 3–53.

[9] S. EILENBERG, *Automata, Languages and Machines*, Vol. A, Academic Press, New York, 1974.

[10] E. A. EMERSON, *Characterizing correctness properties of parallel programs using fixpoints*, in Internat. Colloquium on Automata, Languages and Programming, 1980 (Lecture Notes in Computer Science, No. 85), 1980, Springer-Verlag, New York, pp. 169–181.

[11] ———, *Automata, tableaux and temporal logics*, in Logics of Programs (Lecture Notes in Computer Science, Vol. 193), R. Parikh, ed., Springer-Verlag, Berlin, New York, June 1985, pp. 79–87.

[12] ———, *Temporal and modal logic*, in Handbook of Theoretical Computer Science, Vol. B: Formal Models and Semantics, J. van Leeuwen, ed., Elsevier, The MIT Press, Cambridge, MA, 1990, pp. 995–1072.

[13] E. A. EMERSON AND C. S. JUTLA, *The complexity of tree automata and logics of programs (extended abstract)*, in 29th Ann. Sympos. on Foundations of Computer Science, White Plains, NY, Oct. 24–26, 1988, pp. 328–337.

[14] ———, *On simultaneously determinizing and complementing ω-automata (extended abstract)*, in IEEE Sympos. on Logic in Computer Science, Asilomar, CA, 1989, pp. 333–342.

[15] E. A. EMERSON AND C.-L. LEI, *Modalities for model checking: Branching time strikes back*, in Proc. 12th ACM Sympos. on Principles of Programming Languages, New Orleans, LA, 1985, pp. 84–96.

[16] ———, *Efficient model checking in fragments of the propositional mu-calculus (extended abstract)*, in Proc. of Sympos. on Logic in Computer Science, IEEE, Cambridge, MA, June 16–18, 1986, pp. 267–278.

[17] A. FUSAOKA, H. SEKI, AND K. TAKAHASHI, *A description and reasoning of plant controllers in temporal logic*, in Proc. 8th Internat. Joint Conference on Artificial Intelligence, Karlsruhe, Germany, Aug. 1983, pp. 405–408.

[18] C. GOLASZEWSKI AND P. RAMADGE, *Mutual exclusion problems for discrete event systems with shared events*, in Proc. 27th IEEE Conference on Decision and Control, Austin, TX, Dec. 7–9, 1988, pp. 234–239.

[19] Y. GUREVICH AND L. HARRINGTON, *Trees, automata and games*, in Proc. of Sympos. on the Theory of Computing, ACM, San Francisco, CA, May 5–7, 1982, pp. 60–65.

[20] R. HOSSLEY AND C. RACKOFF, *The emptiness problem for automata on infinite trees*, in Proc. of Switching and Automata Theory Sympos., IEEE, University of Maryland, Oct. 1972, pp. 121–124.

[21] D. KOZEN, *Results on the propositional μ-calculus*, Theoretical Comput. Sci., 27 (1983), pp. 333–354.

[22] R. KURSHAN, *Testing Containment of ω-Regular Languages*, Tech. Rep., AT&T Bell Laboratories, Murray Hill, NJ, Oct. 1986.

[23] ———, *Reducibility in analysis of coordination*, in Discrete Event Systems: Models and Applications, IIASA Conference, Sopron, Hungary, Aug. 3–7, 1987, (Lecture Notes in Control and Information Sciences, Vol. 103), P. Varaiya and A. Kurzhanski, eds., 1988, Springer-Verlag, New York, pp. 19–39.

[24] L. H. LANDWEBER, *Synthesis algorithms for sequential machines*, in Information Processing 68, 1969, North-Holland, Amsterdam, pp. 300–304.

[25] J.-L. LASSEZ, V. NGUYEN, AND E. SONENBERG, *Fixed point theorems and semantics: A folk tale*, Inform. Process. Let., 14 (1982), pp. 112–116.

[26] R. MCNAUGHTON, *Testing and generating infinite sequences by a finite automaton*, Inform. and Control, 9 (1966), pp. 521–530.

[27] D. E. MULLER, *Infinite sequences and finite machines*, in Proc. 4th Annual Sympos. on Switching Circuit Theory and Logical Design, IEEE, Chicago, IL, Oct. 1963, pp. 3–16.

[28] A. NERODE, A. YAKHNIS, AND V. YAKHNIS, *Concurrent programs as strategies in games*, in Logic from Computer Science: Proceedings of a Workshop held November 13-17, 1989, Mathematical Sciences Research Institute Publications Vol. 21, 1992, Springer, Berlin, New York, pp. 405–479.

[29] A. PNUELI AND R. ROSNER, *On the synthesis of a reactive module*, in Proc. 16th Annual Sympos. on Principles of Programming Languages, Association for Computing Machinery, Austin, TX, Jan. 1989, pp. 179–190.

[30] V. PRATT, *A decidable mu-calculus: Preliminary report*, in Proc. 22nd Annual Sympos. on Foundations of Computer Science, IEEE, Nashville, TN, Oct. 28–30, 1981, pp. 421–427.

[31] M. O. RABIN, *Decidability of second-order theories and automata on infinite trees*, Trans. Amer. Math. Soc., 141 (1969), pp. 1–35.

[32] ———, *Weakly definable relations and special automata*, in Mathematical Logic and Foundations of Set Theory, Y. Bar-Hillel, ed., North-Holland, Amsterdam, 1970, pp. 1–23.

[33] ———, *Automata on Infinite Objects and Church's Problem*, Conference Board of the Mathematical Sciences Regional Conference Series in Mathematics No. 13, American Mathematical Society, Providence, RI, 1972; Lectures from the CBMS Regional Conference held at Morehouse College, Atlanta, GA, September 8–12, 1969.

[34] P. RAMADGE AND W. WONHAM, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.

[35] ———, *The control of discrete event systems*, Proc. IEEE, 77 (1989), pp. 81–98.

[36] P. J. RAMADGE, *Some tractable supervisory control problems for discrete-event systems modeled by Büchi automata*, IEEE Trans. Automatic Control, 34 (1989), pp. 10–19.

[37] S. SAFRA, *On the complexity of ω-automata*, in Proc. 29th Annual Sympos. on the Foundations of Computer Science, White Plains, NY, Oct. 24–26, 1988, pp. 319–327.

[38] R. STREETT, *A Propositional Dynamic Logic of Looping and Converse*, Tech. Rep. TR-263, MIT Laboratory for Computer Science, Cambridge, MA, 1981.

[39] A. TARSKI, *A lattice-theoretical fixpoint theorem and its applications*, Pacific J. Math., 5 (1955), pp. 285–309.

[40] J. THISTLE AND W. WONHAM, *On the synthesis of supervisors subject to ω-language specifications*, in Proc. of the 1988 Conference on Information Sciences and Systems, Princeton University, Princeton, NJ, March 1988, pp. 440–444.

[41] ———, *Control of ω-automata, Church's problem, and and the emptiness problem for tree ω-automata*, in Computer Science Logic: 5th Workshop, CSL '91, Berne, Switzerland, October 1991, Proceedings (Lecture Notes in Computer Science, Vol. 626), E. Börger, G. Jäger, H. K. Büning, and M. Richter, eds., Springer-Verlag, Berlin, Heidelberg, 1992, pp. 367–381.

[42] ———, *Supervision of infinite behaviour of discrete-event systems*, SIAM J. Control Optim., 32 (1994), pp. 1098–1113, this issue.

[43] J. G. THISTLE, *Control of Infinite Behaviour of Discrete-Event Systems*, Ph.D. thesis, University of Toronto, Toronto, Canada, Jan. 1991; Systems Control Group Report No. 9012, Systems Control Group, Dept. of Electrical Engineering, University of Toronto, January

1991.

[44] W. THOMAS, *Automata on infinite objects*, in Handbook of Theoretical Computer Science, Vol. B: Formal Models and Semantics, J. van Leeuwen, ed., Elsevier, The MIT Press, Cambridge, MA, 1990, pp. 134–191.

[45] M. Y. VARDI, *A temporal fixpoint calculus (extended abstract)*, in Proc. 15th Annual ACM SIGACT-SIGPLAN Sympos. on Principles of Programming Languages, San Diego, CA, Jan. 1988, pp. 250–259.

[46] ———, *Verification of concurrent programs: The automata-theoretic framework*, Ann. Pure Appl. Logic, 51 (1991), pp. 79–98.

[47] M. Y. VARDI AND P. WOLPER, *An automata-theoretic approach to automatic program verification (preliminary report)*, in Proc. 1986 IEEE Sympos. on Logic in Computer Science, Cambridge, MA, June 16–18, 1986, pp. 332–344.

[48] P. WOLPER, *Temporal logic can be more expressive*, Inform. Control, 56 (1983), pp. 72–99.

[49] H. WONG-TOI AND D. L. DILL, *Synthesizing processes and schedulers from temporal specifications*, in Computer-Aided Verification (Proc. of the CAV 90 Workshop) DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 3, American Mathematical Society, Providence, RI, 1991, pp. 272–281.

[50] A. YAKHNIS AND V. YAKHNIS, *Extension of Gurevich-Harrington's restricted memory determinancy theorem*, Ann. Pure Appl. Logic, 48 (1990), pp. 277–297.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1098–1113, July 1994

012

SUPERVISION OF INFINITE BEHAVIOR OF DISCRETE-EVENT SYSTEMS*

J. G. THISTLE† AND W. M. WONHAM‡

Abstract. Some basic results of supervisory control theory are extended to the setting of ω-languages, formal languages consisting of infinite strings. The extension permits the investigation of both liveness and safety issues in the control of discrete-event systems. A new controllability property appropriate to the infinitary setting (*ω-controllability*) is defined; this language property captures in a natural way the limitations of available control actions. It is shown that every specification language contains a unique maximal ω-controllable sublanguage, representing the least upper bound of the set of achievable closed-loop sublanguages. This supremal ω-controllable sublanguage allows a simple formulation of necessary and sufficient conditions for the solvability of an infinitary supervisory control problem.

The problems of effectively deciding solvability of the control problem and of effectively synthesizing appropriate supervisors are solved for the case where the plant is represented by a deterministic Büchi automaton and the specification of legal behavior by a deterministic Rabin automaton.

Key words. discrete-event systems, supervisory control, controllable languages, synthesis, ω-languages, ω-automata

AMS subject classifications. 93B50, 93B99, 68Q45, 68Q60, 93A30

1. Introduction.

This paper extends basic results of the supervisory control theory of Ramadge and Wonham [17] and others to infinite-string languages and the corresponding finite automata on infinite strings; in particular, it generalizes results of [20] to the case in which specification languages need not be topologically closed relative to plant behavior.

Infinite-string languages, or ω-languages, provide richer models and specifications of DES than do their finite-string counterparts [3], [5], [19], [20]. Moreover, automata on infinite inputs, or ω-automata, form the basis of an extensive theory of automaton synthesis [26] having applications to control [24].

The use of ω-languages and ω-automata in modeling and specifying dynamic systems is well established. First proposed by Muller [15] as a means of describing the infinite behavior of asynchronous switching circuits, they have since been applied to studies of digital hardware and computer software (see, for example, [8]) and, to a smaller extent, discrete-event control systems [27], [13], [20], [11], [22], [2]. Such languages provide a natural means of modeling nonterminating systems and perhaps more importantly, they offer greater expressive power than $*$-languages [15], [3], [5].

This difference in expressive power is best described in the terminology of software verification, whereby a "safety" property is one that states that some condition(s) will *not* occur (ever), and a "liveness" property states that some condition(s) *must* occur (eventually) [14]. (In control-theoretic terms, safety corresponds roughly to invariance or stability; liveness is comparable to reachability or asymptotic stability.)

* Received by the editors July 19, 1991; accepted for publication (in revised form) January 20, 1993.

† Département de génie electrique et de génie informatique, Ecole Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Canada H3C 3A7. The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant OGP0007399 and a Postdoctoral Fellowship.

‡ Systems Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Canada M5S 1A4. The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant OGP0007399.

Whereas safety properties can be specified in terms of $*$-languages, the expression of liveness properties requires the use of ω-languages. Indeed, according to Alpern and Schneider's formal definitions, a safety property is one that can be expressed as a condition on the set of finite event sequences, while a liveness property is one that restricts only the set of infinite event sequences [1].

This paper does not represent the first use of ω-languages and ω-automata within the context of the control theory of Ramadge and Wonham. Ramadge and Golaszewski have already employed ω-language models in this setting [17]–[20], [11], but their main application of ω-languages is in the modeling of the uncontrolled system (namely, in the expression of fairness assumptions); their specifications of controlled behavior are all safety properties. The same applies to the study of Kumar, Garg, and Marcus [12], [13]. Young, Spanjol, and Garg consider systems modeled and specified by deterministic Büchi automata [28], [29]; the results of this paper are more general [22], [25].

Section 2 discusses the necessary preliminaries from the theory of formal languages. Section 3 describes a model of a discrete-event system (DES) as a controlled language generator and defines the property of *deadlock-freedom* [19], [20]. Section 4 defines the key language property of *ω-controllability*, which serves to characterize the limitations of available control actions. This infinitary controllability property is stronger than the essentially finitary notion defined elsewhere and is particularly useful in studying the supervision of infinite behavior, as in §5. Here, a natural ω-language analogue of the original supervisory control problem of [16] is considered, allowing the specification of liveness as well as safety properties.

The effective solution of the synthesis problem is considered in §6, where it is assumed that specifications of legal behavior are represented by deterministic Rabin automata and DESs are modeled as deterministic Büchi automata [20]. More general classes of system models will be considered in future reports.

The results of the article are summarized and compared with related work in §7.

2. Language preliminaries. This section establishes notation and terminology for formal languages. For further background material, refer to [26], [4], [6].

Let Σ be a finite alphabet. Then Σ^* denotes the set of all finite strings over Σ, including the empty string 1. The expression Σ^ω represents the set of infinite strings over Σ. The union of Σ^* and Σ^ω is denoted by Σ^∞.

Any $L \subseteq \Sigma^*$ is called a *$*$-language* over Σ, and any $S \subseteq \Sigma^\omega$ is an *ω-language* over Σ. When dealing with singleton languages, we generally omit braces if such omission is unlikely to lead to confusion; thus $\{v\} \subset \Sigma^\infty$ is typically represented by v.

For any $k \in \Sigma^*$ and $v \in \Sigma^\infty$, kv denotes the catenation of the two strings. If K is a $*$-language and V is a $*$- or ω-language, then $KV := \{kv \in \Sigma^\infty : k \in K \text{ and } v \in V\}$ is called the *product* of K and V. The *quotient*[1] V/K is defined by $V/K := \{w \in \Sigma^\infty : (\text{there exists } k \in K)(kw \in V)\}$.

The *Kleene closure* K^* of a $*$-language $K \subseteq \Sigma^*$ is given by $K^* := \bigcup_{i=0}^{\infty} K^i$, where K^i denotes the i-fold product of K with itself. (K^0 denotes $\{1\}$.) Thus K^* is the set of all catenations of strings in K. The *ω-Kleene closure* K^ω of a $*$-language $K \subseteq \Sigma^*$ is the "infinite product" of K with itself, that is, the set of all strings $s = k_1 k_2 k_3, \ldots, k_i \in K \setminus \{1\}$.

[1] The notation V/K for the quotient is not to be confused with other usage whereby $V/K = \{w \in \Sigma^* : (\exists k \in K)(wk \in V)\}$, that is, where the order of k and w in the concatenation is reversed.

For any $k \in \Sigma^*$, $v \in \Sigma^\infty$, we write $k \leq v$ if k is a *prefix* of v, i.e., if there exists $t \in \Sigma^\infty$ such that kt is identical to v. Define the map pre: $2^{\Sigma^\infty} \longrightarrow 2^{\Sigma^*}$ by pre: $V \mapsto \{k \in \Sigma^* : (\text{there exists } v \in V)(k \leq v)\}$. Thus pre(V) is the set of all finite prefixes of strings in V. For any $R \subseteq \Sigma^\omega$, we call pre$(R) \subseteq \Sigma^*$ the *prefix* of R. For any $K \subseteq \Sigma^*$, we call pre(K) the $*$-*closure* of K. If $K = \text{pre}(K)$, we say that K is $*$-*closed.*

The *limit* [7] of a $*$-language is given by[2] $\lim(K) := \text{pre}^{-1}(K) \cap \Sigma^\omega$, where pre^{-1}: $2^{\Sigma^*} \to 2^{\Sigma^\infty}$ is the inverse of pre: $2^{\Sigma^\infty} \to 2^{\Sigma^*}$, i.e., $\text{pre}^{-1}(K) := \{v \in \Sigma^\infty : \text{pre}(v) \subseteq K\}$. Thus the limit of $K \subseteq \Sigma^*$ is the set of all infinite strings whose prefixes are all contained in K. For example, if K is $\alpha^*[1 \cup \beta]$, then $\lim(K) = \alpha^\omega$.

The operator clo : $2^{\Sigma^\omega} \longrightarrow 2^{\Sigma^\omega}$ is defined by

$$\text{clo}: \ R \mapsto \lim(\text{pre}(R)) = \text{pre}^{-1}(\text{pre}(R)) \cap \Sigma^\omega.$$

In other words, clo(R) is the set of all infinite strings, all of whose prefixes are contained in pre(R). Thus, if $R = \alpha^*\beta^\omega$, then clo$(R) = \alpha^\omega \cup \alpha^*\beta^\omega$. We call clo$(R)$ the ω-*closure* of R. If $R = \text{clo}(R)$, we say that R is ω-*closed*; if $R = \text{clo}(R) \cap S$ (where $S \subseteq \Sigma^\omega$), we say that R is closed *relative to* or *with respect to* S.[3] It can be seen from the definition of clo : $2^{\Sigma^\omega} \longrightarrow 2^{\Sigma^\omega}$ that ω-closed languages are completely determined by their prefixes.

3. Discrete-event systems and supervisors. The following sections describe a basic model (essentially due to Ramadge [20]) of DESs as controlled generators of finite and infinite event sequences.

3.1. Discrete-event systems. We model a DES as a pair $\mathbf{G} = (L, S) \in 2^{\Sigma^*} \times 2^{\Sigma^\omega}$ consisting of a $*$-language L called the $*$-*behavior* of $\mathbf{G}$ and an ω-language S called the ω-*behavior* of $\mathbf{G}$ [19], [20].[4]

The $*$-behavior L is assumed to be $*$-closed, i.e., pre$(L) = L$. We also assume that pre$(S) \subseteq L$. If the reverse inclusion holds (that is, if pre$(S) = L$), then we say that the DES (L, S) is *deadlock-free*.[5]

3.2. Supervisors. We adjoin to the DES model the control feature proposed in [10]; namely, we associate with the alphabet Σ a nonempty family $\mathbf{C} \subseteq 2^\Sigma$ of *control patterns*. A *supervisor* is a partial function $f : \Sigma^* \to \mathbf{C}$.

We assume that $\mathbf{C}$ is closed under union, that is, if $\Gamma, \Gamma' \in \mathbf{C}$, then $\Gamma \cup \Gamma' \in \mathbf{C}$. This property ensures the existence of supremal controllable sublanguages but entails no loss of generality, in the sense that control schemes devised under the assumption can always be implemented nondeterministically [10].

---

[2] This definition, due to Elgot, should not be confused with similar ones appearing in the literature. Elgot's *limit* [7] of a $*$-language is a subset of Eilenberg's *closure* [6], which in turn is contained in Boasson and Nivat's *adherence* [4]. (The three definitions coincide when applied to $*$-closed languages.)

[3] Our definition of ω-closure is equivalent to that of Ramadge [20], which involves a metric; see [4].

[4] We use the symbol L, for "language," to represent $*$-behavior in a manner consistent with the usual notation of the Ramadge–Wonham theory; we follow Ramadge in letting S, for "sequences," represent ω-behavior. These symbols should not be misconstrued as standing, respectively, for liveness and safety. On the contrary, the $*$-behavior L is the more closely connected with safety properties and S with liveness. The authors are grateful to Professor Amir Pnueli for pointing out this potential source of confusion.

[5] Ramadge [20] uses the term "nonblocking."

For any DES $\mathbf{G} = (L, S)$ and any supervisor f, the *controlled discrete-event system* $\mathbf{G}^f$, representing the action of the supervisor $f : L \to \mathbf{C}$ on the DES $\mathbf{G} = (L, S)$, is given by $\mathbf{G}^f = (L^f, S^f)$, where

(i) L^f, the $*$-language *synthesized by* f, is defined by the following recursion:[6]

(a) $1 \in L^f$,

(b) For all $k \in \Sigma^*, \sigma \in \Sigma$

$$k\sigma \in L^f \iff k \in L^f \cap \mathrm{dom}(f) \quad \text{and} \quad k\sigma \in L \quad \text{and} \quad \sigma \in f(k);$$

(ii) S^f, the ω-language *synthesized by* f, is given by $S^f := \lim(L^f) \cap S$.

The definition of L^f means that a sequence of events $k\sigma$ can occur under supervision if and only if the sequence k can occur under supervision, and, once it has, the event σ can take place without violating either the "physical" constraints embodied by L or the control pattern imposed by the supervisor. This interpretation of L^f is valid only if the map f is defined for all strings in L^f. We therefore say that a map $f : L \to \mathbf{C}$ is a *complete*[7] supervisor for the DES (L, S) if and only if $L^f \subseteq \mathrm{dom}(f)$. In the following, we deal exclusively with complete supervisors.

PROPOSITION 3.1. *For any* DES $\mathbf{G}$ *and any supervisor* $f : L \to \mathbf{C}$, $\mathbf{G}^f = (L^f, S^f)$ *is indeed a* DES, *i.e.,* $\mathrm{pre}(L^f) = L^f$ *and* $\mathrm{pre}(S^f) \subseteq L^f$. *Furthermore, the* $*$- *and* ω*-behaviors of* $\mathbf{G}^f$ *are sublanguages of those of* $\mathbf{G}$; *that is,* $L^f \subseteq L$ *and* $S^f \subseteq S$; S^f *is* ω*-closed relative to* S.

We say that $f : L \to \mathbf{C}$ is a *deadlock-free* supervisor for $\mathbf{G} = (L, S)$ if $\mathbf{G}^f$ is a deadlock-free DES.

4. Closed-loop behavior and controllability. The following sections characterize achievable closed-loop system behavior by identifying the $*$- and ω-languages that can be synthesized by supervisors.

4.1. Closed-loop $*$-behaviors and $*$-controllability. We begin with a review of Golaszewski and Ramadge's results on the synthesis of $*$-languages [10].

For any $*$-language $V \subseteq \Sigma^\infty$ and any $s \in \mathrm{pre}(V)$, the *active set of* V *after* s, $\Sigma_V(s) \subseteq \Sigma$, is given by

$$\Sigma_V(s) := \Sigma \cap (\mathrm{pre}(V)/s).$$

Thus $\Sigma_V(s)$ is the set of all $\sigma \in \Sigma$ such that $s\sigma \in \mathrm{pre}(V)$.

Given languages $V \subseteq \Sigma^\infty$, $L \subseteq \Sigma^*$ with $\mathrm{pre}(V) \subseteq L$, V is $*$-*controllable* with respect to L if and only if

$$\forall s \in \mathrm{pre}(V)\ \exists \Gamma \in \mathbf{C} : \Gamma \cap \Sigma_L(s) = \Sigma_V(s).$$

In other words, V is $*$-controllable with respect to L if and only if the extensions $s\sigma \in s\Sigma \cap \mathrm{pre}(L)$ of any $s \in \mathrm{pre}(V)$ can be restricted through control to exactly those $s\sigma$ that belong to $\mathrm{pre}(V)$.

PROPOSITION 4.1 (Golaszewski–Ramadge–Wonham). *For any* DES $\mathbf{G} = (L, S)$ *and any nonempty* $*$*-language* $M \subseteq L$, *there exists a complete supervisor for* $\mathbf{G}$ *that synthesizes* M *if and only if* M *is* $*$*-controllable with respect to* L *and* $*$*-closed.*

[6] This definition differs slightly from that of [20].

[7] This is not an exact analogue of the automaton-based definition of [16], but it has a similar interpretation.

4.2. Closed-loop ω-behaviors: $*$- and ω-controllability. The $*$-controllability property also characterizes the class of ω-languages that are synthesized by deadlock-free supervisors.

PROPOSITION 4.2 (Ramadge [20]). *For any* DES $\mathbf{G} = (L, S)$ *and any nonempty* $T \subseteq S$, *there exists a complete, deadlock-free supervisor* f *for* $\mathbf{G}$ *that synthesizes* T *if and only if* T *is* $*$*-controllable with respect to* L *and* ω*-closed with respect to* S.

The $*$-controllability of an ω-language $T \subseteq S$ means that all infinite extensions of strings in $\mathrm{pre}(T)$ can be controlled to belong to the ω-closure of T relative to S. A more intuitively satisfying notion of controllability would imply that all such infinite extensions could be controlled to belong to T itself, regardless of whether T were ω-closed relative to S. In this section, we define such a property, called *ω-controllability*. The new property is useful in characterizing the solvability of supervisor synthesis problems, mainly because it separates the issue of controllability from that of closure (see §5).

We first define the *controllability prefix* $\mathrm{pre}_{\mathbf{G}}(T)$ of an ω-language T. For any DES $\mathbf{G} = (L, S)$, define

$$\begin{aligned} \mathrm{pre}_{\mathbf{G}} : \quad & 2^S \to 2^{\mathrm{pre}(S)} \\ & T \mapsto \{t \in \mathrm{pre}(T) \ : \ (\exists T' \subseteq T/t) \\ & \qquad [T' \neq \emptyset \text{ is } * \text{-controllable w.r.t. } L/t \\ & \qquad\qquad \text{and } \omega\text{-closed w.r.t. } S/t] \ \}. \end{aligned}$$

By Proposition 4.2, $\mathrm{pre}_{\mathbf{G}}(T)$ represents the set of all $*$-strings in $\mathrm{pre}(T)$ whose infinite extensions can be controlled to belong to T.

In an alternative interpretation of $\mathrm{pre}_{\mathbf{G}}(T)$, the operation of a controlled system is viewed as an infinite game between supervisor and DES (where the supervisor wins just in case the ω-string generated belongs to T). Then $\mathrm{pre}_{\mathbf{G}}(T)$ is the set of "winning positions" for the supervisor.

PROPOSITION 4.3. *For any* DES (L, S) *and any* $T, T' \subseteq S$ *and* $t \in \mathrm{pre}(T)$,

(a) $\quad T \subseteq T' \Longrightarrow \mathrm{pre}_{\mathbf{G}}(T) \subseteq \mathrm{pre}_{\mathbf{G}}(T')$;

(b) $\quad \mathrm{pre}_{(L/t, S/t)}(T/t) = (\mathrm{pre}_{(L,S)}(T))/t$;

(c) *If* T *is* $*$*-controllable w.r.t.* L *and* ω*-closed w.r.t.* S,
$$\mathrm{pre}_{\mathbf{G}}(T) = \mathrm{pre}(T);$$

(d) $\quad \forall k \in \mathrm{pre}_G(T) : \exists \Gamma \in \mathbf{C} : \Gamma \cap \Sigma_L(k) \subseteq \Sigma_{\mathrm{pre}_G(T)}(k)$.

Proof. (a) The proof follows by definition.
(b) It holds that

$$\begin{aligned} & t' \in (\mathrm{pre}_{(L,S)}(T))/t \\ \Longleftrightarrow \ & tt' \in \mathrm{pre}_{(L,S)}(T) \\ \Longleftrightarrow \ & tt' \in \mathrm{pre}(T) \quad \text{and} \quad \exists T' \subseteq T/tt' \ : \ T' \neq \emptyset \text{ is} \\ & *\text{-controllable w.r.t. } L/tt', \omega\text{-closed w.r.t. } S/tt' \\ \Longleftrightarrow \ & t' \in \mathrm{pre}(T/t) \quad \text{and} \quad \exists T' \subseteq (T/t)/t' \ : \ T' \neq \emptyset \text{ is} \\ & *\text{-controllable w.r.t. } (L/t)/t', \omega\text{-closed w.r.t. } (S/t)/t' \\ \Longleftrightarrow \ & t' \in \mathrm{pre}_{(L/t, S/t)}(T/t). \end{aligned}$$

(c) It is easily shown that $*$-controllability and ω-closure are "preserved under quotients" in the following sense:

$$\begin{aligned} T \ *\text{-controllable w.r.t. } L &\implies T/t \ *\text{-controllable w.r.t. } L/t, \\ T \ \omega\text{-closed w.r.t. } S &\implies T/t \ \omega\text{-closed w.r.t. } S/t, \end{aligned}$$

where $T \subseteq \Sigma^\omega$. The result follows.

(d) Suppose that $t \in \mathrm{pre}_{\mathbf{G}}(T)$. Then there exists a nonempty $T' \subseteq T/t$ that is $*$-controllable with respect to L/t and ω-closed with respect to S/t. By $*$-controllability, there exists $\Gamma \in \mathbf{C}$ such that

$$\begin{aligned} & \Gamma \cap \Sigma_{L/t}(1) = \Sigma_{\mathrm{pre}(T')}(1) \\ \iff \quad & \Gamma \cap L/t = \Sigma \cap \mathrm{pre}(T') \\ \iff \quad & \Gamma \cap L/t = \Sigma \cap \mathrm{pre}_{(L/t,S/t)}(T') && \text{(part (c))} \\ \implies \quad & \Gamma \cap L/t \subseteq \Sigma \cap (\mathrm{pre}_{(L/t,S/t)}(T/t)) && \text{(part (a))} \\ \iff \quad & \Gamma \cap L/t \subseteq \Sigma \cap (\mathrm{pre}_{(L,S)}(T))/t && \text{(part (b))} \\ \iff \quad & \Gamma \cap \Sigma_L(t) \subseteq \Sigma_{\mathrm{pre}_{\mathbf{G}}(T)}(t). && \square \end{aligned}$$

A natural controllability property would require the supervisor to be always in a winning position. For any DES $\mathbf{G} = (L, S)$ and any $T \subseteq S$, T is *ω-controllable* with respect to $\mathbf{G}$ if T is $*$-controllable with respect to L and $\mathrm{pre}(T) = \mathrm{pre}_{\mathbf{G}}(T)$. (In the case where $\mathbf{C}$ is closed under containment (i.e., $\Gamma \in \mathbf{C}$ and $\Gamma \subseteq \Gamma' \implies \Gamma' \in \mathbf{C}$), as in the original Ramadge–Wonham theory, $\mathrm{pre}(T) = \mathrm{pre}_{\mathbf{G}}(T)$ implies $*$-controllability of T, by Proposition 4.3(d).)

Let $\Sigma = \{\alpha, \beta\}$, $L = \alpha^*\beta^*$, $S = \lim(L) = \{\alpha^\omega\} \cup \alpha^*\beta^\omega$, and $E = \alpha^*\beta^\omega$. If $\mathbf{C} = \{\Gamma \subseteq \Sigma : \alpha \in \Gamma\}$, then E is $*$-controllable with respect to L, but not ω-controllable with respect to (L, S) (as $\mathrm{pre}_{(L,S)}(E) = \emptyset$).

By Proposition 4.3(c), the two properties coincide for languages that are ω-closed relative to S.

PROPOSITION 4.4. *For any* DES (L, S) *and any* $T \subseteq S$, *if* T *is* ω*-closed with respect to* S,

$$T \text{ is } \omega\text{-controllable w.r.t. } (L, S) \iff T \text{ is } *\text{-controllable w.r.t. } L.$$

By Proposition 4.4, we may replace $*$-controllability with ω-controllability in Proposition 4.2.

PROPOSITION 4.5. *For any* DES (L, S) *and any* $T \subseteq S$, *there exists a complete, deadlock-free supervisor for* (L, S) *that synthesizes* T *if and only if* T *is* ω*-controllable with respect to* (L, S) *and* ω*-closed with resepct to* S.

This new characterization of achievable closed-loop behavior provides the key to explaining solvability of the supervisor synthesis problems of the next section.

5. Supervisor synthesis. We next discuss an ω-language analogue of the supervisory control problem of [16]. The original $*$-language problem is of basic importance in supervisor synthesis and has provided a foundation for numerous extensions.

5.1. The supervisory control problem for ω-languages (SCP$^\omega$).

PROBLEM 5.1 (SCP$^\omega$). *Given a* DES (L, S) *and* ω*-languages* $A, E \subseteq \Sigma^\omega$ *such that* $A \subseteq E \subseteq S$, *construct a complete, deadlock-free supervisor* f *for* (L, S) *such that* $A \subseteq S^f \subseteq E$.

The requirement that the supervisor be deadlock-free does not limit the generality of the problem. It is used to eliminate solutions in which containment in the maximal

legal sublanguage is satisfied vacuously; in cases where deadlocks are to be allowed, problems can be suitably recast; see [25] for an example.

By Proposition 4.5, solvability of SCP$^\omega$ is equivalent to the existence of a nonempty sublanguage $T \subseteq S$, ω-controllable with respect to (L, S) and ω-closed with respect to S, such that $A \subseteq T \subseteq E$. The ω-controllable and the ω-closed languages have different closure properties under union and intersection. Specifically, ω-controllability is preserved under arbitrary unions but not intersections, and ω-closure is preserved under arbitrary intersections but not arbitrary unions. It is therefore convenient to define, below, the following separate language classes.

For any DES $\mathbf{G} = (L, S)$ and any ω-languages $A \subseteq E \subseteq S$,

$$\mathcal{C}^\omega(E) := \{T \subseteq S : T \subseteq E \subseteq S \text{ and } T \text{ is } \omega\text{-controllable w.r.t. } \mathbf{G}\},$$
$$\mathcal{F}^\omega(A) := \{T \subseteq S : A \subseteq T \subseteq S \text{ and } T \text{ is } \omega\text{-closed w.r.t. } S\}.$$

With respect to the DES $\mathbf{G}$, $\mathcal{C}^\omega(E)$ is the class of ω-controllable sublanguages of E, and $\mathcal{F}^\omega(A)$ is the class of ω-closed superlanguages of A. We call an ω-language a *solution* to SCP$^\omega$ if it is a nonempty element of $\mathcal{C}^\omega(E) \cap \mathcal{F}^\omega(A)$.

We also bring in the $*$-language class of [16]. For any $*$-language L and any $L' \subseteq L$,

$$\mathcal{CF}^*(L') := \{K \subseteq L' : K \text{ is } *\text{-controllable w.r.t. } L \text{ and } *\text{-closed}\}.$$

The class $\mathcal{CF}^*(L')$ contains a supremal element $\sup \mathcal{CF}^*(L')$ [16]. For the ω-language classes, we have the following result.

PROPOSITION 5.2. *For any* DES $\mathbf{G} = (L, S)$ *and any* ω*-languages* $A \subseteq E \subseteq S$,

(a) $\sup \mathcal{C}^\omega(E)$ *exists in* $\mathcal{C}^\omega(E)$, *namely,*

$$\sup \mathcal{C}^\omega(E) = \lim(\sup \mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E))) \cap E;$$

(b) $\inf \mathcal{F}^\omega(A)$ *exists in* $\mathcal{F}^\omega(A)$, *namely,*

$$\inf \mathcal{F}^\omega(A) = \mathrm{clo}(A) \cap S.$$

Proof. Part (b) is clear, since clo is a Kuratowski closure operator. For part (a), let

$$E' := \lim(\sup \mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E))) \cap E.$$

We show that E' is the supremal element of $\mathcal{C}^\omega(E)$. To establish ω-controllability we apply the following claim.

CLAIM 1. *It holds that*

$$\sup \mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E)) \subseteq \mathrm{pre}_{\mathbf{G}}(E').$$

Suppose that $k \in \sup \mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E)) \subseteq \mathrm{pre}_{\mathbf{G}}(E)$. Then there exists a nonempty $E_k \subseteq E/k$, $*$-controllable with respect to L/k and ω-closed with respect to S/k. Since k is arbitrary, it suffices to show that $E_k \subseteq E'/k$.

Let $k' \in \mathrm{pre}(E_k)$. Then $E_k/k' \subseteq E/kk'$ is nonempty, $*$-controllable with respect to L/kk', and ω-closed with respect to S/kk'. We therefore have $kk' \in \mathrm{pre}_{\mathbf{G}}(E)$. Because k' is arbitrary, this means that

$$k\mathrm{pre}(E_k) \subseteq \mathrm{pre}_{\mathbf{G}}(E).$$

Now

$$\begin{aligned}
& E_k \subseteq E/k \text{ is } *\text{-controllable w.r.t. } L/k \\
\Longrightarrow\ & \sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E)) \cup k\mathrm{pre}(E_k) \text{ is } *\text{-controllable w.r.t. } L \text{ and } *\text{-closed} \\
& \qquad (\mathbf{C} \text{ is closed under union; } k \in \sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E))) \\
\Longleftrightarrow\ & k\mathrm{pre}(E_k) \subseteq \sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E)) \qquad (k\mathrm{pre}(E_k) \subseteq \mathrm{pre}_{\mathbf{G}}(E)) \\
\Longleftrightarrow\ & kE_k \subseteq \lim(\sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E))) \cap E \qquad (k \in \sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E))).
\end{aligned}$$

Thus $E_k \subseteq E'/k$. This completes the proof of the claim.

We now have

$$\begin{aligned}
\mathrm{pre}(E') &\subseteq \sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E)) \qquad \text{(by definition)} \\
&\subseteq \mathrm{pre}_{\mathbf{G}}(E') \qquad \text{(Claim 1)} \\
&\subseteq \mathrm{pre}(E') \qquad \text{(by definition)}.
\end{aligned}$$

Thus

$$\mathrm{pre}(E') = \sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E)) = \mathrm{pre}_{\mathbf{G}}(E'),$$

so E' is ω-controllable with respect to $\mathbf{G}$.

Now suppose that $E'' \in \mathcal{C}^\omega(E)$. Then

$$\begin{aligned}
E'' &\subseteq \mathrm{clo}(E'') \cap E \\
&= \lim(\mathrm{pre}(E'')) \cap E \\
&= \lim(\sup\mathcal{CF}^*(\mathrm{pre}(E''))) \cap E \qquad (*\text{-controllability}) \\
&= \lim(\sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E''))) \cap E \qquad (\omega\text{-controllability}) \\
&\subseteq E'.
\end{aligned}$$

This establishes that E' is indeed the unique maximal element of $\mathcal{C}^\omega(E)$. □

In the special case where $\mathbf{C}$ is closed under containment, $\sup\mathcal{C}^\omega(E) = \lim(\mathrm{pre}_{\mathbf{G}}(E)) \cap E$ (by Proposition 4.3 (d)) [25]. The solvability of SCP$^\omega$ can be characterized in terms of the extremal elements of $\mathcal{C}^\omega(E)$ and $\mathcal{F}^\omega(A)$.

THEOREM 5.3. SCP$^\omega$ *is solvable if and only if* $\sup\mathcal{C}^\omega(E) \neq \emptyset$ *and*

$$\inf\mathcal{F}^\omega(A) \subseteq \sup\mathcal{C}^\omega(E)\ .$$

Proof. Necessity follows from Proposition 4.5. For sufficiency, let

$$A' := \inf\mathcal{F}^\omega(A) \subseteq \sup\mathcal{C}^\omega(E) =: E'$$

and suppose that $E' \neq \emptyset$.

Because E' is $*$-controllable and nonempty, there exists a complete, deadlock-free supervisor $f_0 : \Sigma^* \to \Gamma$ that synthesizes $\mathrm{clo}(E') \cap S$, by Proposition 4.2.

Because E' is ω-controllable, there exists for every $m \in \mathrm{pre}(E')$ a nonempty ω-language $E'_m \subseteq E'/m$, $*$-controllable with respect to L/m and ω-closed with respect to S/m; let $f_m : \Sigma^* \to \Gamma$ be a corresponding complete, deadlock-free supervisor.

Let M be the set of all elements of $\mathrm{pre}(E') \setminus \mathrm{pre}(A')$ of minimal length. Define the following supervisor $f : \Sigma^* \to \Gamma$:

$$f(s) := \begin{cases} f_0(s) & \text{if } s \in \mathrm{pre}(A'), \\ f_m(s/m) & \text{if } s \in m\ \mathrm{pre}(E'_m), \text{ where } m \in M, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Note that f is thus well defined, since the sets $m\ \mathrm{pre}(E'_m)$, $m \in M$ are pairwise disjoint.

We claim that f is a complete, deadlock-free supervisor for (L, S) and that $A' \subseteq S^f \subseteq E'$. The result is a consequence of the following claim.

CLAIM 2. *It holds that*

$$\text{(a)} \qquad L^f = \mathrm{pre}(A') \cup \bigcup_{m\in M} m\mathrm{pre}(E'_m),$$
$$\text{(b)} \qquad S^f = A' \cup \bigcup_{m\in M} mE'_m.$$

(a) We show by induction on the length of strings that

$$k \in L^f \iff k \in \mathrm{pre}(A') \cup \bigcup_{m\in M} m\mathrm{pre}(E'_m).$$

The result holds by definition when k is the empty string. For the induction step, suppose that the result holds for k. Then

$$\begin{aligned}
& k\sigma \in L^f \\
\iff\ & k\sigma \in L, k \in L^f \text{ and } \sigma \in f(k) \\
\iff\ & k\sigma \in L, k \in \mathrm{pre}(A') \cup \textstyle\bigcup_{m\in M} m\mathrm{pre}(E'_m) \text{ and } \sigma \in f(k) \quad \text{(induction hypothesis)} \\
\iff\ & k\sigma \in L, [(k \in \mathrm{pre}(A') \text{ and } \sigma \in f_0(k)) \\
& \qquad \text{or } (\exists m \in M)(k \in m\mathrm{pre}(E'_m) \text{ and } \sigma \in f_m(k/m))] \\
\iff\ & k\sigma \in L, [(k \in L^{f_0} \cap \mathrm{pre}(A') \text{ and } \sigma \in f_0(k)) \\
& \qquad \text{or } (\exists m \in M)(k \in m(L/m)^{f_m} \text{ and } \sigma \in f_m(k/m))] \\
& \qquad\qquad \text{(deadlock-freedom of } f_m) \\
\iff\ & k\sigma \in L^{f_0} \cap \mathrm{pre}(A')\Sigma \text{ or } (\exists m \in M)(k\sigma/m \in (L/m)^{f_m}) \\
\iff\ & k\sigma \in \mathrm{pre}(A') \text{ or } (\exists m \in M)(k\sigma/m \in \mathrm{pre}(E'_m)) \\
& \qquad\qquad \text{(deadlock-freedom of } f_0 \text{ and } f_m) \\
\iff\ & k\sigma \in \mathrm{pre}(A') \cup \textstyle\bigcup_{m\in M} m\mathrm{pre}(E'_m).
\end{aligned}$$

(b) First, note that

$$\lim(L^f) = \mathrm{clo}(A') \cup \bigcup_{m\in M} m\mathrm{clo}(E'_m)\ .$$

The inclusion ($\supseteq$) is easily proved. For the reverse, suppose that $s \in \lim(L^f)$. Then

$$\mathrm{pre}(s) \subseteq L^f = \mathrm{pre}(A') \cup \bigcup_{m\in M} m\mathrm{pre}(E'_m).$$

Now, if $\mathrm{pre}(s) \subseteq \mathrm{pre}(A')$, then $s \in \mathrm{clo}(A')$, and the inclusion holds; otherwise, $\mathrm{pre}(s) \subseteq \mathrm{pre}(m) \cup m\mathrm{pre}(E'_m)$, where $m \in M$ is the shortest element of $\mathrm{pre}(s)\backslash\mathrm{pre}(A')$. In that case, $s \in m\mathrm{clo}(E'_m)$.

Thus,

$$\begin{aligned}
S^f &= \lim(L^f) \cap S \\
&= (\mathrm{clo}(A') \cup \textstyle\bigcup_{m\in M} m\mathrm{clo}(E'_m)) \cap S \\
&= (\mathrm{clo}(A') \cap S) \cup \textstyle\bigcup_{m\in M} (m\mathrm{clo}(E'_m) \cap S) \\
&= (\mathrm{clo}(A') \cap S) \cup \textstyle\bigcup_{m\in M} m(\mathrm{clo}(E'_m) \cap S/m) \\
&= A' \cup \textstyle\bigcup_{m\in M} mE'_m.
\end{aligned}$$

This completes the proof of the claim. Completeness of f follows from part (a).

The containments $A' \subseteq S^f \subseteq E'$ follow from part (b).
For deadlock-freedom,

$$\begin{aligned} \mathrm{pre}(S^f) &= \mathrm{pre}(A' \cup \bigcup_{m\in M} mE'_m) \qquad \text{(part (b))} \\ &= \mathrm{pre}(A') \cup \bigcup_{m\in M} m\mathrm{pre}(E'_m) \\ &= L^f \qquad \text{(part (a))}. \quad \square \end{aligned}$$

COROLLARY 5.4. *For any* DES (L, S) *and any* $E \subseteq S$, *the* ω*-language* $\sup \mathcal{C}^\omega(E)$ *is the least upper bound of the achievable closed-loop behaviors contained in* E, *that is,*

$$\sup \mathcal{C}^\omega(E) = \bigcup \{R \subseteq E : R \text{ is } *\text{-controllable w.r.t. } L \text{ and } \omega\text{-closed w.r.t. } S\}.$$

Proof. ($\supseteq$) The proof follows by Proposition 4.4. ($\subseteq$) Let $s \in \sup \mathcal{C}^\omega(E)$ and define $A := \{s\}$. Then, by Theorem 5.3, there exists $R \subseteq \sup \mathcal{C}^\omega(E)$, $*$-controllable with respect to L and ω-closed with respect to S, such that $A \subseteq R \subseteq E$. $\square$

Let an ω-language $R \subseteq \Sigma^\omega$ be a *maximal solution* to SCP$^\omega$ if it is a nonempty maximal element (in the sense of set inclusion) of the language class $\mathcal{C}^\omega(E) \cap \mathcal{F}^\omega(A)$.

COROLLARY 5.5. *When* SCP$^\omega$ *is solvable, it has maximal solutions if and only if* $\sup \mathcal{C}^\omega(E)$ *is* ω*-closed with respect to* S, *i.e., if and only if* $\sup \mathcal{C}^\omega(E) \in \mathcal{F}^\omega(A)$. *In this case,* $\sup \mathcal{C}^\omega(E)$ *is the unique maximal solution.*

Proof. The proof follows by Corollary 5.4, since $\mathcal{C}^\omega(E) \cap \mathcal{F}^\omega(A)$ is closed under finite unions. $\square$

COROLLARY 5.6. $\sup \mathcal{C}^\omega(E)$ *is* ω*-closed with respect to* S *whenever* E *is* ω*-closed.*

Proof. The proof follows by Proposition 5.2. $\square$

Intuitively, an ω-language $E \subseteq \Sigma^\omega$ that is ω-closed relative to S is completely determined by its finite prefixes; it represents no restriction on infinite strings beyond that implied by the restriction on finite strings to $\mathrm{pre}(E)$. Hence it is not surprising that, when E is ω-closed relative to S, the existence of maximal solutions carries over to the infinite string case (as Corollaries 5.5 and 5.6 imply). On the other hand, maximal legal sublanguages that are not ω-closed relative to S embody an additional restriction on infinite trajectories beyond that implied by the restriction on finite trajectories; in other words, such languages incorporate liveness specifications. The nonexistence of maximal solutions reflects the open-ended nature of the liveness component of the specification [22], [28], [29].

6. Effective solution of SCP$^\omega$. To study effective supervisor synthesis, we assume that both DES and specification languages are represented by finite automata. We suppose the DES to be modeled as a deterministic Büchi automaton [20]. The Büchi acceptance criterion is chosen for technical simplicity; more general models will be considered in future reports. The maximal legal sublanguage $E \subseteq S$ is assumed to be given by a deterministic Rabin automaton (see [24]). For the minimal acceptable sublanguage $A \subseteq E$, we need only a representation of the prefix $\mathrm{pre}(E)$, by Proposition 5.2; we assume $\mathrm{pre}(A)$ to be given by a finite automaton.

Computation of controllability prefixes. A central aspect of the effective solution of SCP$^\omega$ is the computation of the controllability prefix $\mathrm{pre}_{(L,S)}(E) \subseteq \mathrm{pre}(E)$ of the maximal legal sublanguage $E \subseteq S$. For this, it is convenient to assume that the languages $L \subseteq \Sigma^*$, $S \subseteq \Sigma^\omega$, and $E \subseteq \Sigma^\omega$ are represented by different acceptance criteria based on the same transition structure. We therefore define a *Rabin–Büchi*

automaton to be a 6-tuple

$$(\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\}, R),$$

where the first four components determine a deterministic transition structure in the usual way, the fifth is a set of pairs of state subsets as appears in a Rabin acceptance criterion (as defined in [24]), and the sixth is a state subset as employed in a Büchi acceptance criterion [26].

In particular, we assume that a deterministic Rabin–Büchi automaton $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\}, R)$ is given, such that the Rabin automaton $(\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$ accepts $E \subseteq S \subseteq \Sigma^\omega$, the $*$-automaton $(\Sigma, X, \delta, x_0, X)$ accepts $L \subseteq \Sigma^*$, and the Büchi automaton $(\Sigma, X, \delta, x_0, R)$ accepts $S \subseteq \Sigma^\omega$.[8] The language $\mathrm{pre}_{(L,S)}(E) \subseteq \mathrm{pre}(E)$ then corresponds to the following subset $C^{\mathcal{A}} \subseteq X$ of the state set of $\mathcal{A}$. Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\}, R)$ be a Rabin–Büchi automaton. Then

$$C^{\mathcal{A}} := \{x \in X : E_x \subseteq S_x, (\exists E'_x \subseteq E_x)\ [E'_x \neq \emptyset \text{ is } *\text{-controllable w.r.t. } L_x \text{ and } \omega\text{-closed w.r.t. } S_x]\},$$

where, for any $x \in X$, $E_x \subseteq \Sigma^\omega$ is the ω-language accepted by the Rabin automaton $(\Sigma, X, \delta, x, \{(R_p, I_p) : p \in P\})$, $L_x \subseteq \Sigma^*$ is the $*$-language accepted by the $*$-automaton $(\Sigma, X, \delta, x, X)$, and $S_x \subseteq \Sigma^\omega$ is the ω-language accepted by the Büchi automaton $(\Sigma, X, \delta, x, R)$.

Because $\mathcal{A}$ is deterministic we have, for any $k \in \mathrm{pre}(E)$,

$$k \in \mathrm{pre}_{(L,S)}(E) \iff \delta(k, x_0) \in C^{\mathcal{A}}.$$

Just as $\mathrm{pre}_{(L,S)}(E)$ represents the set of prefixes of E whose infinite extensions "can be controlled to belong to E," under the assumption that all ω-strings generated belong to S, so $C^{\mathcal{A}}$ intuitively represents the set of states of $\mathcal{A}$ from which the automaton "can be controlled to satisfy its Rabin acceptance criterion" under the assumption that all infinite trajectories followed by $\mathcal{A}$ satisfy the Büchi acceptance criterion.

In cases where this liveness assumption is vacuous (for example, where $R = X$), the subset $C^{\mathcal{A}}$ reduces to the "controllability subset" [24] of the Rabin automaton $(\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$. It was shown in [24] that, with the use of a suitable fixpoint calculus, this subset could be represented as a certain fixpoint of an "inverse dynamics operator" based on the one-step dynamics of the controlled automaton and that this characterization allowed for efficient computation and effective control synthesis. This result can be easily extended to the present setting by means of a suitable generalization of the definition of the "inverse dynamics operator" of [24]; see [23].

Computation of supremal ω-controllable sublanguages. Once $\mathrm{pre}_{\mathbf{G}}(E)$ has been computed, it remains to find $\sup \mathcal{C}\mathcal{F}^*(\mathrm{pre}_{\mathbf{G}}(E))$, or alternatively $\sup \mathcal{C}^*(\mathrm{pre}_{\mathbf{G}}(E))$, and take the intersection of its limit with E to yield the supremal ω-controllable sublanguage $\sup \mathcal{C}^\omega(E)$ (by Proposition 5.2).

[8] Such an automaton can, of course, be constructed from separate automata accepting E, L, and S. Actually, we need only assume that the $*$-automaton accepts some $*$-language $L' \subseteq \Sigma^*$ such that $L' \cap \mathrm{pre}(E)\Sigma = L \cap \mathrm{pre}(E)\Sigma$, and the Büchi automaton accepts some ω-language $S' \subseteq \Sigma^\omega$ such that $S' \cap \mathrm{clo}(E) = S \cap \mathrm{clo}(E)$; for any such languages, $\mathrm{pre}_{(L',S')}(E) = \mathrm{pre}_{(L,S)}(E)$.

Owing to the structure of the controllability subset $C^{\mathcal{A}}$, the computation of $\sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E))$ requires only a one-step deletion of state transitions. We first define the following "state feedback" map. Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\}, R)$ be a deterministic Rabin–Büchi automaton. Define

$$\Gamma^{\mathcal{A}} : C^{\mathcal{A}} \longrightarrow \mathbf{C},$$
$$x \mapsto \bigcup\{\Gamma \in \mathbf{C} : (\forall \sigma \in \Gamma)[\delta(\sigma, x)! \Longrightarrow \delta(\sigma, x) \in C^{\mathcal{A}}]\}.$$

The map $\Gamma^{\mathcal{A}}$ can be interpreted as the most liberal state feedback control that ensures the invariance of the subset $C^{\mathcal{A}}$.

Now define a transition function corresponding to $\Gamma^{\mathcal{A}}(x)$. Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\}, R)$ be a deterministic Rabin–Büchi automaton. Define

$$\delta^{\mathcal{A}} : \Sigma \times X \to X,$$
$$(\sigma, x) \mapsto \begin{cases} \delta(\sigma, x) & \text{if } x \in C^{\mathcal{A}} \text{ and } \sigma \in \Gamma^{\mathcal{A}}(x), \\ \text{undefined} & \text{otherwise.} \end{cases}$$

We also let $\delta^{\mathcal{A}}$ denote the following natural extension of the above function:

$$\delta^{\mathcal{A}} : \Sigma^* \times X \longrightarrow X,$$
$$(1, x) \mapsto x,$$
$$(k\sigma, x) \mapsto \delta^{\mathcal{A}}(\sigma, \delta^{\mathcal{A}}(k, x)).$$

The transition function $\delta^{\mathcal{A}}$ produces the freest achievable closed-loop $*$-behavior contained in $\mathrm{pre}_{\mathbf{G}}(E)$, namely, $\sup\mathcal{CF}^*(\mathrm{pre}_{\mathbf{G}}(E))$.

PROPOSITION 6.1. *Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\}, R)$ be a deterministic Rabin–Büchi automaton. Let E be the ω-language accepted by $\mathcal{A}$, and L the $*$-language accepted by $(\Sigma, X, \delta, x_0, X)$. Then the $*$-automaton*

$$(\Sigma, X, \delta^{\mathcal{A}}, x_0, C^{\mathcal{A}})$$

accepts the $*$-language $\sup\mathcal{CF}^*(\mathrm{pre}_{(L,S)}(E))$.

Proof. Let the $*$-language accepted by the $*$-automaton be L'. We must show that

$$L' = \sup\mathcal{CF}^*(\mathrm{pre}_{(L,S)}(E)).$$

Since $\delta^{\mathcal{A}}$ is a restriction of δ, we have $L' \subseteq \mathrm{pre}_{(L,S)}(E)$.

The $*$-controllability and $*$-closure of L' follow from the definition of $\delta^{\mathcal{A}}$. This proves that $L' \in \mathcal{CF}^*(\mathrm{pre}_{(L,S)}(E))$. We now show that, for any $L'' \in \mathcal{CF}^*(\mathrm{pre}_{(L,S)}(E))$, $L'' \subseteq L'$. Specifically, we show by induction on the length of strings that, for all $k \in \Sigma^*$,

$$k \in L'' \Longrightarrow k \in L'.$$

For the base, $1 \in L'' \Longrightarrow 1 \in \mathrm{pre}_{(L,S)}(E) \iff x_0 \in C^{\mathcal{A}} \iff 1 \in L'$. For the induction step, it suffices to show that, for all $k \in L' \cap L''$,

$$\Sigma_{L'}(k) \supseteq \Sigma_{L''}(k).$$

Suppose not. Let $\sigma \in \Sigma_{L''}(k) \setminus \Sigma_{L'}(k)$. Then

$$\begin{aligned} & \sigma \notin \Gamma^{\mathcal{A}}(\delta^{\mathcal{A}}(k, x_0)) \\ \iff & \sigma \notin \Gamma^{\mathcal{A}}(\delta(k, x_0)) \\ \iff & (\forall \Gamma \in \mathbf{C})[\sigma \in \Gamma \Longrightarrow (\exists \sigma' \in \Gamma)[\delta(\sigma', \delta(k, x_0)) \notin C^{\mathcal{A}}]] \\ \iff & (\forall \Gamma \in \mathbf{C})[\sigma \in \Gamma \Longrightarrow \Gamma \cap \Sigma_L(k) \not\subseteq \Sigma_{\mathrm{pre}_{(L,S)}(E)}(k)] \\ \Longrightarrow & (\forall \Gamma \in \mathbf{C})[\Gamma \cap \Sigma_L(k) \neq \Sigma_{L''}(k)]. \end{aligned}$$

This contradicts $*$-controllability of L''. □

THEOREM 6.2. *Let $\mathcal{A} = (\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\}, R)$ be a deterministic Rabin–Büchi automaton. Let E be the ω-language accepted by the Rabin automaton $(\Sigma, X, \delta, x_0, \{(R_p, I_p) : p \in P\})$, L the $*$-language accepted by the $*$-automaton $(\Sigma, X, \delta, x_0, X)$, and $S \subseteq \lim(L)$ the ω-language accepted by the Büchi automaton $(\Sigma, X, \delta, x_0, R)$. Suppose that $E \subseteq S$. Then the deterministic Rabin automaton*

$$(\Sigma, X, \delta^{\mathcal{A}}, x_0, \{(R_p, I_p) : p \in P\})$$

accepts the ω-language $\sup \mathcal{C}^{\omega}(E)$ relative to the DES (L, S).

Proof. The proof follows by Propositions 5.2 and 6.1. □

Testing solvability. Once $\sup \mathcal{C}^{\omega}(E)$ has been computed, the existence of solutions to SCP$^{\omega}$ can be checked by testing the containment

$$\inf \mathcal{F}^{\omega}(A) \subseteq \sup \mathcal{C}^{\omega}(E).$$

Let $\mathcal{A}_{\inf} = (\Sigma, X, \delta, x_0, F)$ be a $*$-automaton accepting $\mathrm{pre}(A)$,[9] $\mathcal{A}_S = (\Sigma, X', \delta', x_0', R')$ a deterministic Büchi automaton accepting S, and $\mathcal{A}_{\sup} = (\Sigma, X''\delta'', x_0'', \{(R_p, I_p) : p \in P\})$ a total, deterministic Rabin automaton accepting $\sup \mathcal{C}^{\omega}(E)$. Then $\inf \mathcal{F}^{\omega}(A) \setminus \sup \mathcal{C}^{\omega}(E)$ is accepted by the Streett automaton[10]

$$\mathcal{A}_{\mathrm{diff}} = (\Sigma, X''', \delta''', x_0''', \{(\emptyset, X \times R \times X''), (\emptyset, F \times X' \times X'')\} \cup \{(I_p''', R_p''') : p \in P\}),$$

where

$$X''' = X \times X' \times X'',$$

$$\delta''' : (\sigma, (x, x', x'')) \mapsto \begin{cases} (\delta(\sigma, x), \delta'(\sigma, x'), \delta''(\sigma, x'')) & \text{if } \mathcal{A}_{\inf} \text{ is deterministic,} \\ \delta(\sigma, x) \times \{\delta'(\sigma, x')\} \times \{\delta''(\sigma, x'')\} & \text{otherwise;} \end{cases}$$

$$x_0''' = (x_0, x_0', x_0''),$$

$$I_p''' = X \times X' \times R_p,$$

$$R_p''' = X \times X' \times I_p.$$

(Streett automata are specified in the same way as Rabin automata, but employ the negation of the Rabin acceptance condition [21], [26]; hence this automaton accepts

[9] Such an automaton can be computed in polynomial time, given an ω-automaton accepting $A \subseteq \Sigma^{\omega}$.

[10] Because $\mathrm{pre}(A)$ is $*$-closed, we may have $F = X$. In this case, the subset pair $(F \times X' \times X'', \emptyset)$ may be omitted from $\mathcal{A}_{\mathrm{diff}}$.

$\inf\mathcal{F}^\omega(A) \cap [\Sigma^\omega \setminus \sup\mathcal{C}^\omega(E)] = \inf\mathcal{F}^\omega(A) \setminus \sup\mathcal{C}^\omega(E)$.) To check the containment, it suffices to test this automaton for emptiness [9]. The complexity of testing solvability is polynomial in the numbers of states of $\mathcal{A}_{\inf}$, $\mathcal{A}_S$, and $\mathcal{A}_{\sup}$ and linear in the number of state subset pairs of $\mathcal{A}_{\sup}$.

Supervisor synthesis. If SCP$^\omega$ is solvable, then the method outlined in the proof of Theorem 5.3 can be used to synthesize a supervisor that solves SCP$^\omega$.

Let $\mathcal{A}_{\sup} = (\Sigma, X'', \delta'', x_0'', \{(R_p, I_p) : p \in P\})$ be the deterministic Rabin automaton of the previous section. A subset $F'' \subseteq X''$ can be computed[11] in polynomial time such that the $*$-automaton $(\Sigma, X'', \delta'', x_0'', F'')$ accepts $\mathrm{pre}(\sup\mathcal{C}^\omega(E))$. Because $\mathrm{pre}(\sup\mathcal{C}^\omega(E))$ is $*$-controllable with respect to L, we may define the feedback map

$$\begin{aligned} \phi_0 &: F'' \longrightarrow \mathbf{C}, \\ x' &\mapsto \{\sigma \in \Sigma : \delta''(\sigma, x') \in F''\}. \end{aligned}$$

The map

$$\begin{aligned} f_0 &: \Sigma^* \longrightarrow \mathbf{C}, \\ k &\mapsto \phi_0(\delta''(k, x_0)) \end{aligned}$$

is a complete supervisor for (L, S) that synthesizes the $*$-language $\mathrm{pre}(\sup\mathcal{C}^\omega(E))$. It follows that f_0 synthesizes the ω-language $\lim(\mathrm{pre}(\sup\mathcal{C}^\omega(E))) \cap S = \mathrm{clo}(\sup\mathcal{C}^\omega(E)) \cap S$. Moreover, because $\mathrm{pre}(\sup\mathcal{C}^\omega(E)) = \mathrm{pre}(\mathrm{clo}(\sup\mathcal{C}^\omega(E)) \cap S)$, f_0 is deadlock-free.

By the results of [23], there exists a feedback map $\phi : X'' \longrightarrow \mathbf{C}$ such that, for any $k \in \mathrm{pre}(\sup\mathcal{C}(E))$, the supervisor

$$\begin{aligned} f_k &: \Sigma^* \longrightarrow \mathbf{C}, \\ l &\mapsto \phi(\delta''(kl, x)) \end{aligned}$$

is a complete, deadlock-free supervisor for $(L/k, S/k)$ that synthesizes some nonempty sublanguage of E/k.

It follows by the proof of Theorem 5.3 that the supervisor

$$\begin{aligned} f &: \Sigma^* \longrightarrow \mathbf{C}, \\ l &\mapsto \begin{cases} f_0(l) & \text{if } l \in \mathrm{pre}(A), \\ f_k(l/k) & \text{if } k \text{ is the shortest element of } \mathrm{pre}(l) \setminus \mathrm{pre}(A) \end{cases} \end{aligned}$$

solves SCP$^\omega$. If the $*$-automaton $\mathcal{A}_{\inf}$ of the previous section is deterministic and total, then f can be defined in terms of a state feedback map based on the transition structure of the automaton

$$\mathcal{A}_{\mathrm{diff}} = (\Sigma, X''', \delta''', x_0''', \{(\emptyset, X \times R \times X''), (\emptyset, F \times X' \times X'')\} \cup \{(I_p''', R_p''') : p \in P\})$$

constructed in the previous section, namely, $f : k \mapsto \psi(\delta'''(k, x_0'''))$, where

$$\begin{aligned} \psi &: X''' \longrightarrow \mathbf{C}, \\ (x, x', x'') &\mapsto \begin{cases} \phi_0(x'') & \text{if } x \in F, \\ \phi(x'') & \text{otherwise.} \end{cases} \end{aligned}$$

[11] Using algorithms for testing emptiness of automata on infinite strings [9].

7. Conclusion. This report extends some of the basic results of the Ramadge–Wonham supervisory control theory to an infinite-string framework. The generalization allows the consideration of liveness as well as safety properties in the formulation of control problems for discrete-event systems. The article introduces the language property of *ω-controllability*, which provides a more precise characterization of the limitations of available control actions than does the earlier, finitary notion of controllability upon which it is based. A central result is that every sublanguage of the language generated by a given DES contains a unique maximal ω-controllable sublanguage (provided that the family of *control patterns* is closed under union). Within the present framework, this supremal ω-controllable sublanguage provides the key to solvability of problems of DES supervision.

Effective supervisor synthesis is studied under the assumption that the DES is modeled as a deterministic Büchi automaton and legal behavior specified by a deterministic Rabin automaton. The computation of the supremal ω-controllable sublanguage is exponential in the size of the Rabin acceptance condition but polynomial in the size of the automaton state sets; the further step of testing for the existence of solutions is polynomial-time.

The ω-language formulation of the article is essentially due to Ramadge. However, Ramadge [19], [20] and Golaszewski and Ramadge [11] consider only specifications that are ω-closed relative to DES behavior, and therefore represent pure safety properties; thus the infinitary setting serves only to provide a model of nonterminating behavior rather than to allow liveness specifications and does not necessitate the infinitary controllability property employed here [22]. The study of Kumar, Garg, and Marcus is similar in this respect [12], [13]. Young, Spanjol, and Garg [28], [29] consider liveness specifications represented by deterministic Büchi automata and introduce the language property of *finite stabilizability*; the conjunction of finite stabilizability and $*$-controllability is within this context equivalent to ω-controllability [22].

The main contribution of the present article is the extension of supervisory control theory to allow the use of liveness properties in the specification of DESs. While the importance of such properties is widely recognized in computer science, it has yet to be thoroughly evaluated from the perspective of control. Liveness properties can be expected to lead to simpler, more modular specifications, owing to their relatively weak, open-ended nature. For the same reason, they allow qualitatively acceptable behavior to be specified as liberally as possible, permitting the existence and form of solutions to be studied before quantitative performance criteria are introduced [14]. This article provides a framework for the study of such potential benefits.

Acknowledgments. The reading of a preliminary version of these results by the first author's thesis examiners, particularly Professors Eric Hehner, Raymond Kwong, and Amir Pnueli, is gratefully acknowledged. Thanks are due also to the referees for helpful comments on the presentation of the material.

REFERENCES

[1] B. Alpern and F. B. Schneider, *Defining liveness*, Inform. Process. Lett., 21 (1985), pp. 181–185.

[2] A. Arnold and M. Nivat, *Controlling behaviours of systems: Some basic concepts and some applications*, in Mathematical Foundations of Computer Science 1980 (Lecture Notes in Computer Science, No. 88), 1980, Springer-Verlag, New York, pp. 113–122.

[3] D. L. Black, *On the existence of delay-insensitive fair arbiters: Trace theory and its limitations*, Distributed Comput., 1 (1986), pp. 205–225.

[4] L. Boasson and M. Nivat, *Adherences of languages*, J. Comput. System Sci., 20 (1980), pp. 285–309.

[5] D. L. DILL, *Trace Theory for Automatic Hierarchical Verification of Speed-Independent Circuits*, ACM Distinguished Dissertations, MIT Press, Cambridge, MA, 1989.
[6] S. EILENBERG, *Automata, Languages and Machines*, Vol. A, Academic Press, New York, 1974.
[7] C. C. ELGOT, *Decision problems of finite automata and related arithmetics*, Trans. Amer. Math. Soc., 98 (1961), pp. 21–51.
[8] E. A. EMERSON, *The role of Büchi's automata in computer science*, in The Collected Works of J. Richard Büchi, S. MacLane and D. Siefkes, eds., Springer-Verlag, Berlin, New York, 1990, pp. 18–22.
[9] E. A. EMERSON AND C.-L. LEI, *Modalities for model checking: Branching time strikes back*, in Proc. 12th ACM Sympos. on Principles of Programming Languages, New Orleans, LA, 1985, pp. 84–96.
[10] C. GOLASZEWSKI AND P. RAMADGE, *Control of discrete event processes with forced events*, in Proc. 26th IEEE Conference on Decision and Control, Los Angeles, CA, Dec. 9–11, 1987, pp. 247–251.
[11] ———, *Mutual exclusion problems for discrete event systems with shared events*, in Proc. 27th IEEE Conference on Decision and Control, Austin, TX, Dec. 1988, pp. 234–239.
[12] R. KUMAR, V. GARG, AND S. I. MARCUS, *On ω-controllability and ω-normality of deds*, in Proc. 1991 American Control Conference, Boston, MA, June 26–28, 1991, pp. 2905–2910.
[13] ———, *On supervisory control of sequential behaviors*, IEEE Trans. Automat. Control, 37 (1992), pp. 1978–1985.
[14] L. LAMPORT, *Proving the correctness of multiprocess programs*, ACM Trans. Software Engrg., SE-3 (1977), pp. 125–143.
[15] D. E. MULLER, *Infinite sequences and finite machines*, in Proc. 4th Annual Sympos. on Switching Circuit Theory and Logical Design, Chicago, IL, IEEE, New York, Oct. 28–30, 1963, pp. 3–16.
[16] P. RAMADGE AND W. WONHAM, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.
[17] ———, *The control of discrete event systems*, Proc. IEEE, 77 (1989), pp. 81–98.
[18] P. J. RAMADGE, *Supervisory control of discrete event systems: A survey and some new results*, in Discrete Event Systems: Models and Applications, IIASA Conference, Sopron, Hungary, Aug. 3–7, 1987 (Lecture Notes in Control and Information Sciences, Vol. 103), P. Varaiya and A. Kurzhanski, eds., 1988, Springer-Verlag, New York, pp. 69–80.
[19] ———, *Tractable supervisory control problems for discrete-event systems*, in Analysis and Control of Nonlinear Systems, C. Byrnes, C. Martin, and R. Saeks, eds., 1988, North-Holland, New York, pp. 359–368.
[20] P. J. RAMADGE, *Some tractable supervisory control problems for discrete-event systems modeled by Büchi automata*, IEEE Trans. Automat. Control, 34 (1989), pp. 10–19.
[21] R. STREETT, *A Propositional Dynamic Logic of Looping and Converse*, Tech. Rep. TR-263, MIT Laboratory for Computer Science, Cambridge, MA, 1981.
[22] J. THISTLE AND W. WONHAM, *On the synthesis of supervisors subject to ω-language specifications*, in Proc. 1988 Conference on Information Sciences and Systems, Princeton University, Princeton, NJ, March 1988, pp. 440–444.
[23] ———, *Control of ω-automata, Church's problem, and the emptiness problem for tree ω-automata*, in Computer Science Logic: 5th Workshop, CSL '91, Berne, Switzerland, October 1991, Proceedings (Lecture Notes in Computer Science, Vol. 626), E. Börger, G. Jäger, H. K. Büning, and M. Richter, eds., Springer-Verlag, Berlin, Heidelberg, 1992, pp. 367–381.
[24] ———, *Control of infinite behaviour of finite automata*, SIAM J. Control Optim., 32 (1994), pp. 1075–1097, this issue.
[25] J. G. THISTLE, *Control of Infinite Behaviour of Discrete-Event Systems*, Ph.D. thesis, University of Toronto, Toronto, Canada, Jan. 1991; Systems Control Group Report No. 9012, Systems Control Group, Department of Electrical Engineering, University of Toronto, January 1991.
[26] W. THOMAS, *Automata on infinite objects*, in Handbook of Theoretical Computer Science, Vol. B: Formal Models and Semantics, J. van Leeuwen, ed., Elsevier, The MIT Press, Cambrige, MA, 1990, pp. 134–191.
[27] H. WONG-TOI AND G. HOFFMANN, *The control of dense real-time discrete-event systems*, preprint, 1992.
[28] S. YOUNG, D. SPANJOL, AND V. GARG, *Control of Discrete Event Systems Modeled with Infinite Strings*, Tech. Rep., University of Texas at Austin, Austin, TX, 1990.
[29] ———, *Control of discrete event systems modeled with deterministic Büchi automata*, in Proc. 1992 American Control Conference, Chicago, IL, June 24–26, 1992, pp. 2809–2813.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1114–1127, July 1994

013

A VERSION OF OLECH'S LEMMA IN A PROBLEM OF THE CALCULUS OF VARIATIONS*

ARRIGO CELLINA† AND SANDRO ZAGATTI‡

Abstract. This paper studies the solutions of the minimum problem for a functional of the gradient under linear boundary conditions. A necessary and sufficient condition, based on the facial structure of the epigraph of the integrand, is provided for the continuous dependence of the solutions on boundary data.

Key words. calculus of variations, extremality, strong convergence, weak convergence

AMS subject classification. 49A50

1. Introduction. A well-known result in the framework of integrals of multifunctions, Olech's lemma, gives a condition implying strong convergence out of a very weak form of convergence to extreme points [O1]. Namely, if e is an extreme point of the closure of the integral of a multifunction, there is a unique integrand in the multifunction that gives e; moreover, if u and v are arbitrary selections and $\int u$ and $\int v$ are sufficiently close to e, then u and v are close to each other in L^1. Hence this result exhibits a condition (extremality) that implies both uniqueness and continuous dependence that has been investigated, in the context of n-dimensional integration, by various authors (see [A], [AR], [Re], [Ba], [V] and the references quoted there); the purpose of this note is to investigate a similar property in the context of the calculus of variations. More precisely, we consider the problem (studied in the context of crystallography) of minimizing a functional of the gradient under linear boundary conditions:

$$\mathcal{P}_a: \qquad \text{Minimize } \int_\Omega g(\nabla u(x))dx; \qquad u \in \langle a, \cdot \rangle + W_0^{1,1}(\Omega); \quad (\Omega \subset \mathbb{R}^n)$$

and study the dependence on $a \in \mathbb{R}^n$ of the solutions to $\mathcal{P}_a$.

Analogously to the case of Olech's lemma, which infers strong convergence of the selections from the convergence of their integrals to the extreme points of the integral of the multifunction, here we have a vector parameter a playing the role of the integral, in the sense that the location of $(a, g^{**}(a))$ with respect to the facial structure of the epigraph of g^{**} (the bipolar of g) determines whether continuous dependence of the solutions of $\mathcal{P}_a$ with respect to boundary data holds.

As shown in [C1] and in [C2], uniqueness for problem $\mathcal{P}_a$ holds if and only if the dimension d of the face of the epigraph of g^{**} to whose relative interior $(a, g^{**}(a))$ belongs is strictly less than n, the dimension of the space, and in this case the solution is $u_a = \langle a, \cdot \rangle$. Hence we might ask the following question: given a point a such that the previous uniqueness condition holds, is it true that whenever a point a' is sufficiently close to a, solutions of $\mathcal{P}_{a'}$ are close to u_a in $W^{1,1}$? This is certainly true in a special

*Received by the editors July 20, 1992; accepted for publication (in revised form) January 22, 1993.

†Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Beirut 2–4, I34014 Trieste, Italy.

‡Dipartimento di Matematica, Universitá di Bologna, Piazza di Porta S. Donato 5, I40127 Bologna, Italy.

case: assume indeed that, given a, there exists a neighbourhood U of a such that for any point a' in U, $\mathcal{P}_{a'}$ has the unique solution $u_{a'}$; in this case continuous dependence follows from the explicit form of the solutions.

Hence the problem arises whenever the point a is such that $\mathcal{P}_a$ admits the unique solution u_a and there are points a_k, arbitrarily close to a, for which the corresponding problem $\mathcal{P}_{a_k}$ has infinitely many solutions. This happens when $(a_k, g^{**}(a_k))$ belongs to an n-dimensional face F of $\mathrm{epi}(g^{**})$ and $(a, g^{**}(a))$ belongs to a face F_1 of dimension less than n contained in the relative boundary of F. It is to this case we will refer in our main result, according to which the following conditions are equivalent.

(i) All the solutions u^k of $\mathcal{P}_{a_k}$ are close to u_a in $W^{1,1}$ whenever a_k is close to a;

(ii) $(a, g^{**}(a))$ is an extreme point of the epigraph of g^{**}.

As such, our result is the exact *replica* to Olech's lemma, but it is not true, in general, that uniqueness always implies continuous dependence. Indeed uniqueness holds whenever the dimension d of the face F_1 is in $\{0, 1, \dots, n-1\}$ while for $d = 1, \dots, n-1$, continuous dependence does not hold. Hence our result provides a characterization of extreme points in the sense that whenever $\mathcal{P}_{a_k}$ admits solutions different from the affine one (i.e., when $(a_k, g^{**}(a_k))$ belongs to an n-dimensional face) and $a_k \to a$, then a sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}_{a_k}$ converges strongly to u_a if and only if $(a, g^{**}(a))$ is extremal. Moreover, our result provides a precise definition of the type of convergence (partially weak, partially strong) that occurs for $1 \le d \le n-1$.

The previous analysis applies in particular to the special case of a rotationally symmetric function g. In Remark 4.2 we present a detailed description of this case.

2. Preliminaries and notation. In this paper we study the solutions of the following problems:

$$\mathcal{P}_a: \qquad \text{Minimize } \int_\Omega g(\nabla u(x))dx; \qquad u \in u_a + W_0^{1,1}(\Omega);$$

$$\mathcal{P}_a^{**}: \qquad \text{Minimize } \int_\Omega g^{**}(\nabla u(x))dx; \qquad u \in u_a + W_0^{1,1}(\Omega),$$

where g is a lower semicontinuous (l.s.c.), not necessarily convex, function defined on $\mathbb{R}^n$ with values in $\mathbb{R}$ bounded from below and g^{**} is its bipolar (see [ET] for a definition). Ω is an open, bounded subset of $\mathbb{R}^n$ with piecewise C^1 boundary and $u_a \equiv \langle a, x\rangle$ ($a \in \mathbb{R}^n$). By $u_a + W_0^{1,1}(\Omega)$ we mean the set of functions u that can be written as $u = u_a + v$, where $v \in W_0^{1,1}(\Omega)$. Here and in the following, $\langle\cdot,\cdot\rangle$ denotes the scalar product in $\mathbb{R}^n$ and $|\cdot|$ the associated norm. A point in $\mathbb{R}^n \times \mathbb{R}$ is denoted as a pair (x, z) with $x \in \mathbb{R}^n$ and $z \in \mathbb{R}$. We use the spaces $L^1(\Omega)$ and $W_0^{1,1}(\Omega)$ endowed with the usual norm $\|\cdot\|_{L^1(\Omega)}$, $\|u\|_{W_0^{1,1}(\Omega)} = \|\nabla u\|_{L^1(\Omega)}$. The weak convergence in such spaces is denoted with the half arrow $\rightharpoonup$.

For S subset of $\mathbb{R}^n$ and $x \in \mathbb{R}^n$, $\mathrm{dist}(x, S)$ is the distance of x from S, S^c is the complement, $\mathrm{co}(S)$ is the convex hull and $\mu(\cdot)$ is the Lebesgue measure. When zero belongs to S, the smallest linear manifold containing S is denoted by $\mathrm{span}(S)$; the dimension of an affine set is the dimension of the subspace parallel to it, and we say that a subset S of $\mathbb{R}^n$ has dimension p if the dimension of the affine hull of S is p, and write $\dim(S) = p$. For a scalar function f we define the negative and the positive parts $f^- \equiv \max(-f, 0)$ and $f^+ \equiv \max(f, 0)$.

We make use in this paper of basic elements of convex analysis such as the notions of face, extreme point of a convex set, relative boundary (r.b.), relative interior (r.i.) and polytope, following the notations contained in [R]; we call $\mathrm{extr}(C)$ the set of extreme points of a convex set C.

Given a subset S of $\mathbb{R}^n \times \mathbb{R}$ we denote by $\hat{S}$ the projection of S on $\mathbb{R}^n$, i.e.,

$$\hat{S} = \{x \in \mathbb{R}^n : \exists z \in \mathbb{R} : \quad (x,z) \in S\}.$$

The study of problems $\mathcal{P}_a$ and $\mathcal{P}_a^{**}$ involves the properties of the epigraph of g^{**}, $\mathrm{epi}(g^{**})$, which is a convex subset of $\mathbb{R}^n \times \mathbb{R}$; we recall now some properties of the epigraph of a convex function (see [C1], [C2]).

PROPOSITION 2.1. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex l.s.c. function. Then we have the following*:

(i) *The collection of the relative interior of the faces of* $\mathrm{epi}(h)$ *is a partition of* $\mathrm{epi}(h)$.

(ii) *If F is a face of* $\mathrm{epi}(h)$ *containing a point $(x, h(x))$ in its relative interior, F is a proper face and* $\dim(F) \leq n$; *moreover* $\dim(F) = \dim(\hat{F})$.

(iii) *If F_1 is a proper face of a proper face F of* $\mathrm{epi}(h)$ *and* r.i. (F_1) *contains a point $(x, h(x))$, then $\hat{F}_1$ is a proper face of $\hat{F}$. Moreover a point $(x, h(x))$ is an extreme point of* $\mathrm{epi}(h)$ *if and only if x is an extreme point of all the projections of the faces that contain $(x, h(x))$.*

Proof. Statement (i) is a particular case of [R, Thm. 18.2]. To prove (ii) we simply note that $(x, h(x))$ cannot belong to the relative interior of $\mathrm{epi}(h)$; then $\dim(F) \leq n$. Moreover F cannot contain a point (x, z) with $z > h(x)$; hence $\dim(F) = \dim(\hat{F})$. Statement (iii) is trivial. □

We will need the following characterization of faces of a convex set (see [O2]).

LEMMA 2.1. *Let F be a convex subset of $\mathbb{R}^n$ and F_d a d-dimensional face of F such that zero belongs to* r.i.(F_d). *Then there exist $n - d$ orthonormal vectors $h_1, \ldots, h_{n-d}$ such that F is contained in the cone*

$$\begin{aligned} C := &\{x : \langle h_1, x\rangle > 0\} \bigcup \{x : \langle h_1, x\rangle = 0,\ \langle h_2, x\rangle > 0\} \\ &\bigcup \{x : \langle h_1, x\rangle = \langle h_2, x\rangle = \cdots = \langle h_{n-d-1}, x\rangle = 0,\ \langle h_{n-d}, x\rangle > 0\} \bigcup \\ &\bigcup \{x : \langle h_1, x\rangle = \langle h_2, x\rangle = \cdots = \langle h_{n-d}, x\rangle = 0\} \end{aligned}$$

and

$$F_d = F \bigcap \{x : \langle h_1, x\rangle = \langle h_2, x\rangle = \cdots = \langle h_{n-d}, x\rangle = 0\}.$$

We remind the reader now of the well-known criteria of weak convergence in $L^1(\Omega)$ and $W_0^{1,1}(\Omega)$ (see [D, p. 19]).

THEOREM 2.1. *Let Ω be a bounded open subset of $\mathbb{R}^n$, and $\{f_k\}_{k\in\mathbb{N}}$ be a sequence in $L^1(\Omega)$; then*

$$f_k \rightharpoonup f \quad \text{in } L^1(\Omega)$$

if and only if

(i) $\|f_k\|_{L^1(\Omega)} \leq M$,

(ii) f_k *is absolutely equiintegrable,*

(iii) $\lim_{k\to\infty} \int_D [f_k(x) - f(x)]dx = 0$ *for any cube $D \subset \Omega$.*

THEOREM 2.2. *Let Ω be a bounded open subset of $\mathbb{R}^n$, and $\{f_k\}_{k\in\mathbb{N}}$ be a sequence in $W_0^{1,1}(\Omega)$; then*

$$f_k \rightharpoonup f \quad \text{in } W_0^{1,1}(\Omega)$$

if and only if

$$D_i f_k \rightharpoonup D_i f \quad \text{in } L^1(\Omega)$$

for $i = 1, \ldots, n$.

(See [B, p. 175].) We end this section with the following definition.

DEFINITION 2.1. *We say that a function* $g : \mathbb{R}^n \to \mathbb{R}$ *satisfies the growth condition* (C) *if there exists a function* $\phi : \mathbb{R} \to \mathbb{R}$ *such that* $\lim_{t\to+\infty} \frac{\phi(t)}{t} = +\infty$ *and* $g(y) \geq \phi(|y|)$ *for any* $y \in \mathbb{R}^n$.

3. Existence and uniqueness. In [C1] and [C2] Cellina gives sufficient and necessary conditions on the affine boundary datum u_a for the existence and the uniqueness of solutions of $\mathcal{P}_a$ and $\mathcal{P}_a^{**}$ investigating the facial structure of the epigraph of g^{**}. The main results stated in the quoted papers can be summarized as follows. We emphasize that a solution of $\mathcal{P}_a$ is a solution of $\mathcal{P}_a^{**}$ as well.

THEOREM 3.1. *Let* $g : \mathbb{R}^n \to \mathbb{R}$ *be l.s.c. (not necessarily convex), bounded from below, satisfying growth condition* (C); *let* Ω *be a bounded open subset of* $\mathbb{R}^n$ *with piecewise* C^1 *boundary.*

(i) *If* $\mathcal{P}_a$ *admits a solution, then*

(1) *either* $g(a) = g^{**}(a)$ *or the face of* $\mathrm{epi}(g^{**})$ *to whose relative interior* $(a, g^{**}(a))$ *belongs has dimension* n.

(ii) *Conversely, if condition* (1) *holds then* $\mathcal{P}_a$ *admits at least one solution.*

THEOREM 3.2. *Assume the hypotheses of Theorem* 3.1. *Then*

(i) $\mathcal{P}_a^{**}$ *admits the unique solution* u_a *if and only if* $(a, g^{**}(a))$ *belongs to the relative interior of a face of* $\mathrm{epi}(g^{**})$ *of dimension strictly less than* n.

(ii) $\mathcal{P}_a$ *admits the unique solution* u_a *if and only if* $g^{**}(a) = g(a)$ *and* $(a, g^{**}(a))$ *belongs to the relative interior of a face of* $\mathrm{epi}(g^{**})$ *of dimension strictly less than* n.

The proof of the second part of Theorem 3.1 consists essentially of the explicit construction of the solution of $\mathcal{P}_a$ in the case in which $(a, g^{**}(a))$ belongs to the relative interior of an n-dimensional face of $\mathrm{epi}(g^{**})$. Since we need this construction in the proof of our main result, we recall it in its main steps and refer to [C2] for details.

We begin with a lemma.

LEMMA 3.1. *Let* $\{y_i,\ i = 1, \ldots, m\}$ *be a set of vectors in* $\mathbb{R}^n$, *and consider* $S = \mathrm{co}\{y_i,\ i = 1, \ldots, m\}$. *Suppose* $\dim(S) = n$, $0 \in \mathrm{int}(S)$ *and call* S^* *the polar set of* S. *Then there exists a finite partition* $\{S_i^*,\ i = 1, \ldots, m\}$ *of* S^* *and a Lipschitz continuous function* w, *defined on* $\mathbb{R}^n$, *such that*

(i) $w{=}0$ *on* $(S^*)^c$;

(ii) $\nabla w = y_i$ *almost everywhere in* S_i^*, $i = 1, \ldots, m$;

(iii) *there exists an index set* I *contained in* $\{1, \ldots, m\}$ *such that the set* $\{y_i, i \in I\}$ *contains a system of* n *linearly independent vectors and* $\mu(S_i^*) > 0$ *for* $i \in I$.

Proof. The proof of (i) and of (ii) can be found in [C2]. To prove statement (iii) we recall that since zero belongs to the interior of S, $m > n$, the polar S^* is bounded and it can be written as

$$S^* = \bigcap_{i=1}^{m} \{x : \langle y,x\rangle \leq 1\}.$$

We also recall that the sets S_i^* are defined by

$$S_i^* := co\{F_i^*, 0\},$$

where $F_i^* = S^* \cap \{x : \langle y_i, x\rangle = 1\}$. Since $\mathrm{dist}(0, F_i^*) > 0$ for any index i, $\mu(S_i^*) > 0$ if and only if $\dim(F_i^*) = n - 1$. S^* has at least $n + 1$ faces of dimension $n - 1$; hence we may assume, renaming the indices, that there exists $p \geq n + 1$ such that $\dim(F_i^*) = n - 1$ for $i = 1, \ldots, p$ and $\dim(F_i^*) < n - 1$ for $i = p + 1, \ldots, m$. A face F_j^*

with $j > p$ is a proper face of a face F_i^* with $i < p$, hence $S_j^* \subset S_i^*$, and we can write

$$S^* = \bigcap_{i=1}^{p} \{x : \langle y,x\rangle \leq 1\}.$$

Then the set $\{y_i, i = 1, ..., p\}$ contains a system of n linearly independent vectors since otherwise S^* would be unbounded. □

Proof of Theorem 3.1 (ii). Assume that $(a, g^{**}(a))$ belongs to r.i.(F), the relative interior of F, where F is an n-dimensional face of epi(g^{**}). Our goal is to construct a solution of $\mathcal{P}_a$ (different from u_a). Since g satisfies the growth condition (C), F is bounded and is contained in a hyperplane H separating it from epi(g^{**}). According to Proposition 2.1, H cannot be vertical, i.e.,

$$H = \{(x,z) \in \mathbb{R}^n \times \mathbb{R} : z = \langle h, x\rangle + k\} \qquad (h \in \mathbb{R}^n, k \in \mathbb{R})$$

and, since the extreme points of F are of the form $(y, g(y))$,

$$\operatorname{extr}(\hat{F}) = \{y \in \mathbb{R}^n : \ (y, g(y)) \in \operatorname{extr}(F)\}.$$

Consider a subset $\{y_i,\ i = 1, \ldots, m\}$ of extr$(\hat{F})$ such that $\dim(\operatorname{co}\{y_i,\ i = 1, \ldots, m\}) = n$ and $a \in$ r.i.$(\operatorname{co}\{y_i,\ i = 1, \ldots, m\})$; we remark that whenever

$$a = \sum_{i=1}^{m} \lambda_i y_i, \quad 0 < \lambda_i < 1, \quad \sum_{i=1}^{m} \lambda_i = 1, \tag{3.1}$$

it is

$$g^{**}(a) = \langle h, a\rangle + k = \sum_{i=1}^{m} \lambda_i(\langle h, y_i\rangle + k) = \sum_{i=1}^{m} \lambda_i g^{**}(y_i) = \sum_{i=1}^{m} \lambda_i g(y_i); \tag{3.2}$$

we define the polytope $S(a) := \operatorname{co}\{y_i - a,\ i = 1, ..., m\}$. We can apply Lemma 3.1, defining a partition $\{S_i^*(a),\ i = 1, ..., m\}$ of $S^*(a)$ and a Lipschitz function w^a such that $w^a = 0$ on $(S^*(a))^c$ and $\nabla w^a = y_i - a$ almost everywhere on $S_i^*(a)$.

We now consider the collection of subsets of Ω

$$\mathcal{U} = \{z + rS^*(a),\ z \in \Omega,\ r \in \mathbb{R},\ r < \operatorname{dist}(z, \Omega^c)\}.$$

$\mathcal{U}$ is a Vitali covering of Ω, and we can select a countable subcovering $\{\Omega_j(a),\ j \in \mathbb{N}\}$ such that

(1) $\Omega_j(a) = z_j + r_j S^*(a) \subset \Omega$ for all $j \in \mathbb{N}$;
(2) $\Omega_j(a) \bigcap \Omega_k(a) = \emptyset$, if $j \neq k$;
(3) $\Omega = N \bigcup (\bigcup_{j=1}^{\infty} \Omega_j(a))$ where $\mu(N) = 0$;
(4) $\Omega_j(a) = \bigcup_{i=1}^{m} \Omega_j^i(a)$ (disjoint union); where $\Omega_j^i(a) = z_j + r_j S_i^*(a)$. We set also $\Omega^i(a) = \bigcup_{j=1}^{\infty} \Omega_j^i(a)$, obtaining $\Omega = \bigcup_{i=1}^{m} \Omega^i(a)$, and define

$$w_j^a(x) = r_j w^a\left(\frac{x - z_j}{r_j}\right) \qquad \text{a.e. } x \in \Omega_j(a), \qquad j \in \mathbb{N}$$

and

$$v^a(x) = \sum_{j=1}^{\infty} w_j^a(x), \quad \text{a.e. } x \in \Omega.$$

v^a belongs to $W_0^{1,1}(\Omega)$ and it is

$$\nabla v^a = y_i - a \quad \text{a.e. on } \Omega^i(a), \qquad i = 1, \dots, m. \tag{3.3}$$

Then

$$0 = \int_\Omega \nabla v^a = \sum_{i=1}^m \int_{\Omega^i(a)} (y_i - a),$$

i.e.,

$$a = \sum_{i=1}^m \frac{\mu(\Omega^i(a))}{\mu(\Omega)} y_i. \tag{3.4}$$

We set $u(x) := v^a(x) + \langle a, x\rangle = v^a(x) + u_a(x)$; first, (3.3) implies

$$\nabla u = y_i \quad \text{a.e. on } \Omega^i(a), \tag{3.5}$$

and by virtue of (3.1), (3.2), and (3.4), u is a solution of $\mathcal{P}_a$. □

We are interested in the following question. Consider a sequence $\{a_k\}_{k\in\mathbb{N}}$ converging to a point a such that $\mathcal{P}_a$ admits the unique solution u_a, and a sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}_{a_k}$ $(\mathcal{P}_{a_k}^{**})$ (in the sense that for any $k \in \mathbb{N}$, u^k is a solution of $\mathcal{P}_{a_k}$ $(\mathcal{P}_{a_k}^{**})$). We ask whether $\{u^k\}_{k\in\mathbb{N}}$ converges (in some topology) to u_a as $k \to \infty$. When it happens that, for any $k \in \mathbb{N}$, $(a_k, g^{**}(a_k))$ belongs to the relative interior of a face of dimension strictly less than n, the question is trivial because $u^k \equiv u_{a_k}$ and converges to u_a strongly in $W_0^{1,1}(\Omega)$. The interesting case is when $(a, g^{**}(a))$ belongs to the relative interior of a face F_1 of $\operatorname{epi}(g^{**})$ of dimension strictly less than n, and, for an infinite number of indices $k \in \mathbb{N}$, $(a_k, g^{**}(a_k))$ belongs to the relative interior of at least one n-dimensional face of $\operatorname{epi}(g^{**})$ containing $(a, g^{**}(a))$ (and also F_1) in its relative boundary. According to our main result $(a, g^{**}(a))$ is an extreme point of $\operatorname{epi}(g^{**})$ if and only if any sequence $\{u^k\}_{k\in\mathbb{N}}$ converges strongly to u_a in $W_0^{1,1}(\Omega)$.

4. Main result. We will need the following technical lemmas.

LEMMA 4.1. *Let Ω be an open bounded subset of $\mathbb{R}^n$ and $\{v^k\}_{k\in\mathbb{N}}$ be a sequence in $W_0^{1,1}(\Omega)$. Suppose that $v^k \to 0$ in $L^1(\Omega)$ and that, for some $i \in \{1, \dots, n\}$, $|D_i v^k| \le M$ almost everywhere in Ω, where M is a positive constant. Then*

$$D_i v^k \rightharpoonup 0 \quad \text{in } L^1(\Omega).$$

Proof. We can suppose $i = 1$ and $\Omega = I \times S$, where I is an open bounded interval of $\mathbb{R}$ and S is an open bounded subset of $\mathbf{R}^{n-1}$, since v^k can be extended as zero out of Ω. Let us write $v^k = v^k(x_1, x')$ with $x_1 \in I$ and $x' \in S$; the uniform boundedness of $|D_1 v^k|$ implies that the sequence $\{D_1 v^k\}_{k\in\mathbb{N}}$ is bounded in L^1-norm and is absolutely equi-integrable. According to Theorem 2.1 it is sufficient to prove that for any cube $D \subset \Omega$,

$$\lim_{k\to\infty} \int_D D_1 v^k \to 0.$$

(1) Suppose first $v^k \in C_0^1(\Omega)$ and define

$$\varphi^k(x_1) = \int_S |v^k(x_1, x')| dx', \qquad k \in \mathbb{N}, \quad x_1 \in I.$$

$\{\varphi^k\}_{k\in\mathbb{N}}$ is a sequence of nonnegative continuous functions on I differentiable almost everywhere for any k and the sequence of derivatives is uniformly bounded; hence they are equicontinuous and equibounded.

Moreover $\lim_{k\to\infty}\|\varphi^k\|_{L^1(I)} = \lim_{k\to\infty}\|v^k\|_{L^1(\Omega)} = 0$, then $\varphi^k \to 0$ uniformly on I.

Consider a cube $D \subset \Omega$, $D = (\xi,\eta)\times Q$ where $\xi,\eta \in I$ and Q is an $(n-1)$-dimensional cube contained in S. It is

$$\left|\int_D D_1 v^k(x_1,x')dx_1dx'\right| = \left|\int_Q\left(\int_\xi^\eta D_1 v^k(x_1,x')dx_1\right)dx'\right| \le 2\sup_{x_1\in I}\left(\int_S |v^k(x_1,x')|dx'\right).$$

Hence $D_1 v^k \to 0$ in $L^1(\Omega)$.

(2) Consider now the general case $v^k \in W_0^{1,1}(\Omega)$.

By density there exists a sequence $w^k \in C_0^1(\Omega)$ such that $|D_1 w^k|$ is uniformly bounded in Ω and

$$\|v^k - w^k\|_{W_0^{1,1}} \le \frac{1}{k} \quad \forall k \in \mathbb{N}.$$

Obviously $w^k \to 0$ in $L^1(\Omega)$ and the previous arguments show that $D_1 w^k \to 0$ in $L^1(\Omega)$; hence $D_1 v^k \to 0$ in $W_0^{1,1}(\Omega)$. □

LEMMA 4.2. *Let P be a polytope in $\mathbb{R}^n$ and F be a proper face of P. Then F is exposed, i.e., there exists a supporting hyperplane π of P such that $F = P\cap\pi$.*

Proof. We can assume $0 \in F$ as well.

Set $P = \mathrm{co}\{v_1,\dots,v_m\}$ and $V = \max\{|v_1|,\dots,|v_m|\}$. Consider the collection of all nontrivial hyperplanes H_α separating F from P. Let ν_α be the number of vectors $v_1,\dots,v_m$ contained in H_α but not belonging to F and call ν_0 the minimum, attained for some hyperplane H_0 defined by $H_0 = \{x : \langle h_0, x\rangle = 0\}$. We wish to show that $\nu_0 = 0$. Assume, by contradiction, that it is positive. Set P_0 to be $P\cap H_0$. Note that there is $\eta > 0$ such that for every v_i in P but not in P_0, $\langle h_0, v_i\rangle \ge \eta$. Also, $F\cap H_0$ is a proper face of $P\cap H_0$, so that there is a unit vector k in H_0 separating $F\cap H_0$ from $P\cap H_0$, i.e., $\langle k,x\rangle = 0$ for $x \in F\cap H_0$ and for some y in $P\cap H_0$, $\langle k,y\rangle > 0$. Since $y \in \mathrm{co}\{v_1,\dots,v_m\}$, there is a v_j in $P\cap H_0$ such that $\langle k, v_j\rangle > 0$. Consider $h_1 = h_0 + (\eta/2V)k$. We have that $\langle h_1, x\rangle = 0$ for $x \in F$, that for v_i in P but not in $P\cap H_0$, $\langle h_1, v_i\rangle \ge \eta - (\eta/2V)|v_i| \ge \eta/2$. and that $\langle h_1, v_j\rangle > 0$, contradicting the definition of ν_0. □

The following is our main result; it is convenient to introduce the following definition.

DEFINITION 4.1. *Let $\{v^k\}_{k\in\mathbb{N}}$ be a sequence in $W^{1,1}(\Omega)$ and $v \in W^{1,1}(\Omega)$, where Ω is an open bounded subset of $\mathbb{R}^n$. Let d be the largest integer such that there exists a d-dimensional subspace L of $\mathbb{R}^n$ such that, given any vector e in L, it is*

$$\langle \nabla v^k - \nabla v, e\rangle \overset{k\to\infty}{\longrightarrow} 0 \quad \text{in } L^1(\Omega).$$

We say that $\{v^k\}_{k\in\mathbb{N}}$ converges *d-strongly to v in $W^{1,1}(\Omega)$.*

Remark 4.1. To prove that a sequence converges d-strongly, it is sufficient to find a system E of d independent vectors such that the condition expressed in Definition 4.1 holds and that for any vector e in the orthogonal complement of E, $\langle \nabla v^k - \nabla v, e\rangle$ does not converge to zero in $L^1(\Omega)$. We should also note that $\{v^k\}_{k\in\mathbb{N}}$ is a d-strongly converging sequence in $W^{1,1}(\Omega)$ if and only if there exists a nonsingular change of

coordinates U such that, setting $w^k(x) = v^k(Ux)$, $D_j w^k$ converges strongly in $L^1(U\Omega)$ for $j = 1, \dots, d$ and does not converge in $L^1(U\Omega)$ for $j = d+1, \dots, n$. Obviously d-strong convergence in $W_0^{1,1}(\Omega)$ implies strong convergence in $L^1(\Omega)$ for any $d \geq 1$ (Poincaré inequality, see [B, p. 174]) and it is equivalent to strong convergence in $W_0^{1,1}(\Omega)$ when $d = n$.

THEOREM 4.1. *Let g be l.s.c. satisfying the growth condition* (C), *and let Ω be a bounded open subset of $\mathbb{R}^n$ with piecewise C^1 boundary. Suppose that $g(a) = g^{**}(a)$ and that $(a, g^{**}(a))$ belongs to the relative interior of a proper face F_1 of an n-dimensional face F of* $\operatorname{epi}(g^{**})$, *and let $\{a_k\}_{k\in\mathbb{N}}$ be any sequence such that $(a_k, g^{**}(a_k))$ belongs to the relative interior of F for any $k \in \mathbb{N}$ and $\lim_{k\to\infty} a_k = a$.*

(i) *If any sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}_{a_k}$ converges $(n-r)$-strongly to u_a in $W_0^{1,1}(\Omega)$, then $\dim(F_1) = r$. In particular, if any sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}_{a_k}$ converges strongly to u_a in $W_0^{1,1}(\Omega)$, then $(a, g^{**}(a))$ is an extreme point of* $\operatorname{epi}(g^{**})$.

(ii) *If $\dim(F_1) = r$ then any sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}^{**}_{a_k}$ converges $(n-r)$-strongly to u_a in $W_0^{1,1}(\Omega)$. In particular, if $(a, g^{**}(a))$ is an extreme point of* $\operatorname{epi}(g^{**})$, *then any sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}^{**}_{a_k}$ converges strongly to u_a in $W_0^{1,1}(\Omega)$.*

Proof. First, since $\dim(F_1) < n$ and $g(a) = g^{**}(a)$, $\mathcal{P}_a$, as well as $\mathcal{P}^{**}_a$, admits the unique solution u_a; moreover, growth condition (C) implies that $\hat{F}$ is bounded and we set

$$L = \sup_{y\in\hat{F}} |y|.$$

In the proof we assume, without losing generality, $a = 0$.

(i) Suppose that any sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}_{a_k}$ converges $(n-r)$-strongly to zero. We proceed by contradiction: we assume that $\dim(\hat{F}_1)$ is greater than r and show that there exists a sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}_{a_k}$ that does not converge $(n-r)$-strongly to zero in $W_0^{1,1}(\Omega)$.

(1) Set $d = r+1$ and consider $y_1, \dots, y_p \in \operatorname{extr}(\hat{F}_1)$ $(p \geq d+1)$ such that $a = 0 \in \text{r.i.}(\operatorname{co}\{y_1, \dots, y_p\})$ and $\dim(\operatorname{span}\{y_1, \dots, y_p\}) = d$. Since $a_k \in \text{r.i.}(\hat{F})$, for any k there are $n+1$ extreme points of $\hat{F}$, $v^k_{p+1}, \dots, v^k_{p+n+1}$ such that, setting $m = p+n+1$, the polytope $P_k := \operatorname{co}\{y_1, \dots, y_p, v^k_{p+1}, \dots, v^k_m\}$ has dimension n and $a_k \in \operatorname{int}(P_k)$. Setting $y^k_i = y_i - a_k$ for $i = 1, \dots, p$ and $y^k_i = v^k_i - a_k$ for $i = p+1, \dots, m$, and considering the polytope $P_k - a_k := \operatorname{co}\{y^k_i, i = 1, \dots, m\}$, we can define the (bounded) polar $S^*(a_k)$ of $P_k - a_k$ and a solution u^k of $\mathcal{P}_{a_k}$, defined as in §3, whose gradient, recalling (3.5), takes the values y_i or v^k_i on the sets $\Omega^i(a_k)$.

(2) Extracting subsequences if necessary, we may assume that y^k_i converges to a vector $y_i \in \partial\hat{F}$ for $i = p+1, \dots, m$ and $y_i \in \hat{F}_1$ for $i = p+1, \dots, s$, $y_i \in \partial\hat{F} \setminus \hat{F}_1$ for $i = s+1, \dots, m$, where $p \leq s \leq m$. We define the limit polytope $P := \operatorname{co}\{y_i, i = 1, \dots, m\}$ and its polar S^* written as

$$S^* = \bigcap_{i=1}^{m} \{x : \langle y_i, x\rangle \leq 1\}.$$

We also define

$$C^* = \bigcap_{i=1}^{s} \{x : \langle y_i, x\rangle \leq 1\}, \qquad T^* = \bigcap_{i=s+1}^{m} \{x : \langle y_i, x\rangle \leq 1\},$$

so that $S^* = C^* \bigcap T^*$, and remark that when $s = m$, $T^* = \mathbb{R}^n$ and $S^* = C^*$; in the following we consider $s < m$ since otherwise the proof proceeds in a similar and simpler way. Recalling the definition of the partition of a polar set (Lemma 3.1), we can apply the same definition to C^*, obtaining $C^* = \bigcup_{i=1}^{s} C_i^*$ and $S_i^* \subseteq C_i^*$.

Set L_d to be $\operatorname{span}\{y_i, i = 1, \dots, s\}$ and write $\mathbb{R}^n = L_d \oplus L_d^{\perp}$; the cylinder C^* can be written $C^* = S^{d*} \oplus L_d^{\perp}$, and, analogously, $C_i^* = S_i^{d*} \oplus L_d^{\perp}$; where we have set $S^{d*} := C^* \bigcap L_d$ and $S_i^{d*} := C_i^* \bigcap L_d$. We remark that S^{d*} is the polar of the d-dimensional set $\operatorname{co}\{y_1, \dots, y_s\} \cap L_d$ and $\{S_i^{d*}, i = 1, \dots, s\}$ is the relative partition; since $0 \in \text{r.i.}(\operatorname{co}\{y_i, i = 1, \dots, s\})$, S^{d*} is bounded and by point (iii) of Lemma 3.1 there exists a set I of d indices such that $\{y_i, i \in I\}$ are linearly independent and, calling μ_d the d-dimensional measure in L_d,

$$\frac{\mu_d(S_i^{d*})}{\mu_d(S^{d*})} \geq \lambda > 0 \quad \text{for } i \in I. \tag{4.1}$$

Moreover, boundedness of S^{d*} implies that there exists M ($M \geq -1$) such that

$$C^* = \bigcap_{i=1}^{s} \left[\bigcup_{\alpha_i \in [-M,1]} \{x : \langle y_i, x \rangle = \alpha_i\} \right]. \tag{4.2}$$

(3) Consider now $P_d := \operatorname{co}\{y_i, i = 1, \dots, s\}$: P_d is a face of P and, by Lemma 4.2, it is exposed. Let w be a unit vector in $L_d^{\perp}$ such that $P_d = P \bigcap \{x : \langle w, x \rangle = 0\}$ and $\langle w, x \rangle > 0$ for all $x \in P$. Let $z_{d+1}, \dots, z_n$ be orthonormal vectors in $L_d^{\perp}$ such that $w = (n-d)^{-1/2} \sum_{j=d+1}^{n} z_j$. For every $y \in \{y_{s+1}, \dots, y_m\}$, $\sum_{j=d+1}^{n} \langle z_j, y \rangle = (n-d)^{1/2} \langle w, y \rangle > 0$.

(4) Let us define the family of sets

$$Q_{[R_1,R_2]} := \bigcap_{j=d+1}^{m} \{x : \langle z_j, x \rangle \in [R_1, R_2]\} \qquad R_1, R_2 \in \mathbb{R};$$

it is our purpose to show that there exist $R_0 > 0$ and $\alpha > 1$ such that for any $R > R_0$,

$$S^* \bigcap Q_{[-\alpha R, -R]} = C^* \bigcap Q_{[-\alpha R, -R]}. \tag{4.3}$$

Since $S^* = C^* \bigcap T^*$ it is enough to prove that the right-hand side is contained in the left-hand side (when $s = m$, i.e., $S^* = C^*$, this is obvious). We show that in general $C^* \bigcap Q_{[-\alpha R, -R]}$ is contained in T^*: it follows that a point in $C^* \bigcap Q_{[-\alpha R, -R]}$ is in $C^* \bigcap T^* = S^*$, hence in $S^* \bigcap Q_{[-\alpha R, -R]}$. So, let x be any point in $C^* \bigcap Q_{[-\alpha R, -R]}$; recalling (4.2), we have

$$\langle y_i, x \rangle \in [-M, 1], \qquad i = 1, \dots, s,$$

$$\langle z_j, x \rangle \in [-\alpha R, -R], \qquad j = s+1, \dots, m.$$

Take $y \in \{y_i, i = s+1, \dots, m\}$; since $\{y_1, \dots, y_s, z_{d+1}, \dots, z_n\}$ contains a system of n independent vectors, y can be written as $y = \sum_{i=1}^{s} \nu_i y_i + \sum_{j=d+1}^{n} \mu_j z_j$, where, by point (3), $\sum_{j=d+1}^{n} \mu_j > 0$. Writing $\mu_j^+ = \max(\mu_j, 0)$ and $\mu_j^- = \max(-\mu_j, 0)$ we have, recalling (4.2),

$$\langle y, x \rangle = \sum_{i=1}^{s} \nu_i \langle y_i, x \rangle + \sum_{j=d+1}^{n} \mu_j \langle z_j, x \rangle$$

$$\leq (|M|+1)\sum_{i=1}^{s}|\nu_i| + \left(-\sum_{j=d+1}^{n}\mu_j^+ + \alpha\sum_{j=d+1}^{n}\mu_j^-\right)R.$$

By choosing α satisfying $1 < \alpha < (\sum_{j=d+1}^{n}\mu_j^+)(\sum_{j=d+1}^{n}\mu_j^-)^{-1}$, the term in parenthesis becomes negative; hence, if R is greater than some R_0 sufficiently large, it turns out that $\langle y, x\rangle \leq 1$. Repeat this choice for every $y \in \{y_i, i = s+1, \ldots, m\}$; take R_0 as the largest value and α as the smallest value so obtained, and have $x \in T^*$. By an analogous procedure we also have

$$S_i^* \bigcap Q_{[-\alpha R,-R]} = C_i^* \bigcap Q_{[-\alpha R,-R]}. \tag{4.4}$$

(5) We now take $R \geq R_0$ and consider the sets $S^* \bigcap Q_{[-\alpha R,R]}$ and $S_i^* \bigcap Q_{[-\alpha R,R]}$ for $i = 1, \ldots, s$. Such sets are bounded; by (4.3) and (4.4), we have

$$\mu\left(S^* \bigcap Q_{[-\alpha R,R]}\right) \leq \mu\left(C^* \bigcap Q_{[-\alpha R,R]}\right) = \mu_d(S^{d*})((\alpha+1)R)^{n-d},$$

and

$$\begin{aligned}\mu\left(S_i^* \bigcap Q_{[-\alpha R,R]}\right) &= \mu\left(S_i^* \bigcap Q_{[-\alpha R,-R]}\right) + \mu\left(S_i^* \bigcap Q_{[-R,R]}\right)\\ &= \mu\left(C_i^* \bigcap Q_{[-\alpha R,-R]}\right) + \mu\left(S_i^* \bigcap Q_{[-R,R]}\right)\\ &\geq \mu_d(S_i^{d*})((\alpha-1)R)^{n-d}.\end{aligned}$$

Hence, recalling (4.1),

$$\frac{\mu\left(S_i^* \bigcap Q_{[-\alpha R,R]}\right)}{\mu\left(S^* \bigcap Q_{[-\alpha R,R]}\right)} \geq \frac{\mu_d(S_i^{d*})}{\mu_d(S^{d*})}\left(\frac{\alpha-1}{\alpha+1}\right)^{n-d} \geq \gamma > 0, \qquad i \in I, \tag{4.5}$$

for some positive γ.

(6) Now consider the sets $S^*(a_k) = \bigcup_{i=1}^{m} S_i^*(a_k)$, polar of $P_k - a_k$ and their decompositions. Given $Q = Q_{[R_1,R_2]}$ the sets $S^*(a_k) \cap Q$ and $S_i^*(a_k) \cap Q$ are bounded polytopes whose vertices converge to the vertices of $S^* \cap Q$ and $S_i^* \cap Q$, respectively, since the vertices of P_k converges to the vertices of P. In particular the measures of $S^*(a_k) \cap Q$ and $S_i^*(a_k) \cap Q$ converge to the measures of $S^* \cap Q$ and $S_i^* \cap Q$. Setting

$$\gamma_i^k(R) := \frac{\mu\left(S_i^*(a_k) \bigcap Q_{[-\alpha R,R]}\right)}{\mu\left(S^*(a_k) \bigcap Q_{[-\alpha R,R]}\right)},$$

we have, by (4.5),

$$\lim_{k\to\infty} \gamma_i^k(R) = \frac{\mu\left(S_i^* \bigcap Q_{[-\alpha R,R]}\right)}{\mu\left(S^* \bigcap Q_{[-\alpha R,R]}\right)}, \qquad i \in I, \quad R \geq R_0. \tag{4.6}$$

The sets $S^*(a_k)$ are bounded for any k; hence there exists a sequence R_k in $\mathbb{R}^+$ such that $R_k \nearrow +\infty$ as $k \to \infty$ and

$$\gamma_i^k(R_k) = \frac{\mu\left(S_i^*(a_k)\right)}{\mu\left(S^*(a_k)\right)};$$

this last equality and (4.6) imply that

$$\liminf_{k\to\infty} \frac{\mu\left(S_i^*(a_k)\right)}{\mu\left(S^*(a_k)\right)} \geq \gamma, \qquad i \in I. \tag{4.7}$$

(7) Take $i \in I$. When $i \leq p$, $y_i^k = y_i - a_k$ and $\nabla u^k = y_i$ almost everywhere on $\Omega^i(a_k)$, then (4.7) implies that, for k sufficiently large,

$$\begin{aligned}\int_\Omega |\langle \nabla u^k(x), y_i\rangle|\, dx &\geq \int_{\Omega^i(a_k)} |y_i|^2\, dx = \mu(\Omega^i(a_k))\, |y_i|^2 \\ &\geq \mu(\Omega)\frac{\mu(S_i^*(a_k))}{\mu(S^*(a_k))}\, |y_i|^2 \geq \mu(\Omega)\frac{\gamma}{2}\, |y_i|^2 .\end{aligned}$$

When $i > p$, $y_i^k = v_i^k - a_k$ and $\nabla u^k = v_i^k$ almost everywhere on $\Omega^i(a_k)$; remarking that $v_i^k \to y_i$ and that $|v_i^k| \leq L$, we have, through similar computations, for k sufficiently large,

$$\int_{\Omega^i(a_k)} \left\langle \nabla u^k(x), v_i^k \right\rangle dx = \int_{\Omega^i(a_k)} |v_i^k|^2 dx \geq \frac{\gamma}{4}\mu(\Omega)|y_i|^2;$$

hence

$$\begin{aligned}\int_{\Omega^i(a_k)} \langle \nabla u^k(x), y_i\rangle\, dx &= \int_{\Omega^i(a_k)} \left\langle \nabla u^k(x), v_i^k \right\rangle dx + \int_{\Omega^i(a_k)} \left\langle \nabla u^k(x), y_i - v_i^k \right\rangle dx \\ &\geq \frac{\gamma}{4}\mu(\Omega)|y_i|^2 - L\mu(\Omega)|y_i - v_i^k| \geq \frac{\gamma}{8}\mu(\Omega)|y_i|^2.\end{aligned}$$

We have shown that

$$\int_\Omega |\langle \nabla u^k(x), y_i\rangle|\, dx \geq \int_{\Omega^i(a_k)} \langle \nabla u^k(x), y_i\rangle\, dx \geq \frac{\gamma}{8}\mu(\Omega)|y_i|^2;$$

hence for any y_i, $i \in I$, $\langle \nabla u^k, y_i\rangle$ does not converge in $L^1(\Omega)$. Since $\{y_i, i \in I\}$ is a system of d linearly independent vectors, u^k cannot converge $n-d+1 = n-r$ strongly to zero in $W_0^{1,1}(\Omega)$ and part (i) of the theorem is proved.

(ii) Suppose now $\dim(F_1) = r$ and consider a sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}_{a_k}^{**}$. We wish to show that there exist $n-r$ orthonormal vectors h_i in $\mathbb{R}^n$ such that $\langle h_i, \nabla u^k\rangle$ goes to zero in $L^1(\Omega)$ as k goes to infinity.

(1) We begin by remarking that $a = 0 \in \text{r.i.}(\hat{F}_1)$ and $a_k \in \text{r.i.}(\hat{F})$; by [C1, Thm. 1], $(\nabla u^k(x), g^{**}(\nabla u^k(x))) \in F$ and by $\nabla u^k(x) \in \hat{F}$ almost everywhere on Ω; hence $|\nabla u^k(x)| \leq M$ almost everywhere on Ω. Recalling Lemma 2.1, let $h_1, \ldots, h_{n-r}$ be the vectors defining the cone C such that $\hat{F} \subset C$. We have $\langle h_1, \nabla u^k\rangle \geq 0$ almost everywhere in Ω; writing $\nabla u^k = a_k + \nabla v^k$ with $v^k \in W_0^{1,1}(\Omega)$ it turns out that almost everywhere in Ω,

$$\langle h_1, \nabla v^k\rangle \geq -\langle h_1, a_k\rangle . \tag{4.8}$$

We extend v^k by setting $\tilde{v}^k = v^k$ on Ω and $\tilde{v}^k = 0$ on Ω^c; $\tilde{v}^k$ is in $W^{1,1}(\mathbb{R}^n)$ with compact support. Take a basis $\{e_1, \ldots, e_n\}$ in $\mathbb{R}^n$ such that $e_i = h_i$, $i = 1, \ldots, n-r$, and write a point ξ of $\mathbb{R}^n$ as $\xi = (\xi_1, \ldots, \xi_n) = (\xi_1, \xi')$, where ξ_i is the component with respect to e_i. Define the functions

$$\varphi_{\xi'}(\xi_1) = \tilde{v}^k(\xi_1, \xi');$$

$\varphi_{\xi'}(\cdot)$ is a function of $W^{1,1}(\mathbb{R})$ with compact support for almost every ξ' (see [Z, p. 44]) and this implies that the integral of its derivative is equal to zero. Since

$$\frac{d}{d\xi_1}\varphi_{\xi'}(\xi_1) = \langle h_1, \nabla\tilde{v}^k(\xi_1,\xi')\rangle\,,$$

this means that

$$\int_{\mathbb{R}} (\langle h_1, \nabla\tilde{v}^k(\xi_1,\xi')\rangle)^-\, d\xi_1 = \int_{\mathbb{R}} (\langle h_1, \nabla\tilde{v}^k(\xi_1,\xi')\rangle)^+\, d\xi_1.$$

By repeated integration and by a unitary change of variables, we obtain

$$\int_{\Omega} (\langle h_1, \nabla v^k(x)\rangle)^-\, dx = \int_{\Omega} (\langle h_1, \nabla v^k(x)\rangle)^+\, dx.$$

Remarking that the right-hand side of (4.8) is negative, we then have

$$\int_{\Omega} |\langle h_1, \nabla v^k(x)\rangle|\, dx = 2\int_{\Omega} (\langle h_1, \nabla v^k(x)\rangle)^-\, dx \le 2|a_k|\mu(\Omega),$$

and

$$\int_{\Omega} |\langle h_1, \nabla u^k(x)\rangle|\, dx \le 3|a_k|\mu(\Omega).$$

Hence $\langle h_1, \nabla u^k\rangle \overset{k\to\infty}{\to} 0$ in $L^1(\Omega)$.

(2) Now let $\epsilon > 0$. By Egorov's theorem there exists a compact subset Ω_ϵ of Ω such that $\mu(\Omega\backslash\Omega_\epsilon) \le \epsilon$ and $\langle h_1, \nabla u^k\rangle \overset{k\to\infty}{\longrightarrow} 0$ uniformly on Ω_ϵ. Let $k_\epsilon \in \mathbb{N}$ such that $|a_k| \le \epsilon$ and $\sup_{\Omega_\epsilon} |\langle h_1, \nabla u^k\rangle| \le \epsilon$ for any $k \ge k_\epsilon$. For $x \in \Omega_\epsilon$ and $k \ge k_\epsilon$, $\nabla u^k(x)$ belongs to an ϵ-neighbourhood of $\hat{F} \cap H_1$, where $H_1 = \{x : \langle h_1, x\rangle = 0\}$. A point y in an ϵ-neighbourhood of $\hat{F} \cap H_1$ can be written as $y = y_1 + y_\epsilon$, where $y_1 \in \hat{F} \cap H_1$ and $|y_\epsilon| \le \epsilon$; we have $\langle h_2, y_1\rangle \ge 0$ and $\langle h_2, y\rangle = \langle h_2, y_1\rangle + \langle h_2, y_\epsilon\rangle \ge \langle h_2, y_\epsilon\rangle \ge -\epsilon$. Hence $\langle h_2, \nabla u^k(x)\rangle \ge -\epsilon$ and $\langle h_2, \nabla v^k(x)\rangle \ge -\epsilon - |a_k| \ge -2\epsilon$ for any $x \in \Omega_\epsilon$. By computations analogous to those of point (1), we obtain, for any $k \ge k_\epsilon$,

$$\int_{\Omega_\epsilon} |\langle h_2, \nabla v^k(x)\rangle|\, dx \le 4\epsilon\mu(\Omega);$$

then,

$$\int_{\Omega} |\langle h_2, \nabla u^k(x)\rangle|\, dx = \int_{\Omega\backslash\Omega_\epsilon} |\langle h_2, \nabla u^k(x)\rangle|\, dx + \int_{\Omega_\epsilon} |\langle h_2, \nabla u^k(x)\rangle|\, dx \le \epsilon M + 5\epsilon\mu(\Omega).$$

Hence $\langle h_2, \nabla u^k\rangle \overset{k\to\infty}{\longrightarrow} 0$ in $L^1(\Omega)$.

This process can be iterated to show that $\langle h_i, \nabla u^k\rangle \overset{k\to\infty}{\longrightarrow} 0$ in $L^1(\Omega)$ for $i = 1, \dots, n-r$ and this proves that $\{u^k\}_{k\in\mathbb{N}}$ converges d-strongly to u_a in $W_0^{1,1}(\Omega)$ for some $d \ge n-r$. By point (i), if $d > n-r$, we would have $\dim(F_1) < r$, a contradiction. Hence $\{u^k\}_{k\in\mathbb{N}}$ converges $(n-r)$-strongly in $W_0^{1,1}(\Omega)$. □

COROLLARY 4.1. *Assume the hypotheses of Theorem* 4.1. *Then any sequence* $\{u^k\}_{k\in\mathbb{N}}$ *of solutions of* $\mathcal{P}_{a_k}^{**}$ *converges weakly to* u_a.

Proof. Take any sequence $\{u^k\}_{k\in\mathbb{N}}$ of solutions of $\mathcal{P}^{**}_{a_k}$. The derivatives of u^k are uniformly bounded almost everywhere, and by point (ii) of Theorem 4.1, $u^k \to 0$ in $L^1(\Omega)$. Then the proof is a straightforward application of Lemma 4.1. □

Remark 4.2 (Rotational symmetry). Theorem 4.2 states that in general continuous dependence of the solutions from boundary data does not hold if the point $(a, g^{**}(a))$ is not extremal. However, in one special case continuous dependence holds whenever the solution is unique: assume indeed that the function g is rotationally symmetric, i.e., there exists $h : \mathbb{R}^+ \to \mathbb{R}^+$ such that $g(\nabla u) = h(|\nabla u|)$. Two cases are possible: either (a) $h(0) < h(r)$ for all $r > 0$ or (b) there exists $R > 0$ such that $h(R) = h(0)$ (assume that R is the largest such point). In case (a) there are no extremal faces of $\mathrm{epi}(g^{**})$ having dimension n; hence, by the previous results, uniqueness and continuous dependence hold for every boundary datum. In case (b) there is a unique n-dimensional face, the ball of radius R, whose relative boundary consists of its extreme points; hence both uniqueness and continuous dependence hold if and only if $|a| \geq R$.

For a more general g we have the following result.

COROLLARY 4.2. *Assume the hypotheses of Theorem* 3.2.

(i) *If the point* $\{a\}$ *is such that* $\mathcal{P}_a$ $(\mathcal{P}^{**}_a)$ *admits the unique solution* u_a *and there exists a neighbourhood* U *of* $\{a\}$ *such that for any* $b \in U$ $(b, g^{**}(b))$ *belongs to a face of* $\mathrm{epi}(g^{**})$ *of dimension strictly less than* n, *then continuous dependence holds in* U.

(ii) *If the point* $\{a\}$ *is such that* $\mathcal{P}_a$ $(\mathcal{P}^{**}_a)$ *admits the unique solution* u_a *and* $(a, g^{**}(a))$ *belongs to the relative boundary of an* n*-dimensional face of* $\mathrm{epi}(g^{**})$, *then continuous dependence holds if and only if* $(a, g^{**}(a))$ *is an extreme point of* $\mathrm{epi}(g^{**})$.

Remark 4.3 (example). Consider the (convex) integrand $g = g^{**} : \mathbb{R}^2 \to \bar{\mathbb{R}}$, defined as

$$g(y_1, y_2) = \begin{cases} |y_1| + |y_2| & \text{if } (y_1, y_2) \in D, \\ +\infty & \text{otherwise,} \end{cases}$$

where $D = \{(y_1, y_2) \in \mathbb{R}^2 : |y_1| + |y_2| \leq 1\}$. Consider the corresponding minimum problem $\mathcal{P}_a$, where a is a point of $\mathbb{R}^2$ and Ω is any open bounded subset of $\mathbb{R}^2$.

By the above results $\mathcal{P}_a$ admits infinite solutions for any $a = (a_1, a_2) \in \mathrm{int}(D)$ such that $a_1 a_2 \neq 0$, while existence and uniqueness hold for $a \in \partial D \bigcup \{a = (a_1, a_2) \in \mathbb{R}^2 : a_1 a_2 = 0\} = E$. Set $b_0 = (0,0)$, $b_1 = (1,0)$, $b_2 = (0,1)$, $b_3 = (-1,0)$, $b_4 = (0,-1)$; the points $(b_i, g(b_i))$ $(i = 0, \ldots, 4)$ are the only extreme points of the epigraph of g, while the line segments joining $(b_0, g(b_0))$ to $(b_i, g(b_i))$ $(i = 1, \ldots, 4)$ are one-dimensional faces of the epigraph of g.

According to our main result, for $a \in E \setminus \bigcup_{i=0}^{4} b_i$ the solution of $\mathcal{P}_a$ is unique but (strong) continuous dependence does not hold, while both uniqueness and continuous dependence hold for $a = b_i$ $(i = 0, \ldots, 4)$.

Remark 4.4. In Theorem 4.1 we assume that the sequence $\{(a_k, g^{**}(a_k))\}_{k\in\mathbb{N}}$ is entirely contained in a *fixed* n-dimensional face F. We can suppose otherwise that, as k goes to infinity, the sequence touches different faces of the epigraph of g^{**}. In this case the proof of statement (i) of Theorem 4.1 does not need any modification since to prove that $\dim(F_1) = r$ it is sufficient to consider a subsequence of $\{(a_k, g^{**}(a_k))\}_{k\in\mathbb{N}}$ entirely contained in the relative interior of one n-dimensional face.

Conversely we may assume $\dim(F_1) = r$ and study the behaviour of a sequence of solutions $\{u^k\}_{k\in\mathbb{N}}$ of $\mathcal{P}^{**}_{a_k}$ when $a^k \to a$ and $(a_k, g^{**}(a_k))$ belongs to more than one face. In general we may assume that there exists a finite collection $\{F_1, \ldots, F_q\}$ of n-dimensional faces that contain $(a, g^{**}(a))$ in their respective relative boundaries

and such that $(a_k, g^{**}(a_k))$ belongs to to the relative interior of each of the F_i for an infinite number of indices k. The sequence $\{u^k\}_{k\in\mathbb{N}}$ can be decomposed in the disjoint union of $q+1$ subsequences $\{u^{k_i}\}_{k_i\in\mathbb{N}}$, $i=1,\dots,q+1$, where the indices $k_1,\dots,k_q$ are those ones for which $(a_{k_i}, g^{**}(a_{k_i}) \in \text{r.i.}(F_i)$ while the values $(a_{k_{q+1}}, g^{**}(a_{k_{q+1}}))$ belong to faces of dimension strictly less than n. Since $u^{k_{q+1}} \equiv \langle a_{k_{q+1}}, \cdot\rangle$ it converges strongly to u_a in $W_0^{1,1}(\Omega)$, while the sequences $\{u^{k_i}\}_{k_i\in\mathbb{N}}$, $i=1,\dots,q$ converge $(n-r)$-strongly to u_a in $W_0^{1,1}(\Omega)$ in the sense that $\langle \nabla u^{k_i}, e_j^i\rangle \overset{k_i\to\infty}{\longrightarrow} \langle \nabla u_a, e_j^i\rangle$ in $L^1(\Omega)$, where $E^i = \{e_1^i,\dots,e_{(n-r)}^i\}$ is an orthonormal system in $(\text{span}(\hat{F}_1 - a))^\perp$ for any $i=1,\dots,q$. Hence the whole sequence $\{u^k\}_{k\in\mathbb{N}}$ converges $(n-r)$-strongly (and also weakly) to u_a in $W_0^{1,1}(\Omega)$. Hence statement (ii) of Theorem 4.1 and Corollary 4.1 remain true.

REFERENCES

[A] Z. Artstein, *A note on Fatou's Lemma in several dimensions.* J. Math. Economics, 6 (1992), pp. 277–282 .

[AR] Z. Artstein and T. Rzezuchowski *A note on Olech's lemma*, Studia Math., 98 (1991), pp. 91–94.

[B] H. Brezis, *Analyse fonctionnelle*, Masson, Paris 1983.

[Ba] E. J. Balder, *On weak convergence implying strong convergence under extremal conditions.* J. Math. Anal. Appl., 163 (1992), pp. 147–156.

[C1] A. Cellina, *On minima of a functional of the gradient: Necessary conditions*, Nonlinear Anal. T.M.A., 20 (1993), pp. 337–341.

[C2] ———, *On minima of a functional of the gradient: Sufficient conditions*, Nonlinear Anal. T.M.A., 20 (1993), pp. 343–347.

[D] B. Dacorogna, *Direct methods in calculus of variations*, Springer Verlag, Berlin, 1989.

[ET] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North Holland, Amsterdam, 1976.

[O1] C. Olech, *Integrals of set-valued functions and linear optimal control problems*, Colloque sur la Théorie Mathématique du Contrôle Optimal, C.B.R.M. Vander Louvain, 1970, pp. 109–125.

[O2] ———, *Lectures on the integration of set-valued functions.* Lectures at Scuola Internazionale Superiore di Studi Avanzati (SISSA), 1987.

[R] R. T. Rockafellar, *Convex Analysis.* Princeton University Press, Princeton, NJ, 1972.

[Re] T. Rzezukowski, *Strong convergence of selections implied by weak*, Bull. Australian Math. Soc., 39 (1989), pp. 201-214.

[V] A. Visintin, *Strong convergence related to strict convexity*, Comm. Partial Differential Equations, 9 (1984), pp. 439-466.

[Z] W. P. Ziemer, *Weakly differentiable functions*, Springer Verlag, New York, 1989.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1128–1150, July 1994

014

CHARACTERIZATION OF THE $\mathcal{L}_2$-INDUCED NORM FOR LINEAR SYSTEMS WITH JUMPS WITH APPLICATIONS TO SAMPLED-DATA SYSTEMS*

N. SIVASHANKAR† AND PRAMOD P. KHARGONEKAR‡

Abstract. This paper considers a continuous-time linear system with finite jumps at discrete instants of time. An iterative method to compute the $\mathcal{L}_2$-induced norm of a linear system with jumps is presented. Each iteration requires solving an algebraic Riccati equation. It is also shown that a linear feedback interconnection of a continuous-time finite-dimensional linear time-invariant (FDLTI) plant and a discrete-time finite-dimensional linear shift-invariant (FDLSI) controller can be represented as a linear system with jumps. This leads to an iterative method to compute the $\mathcal{L}_2$-induced norm of a sampled-data system.

Key words. sampled-data systems, $\mathcal{H}_\infty$ control theory, digital control, Riccati differential equations, optimal control

AMS subject classifications. 93B50, 93C35, 93C05, 49A40

1. Introduction.

Consider a sampled-data system consisting of a continuous-time linear plant and a discrete-time linear controller. This system contains signals that evolve in continuous time as well as signals that evolve in discrete time. It is rather difficult to apply the standard analysis results for linear continuous-time systems and linear discrete-time systems to the analysis of sampled-data systems. This fact has motivated much of the recent research on the analysis and the synthesis of sampled-data control systems. In a recent paper, Sun, Nagpal, and Khargonekar, [32] have shown that *linear systems with jumps* are useful in the synthesis of $\mathcal{H}_\infty$ controllers for sampled-data systems. Roughly speaking, a linear continuous-time system with jumps is a standard linear continuous-time system whose state undergoes finite jump discontinuities at discrete instants of time. It turns out that the class of linear systems with jumps contains standard linear continuous-time systems, linear discrete-time systems, and sampled-data systems. The interested reader is referred to the book by Lakshmikantham, Bainov, and Simeonov [24] for a general introduction to systems with jump discontinuities.

The main problem considered in this paper is the analysis and computation of the $\mathcal{L}_2$-induced norm of linear systems with jumps. The $\mathcal{L}_2$-induced norm is very closely related to the $\mathcal{H}_\infty$ norm; recall that the $\mathcal{L}_2$-induced norm of a linear time-invariant system is the $\mathcal{H}_\infty$ norm of its transfer function. The main results of this paper show that the $\mathcal{L}_2$-induced norm of a linear system with jumps can be computed by solving matrix Riccati equations. Based on comparing our results and Riccati equations with the well-known results on the characterization and computation of the $\mathcal{H}_\infty$ norm of standard linear time-invariant systems [1], [8], [36], we believe that these results are the most natural and direct generalizations of these classical results. Both the finite-

* Received by the editors December 9, 1991; accepted for publication (in revised form) Janaury 25, 1993. This work was supported in part by National Science Foundation grant ECS-9001371, Air Force Office of Scientific Research contract AFOSR-90-0053, and Army Research Office grant DAAL03-90-G-0008.

† Department of Electrical Engingeering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109-2122. Current address, Scientific Research Labs, Ford Motor Company, Dearborn, Michigan 48121.

‡ Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109-2122 (pramod@dip.eecs.umich.edu).

and infinite-horizon cases are treated. The finite-horizon result, Theorem 5.2, is given in terms of existence of solution to a matrix Riccati differential equation with jumps over the finite horizon. The infinite-horizon case is treated by analyzing the limiting behavior of the finite-horizon Riccati equation. This leads to a characterization in Theorem 5.3 of the $\mathcal{L}_2$-induced norm in terms of existence of a stabilizing solution to a Riccati differential equation with jumps on the infinite horizon. This result is intuitivly appealing in that the Riccati equation is given directly in terms of the system parameters and contains the standard continuous and discrete-time Riccati equations as special cases in a very natural manner. However, this characterization is not convenient for computations. For this purpose, we give an equivalent characterization in Theorem 5.1, where we show that in the infinite-horizon case, the $\mathcal{L}_2$-induced norm can be characterized in terms of *existence of a stabilizing solution to one discrete-time algebraic Riccati equation and invertibility of a matrix function.* This theorem immediately leads to an algorithm for computing the $\mathcal{L}_2$-induced norm of a linear system with jumps. As should be expected, the standard analysis results on the $\mathcal{H}_\infty$ norm of linear time-invariant continuous-time and discrete-time systems are simple corollaries of these results.

The $\mathcal{L}_2$-induced norm can be used to provide necessary and sufficient conditions for robust stability of sampled-data systems (see [29] and the references therein for details). Thus, the results in the paper have immediate applications to the robust stability analysis of sampled-data feedback systems.

In addition to being of independent interest, in our view, the framework of linear systems with jumps gives the most natural setting for extending the well-known results on $\mathcal{H}_\infty$ control of standard linear time-invariant systems to the case of hybrid systems containing both discrete-time and continuous-time signals. As will be seen, the framework of systems with jumps contain the standard continuous-time, discrete-time, and sampled-data systems, and in this sense it is a unifying framework. Also, as far as controller synthesis is concerned, Sun, Nagpal, and Khargonekar [32] have used linear systems with jumps in deriving results on $\mathcal{H}_\infty$ controller synthesis for sampled-data systems. It should be noted that our results cannot be obtained as a special case of the $\mathcal{H}_\infty$ synthesis results. This is because our analysis condition, naturally, involves both the continuous-time plant and the discrete-time controller parameters while the synthesis solution depends only on the continuous-time plant parameters.

The last few years have seen a surge in research activity in the analysis and the synthesis of sampled-data control systems. Some basic issues regarding stability of sampled-data systems have been addressed in [15], [10]. Analysis and synthesis of sampled-data systems using various system norms has been studied in [6], [7], [2]–[5], [11], [12], [14], [18], [20], [23], [28], [32]–[34], [37]. Many of these papers use the "lifting technique" (see, for example, [22]) to convert the sampled-data system to an equivalent discrete-time system with infinite-dimensional input and output spaces (see [5], [2], [34], [37] and the references therein for details). More recently, direct state-space solutions to the $\mathcal{H}_\infty$ control problem for sampled-data systems without resorting to the lifting technique have been given in [32], [33].

Our results are most closely related to the results on the computation of the $\mathcal{L}_2$-induced norm of a sampled-data system that has been investigated in [2], [5], [17], [18]. In [18], it has been shown that the $\mathcal{L}_2$-induced norm of a sampled-data system is less than a prespecified number if and only if an associated discrete-time descriptor system has no eigenvalues on the unit circle and the norm of a certain

infinite-dimensional operator is less than one. The approach taken in [17], [18] is based on a representation of the sampled-data system with a state vector that contains continuous- as well as discrete-time signals. In [2], [5], it has been shown that the $\mathcal{L}_2$-induced norm of a sampled-data system is less than a prespecified number if and only if the $\mathcal{H}_\infty$ norm of an associated discrete-time system is less than one and the norm of a related infinite-dimensional operator is less than one. The approach taken in [2], [5] is based on lifting the sampled-data system to a discrete-time system with infinite-dimensional input and output spaces. Our paper provides an alternative solution to the problem treated in [2], [5], [17], [18]. The main difference between our work and these papers lies in the use of linear systems with jumps and the recent state-space time-domain approach to the $\mathcal{H}_\infty$ control theory (see the recent books [7], [31] and the references cited there). As stated above, the resulting characterizations appear to be the most natural generalizations of the classical results on the characterization of the $\mathcal{H}_\infty$ norm of standard continuous- and discrete-time systems using Riccati equations. Since we use a time-domain approach, we get results for the finite- and infinite-horizon cases, which is one distinction between our approach and [2], [5], [17], [18]. Since we do not use the lifting approach, our Riccati equations are given quite directly in terms of the problem data. When our results are specialized to the case of sampled-data systems, they are *mathematically equivalent* to the results of [2], [5], [18], although the form of the results is so different that this equivalence is not easy to see. In this context, our condition on the invertibility of a matrix function is related to the norm condition on a certain infinite-dimensional linear operator in [5], [2], [18]. (See the remarks following Theorem 5.1 for further details.) From a computational point of view, our conditions are as easy to check as those in [5], [2], [18].

The paper is organized as follows. The class of linear systems with jumps is introduced in the next section. We show in §3 that sampled-data systems are a special case of linear systems with jumps. In §4 we introduce the worst case performance measure and give the problem formulation. The main results of this paper are contained in §5. We present the proofs of the main results in §6 and give some concluding remarks in §7.

We end this section with some remarks on the notation used in this paper. Let $\mathcal{C}^n$ denote the space of continuous functions from the time set $[0,\infty)$ to $\Re^n$ and let $\mathcal{PC}^n$ denote the space of piecewise-continuous functions from the time set $[0,\infty)$ to $\Re^n$ that are bounded on compact sets of $[0,\infty)$ and are continuous from the left at every point except the origin. We will denote by $\mathcal{L}_2^n\,[a,b]$ the standard Lebesgue space of square integrable functions over the time interval $[a,b]$ with values in $\Re^n$, and the $\mathcal{L}_2$ norm is defined as

$$\|f\|_{[a,b]} := \left\{\int_a^b f'(t)f(t)dt\right\}^{1/2}.$$

If $f \in \mathcal{L}_2^n\,[0,\infty)$, then its $\mathcal{L}_2$ norm is defined as

$$\|f\|_2 := \left\{\int_0^\infty f'(t)f(t)dt\right\}^{1/2}.$$

Similarly, in discrete-time $\mathcal{S}^n$ denotes the space of $\Re^n$-valued sequences defined on the time set {0,1,2,...}, $\ell_2^n\,[0,k]$ denotes the space of all $(k+1)$ length sequences with

values in $\Re^n$, and the ℓ_2 norm of a $(k+1)$ length sequence $\{\xi_i\}_{i=0}^k$ is defined as

$$\|\xi\|_{[0,k]} := \left[\sum_{i=0}^{k} \xi_i' \xi_i\right]^{1/2}.$$

Again, if $\{\xi_i\}_{i=0}^{\infty} \in \ell_2^n$, then its ℓ_2 norm is defined as

$$\|\xi\|_2 := \left[\sum_{i=0}^{\infty} \xi_i' \xi_i\right]^{1/2}.$$

We will drop the superscript n in the subsequent sections because the dimension of the signal space will be clear from the context.

Finally, for a matrix function M,

$$M : \Re \to \Re^{n \times n},$$

the *left limit of* M *at* $\alpha \in \Re$ is defined as

$$M(\alpha^-) := \lim_{\epsilon \downarrow 0} M(\alpha - \epsilon)$$

if the limit exists, and the *right limit of* M *at* $\alpha \in \Re$ is defined as

$$M(\alpha^+) := \lim_{\epsilon \downarrow 0} M(\alpha + \epsilon)$$

if the limit exists.

2. Linear systems with jumps. Consider the system of equations

$$\Sigma : \begin{cases} \dot{x}(t) = Ax(t) + Bw(t); \quad x(0) = 0; \quad t \neq iT, \\ x(iT^+) = A_d x(iT) + B_d w_d(iT), \\ z(t) = Cx(t), \\ z_d(iT) = C_d x(iT), \end{cases} \tag{1}$$

where i is a nonnegative integer and T is a real number. Here x is the state vector, w and z are the *continuous-time* input and output, and w_d and z_d are the *discrete-time* input and output. It is clear from (1) that the state x of the system jumps at discrete instants of time iT. The state $x(t)$ is left continuous but may be right discontinuous with finite jumps at $t = iT$. The following properties are of significance:

- By setting $A_d = I$, $B_d = 0$, and $C_d = 0$ in (1), we recover standard finite-dimensional linear time-invariant (FDLTI) continuous-time systems.
- By setting $A = 0$, $B = 0$, and $C = 0$ in (1), we recover standard finite-dimension linear shift-invariant (FDLSI) discrete-time systems.
- It will be seen in §3 that a linear feedback interconnection of a FDLTI continuous-time system and a FDLSI discrete-time controller by sample and (zero-order) hold devices leads to a linear system with jumps.

Suitable generalizations of the class of linear systems with jumps also have potential applications in the analysis and synthesis of multirate systems and systems with nonuniform sampling period. This topic will not be pursued in this paper and is left for future research.

Now consider the system Σ with $w = w_d = 0$. The solution $x(t)$ of the unforced system is given by

$$x(t) = \Phi(t, s)x(s); \qquad t \geq s,$$

where the matrix function $\Phi(t, s)$ is the *state transition matrix* of the system Σ. It is piecewise continuous with possible discontinuities at $t, s = iT$, and is characterized by

$$\begin{aligned} \frac{\partial}{\partial t}\Phi(t, s) &= A\Phi(t, s); \qquad t \geq s; \quad t \neq iT, \\ \Phi(iT^+, s) &= A_d\Phi(iT, s); \qquad iT \geq s, \\ \Phi(s, s) &= I. \end{aligned}$$

DEFINITION 2.1 ([32]). *The system Σ in (1) is said to be internally exponentially stable if there exist positive constants c_1, c_2 such that*

$$\|\Phi(t, s)\| \leq c_1 e^{-c_2(t-s)}, \quad \textit{for all } t \geq s.$$

We will abbreviate *internally exponentially stable* as *stable.*

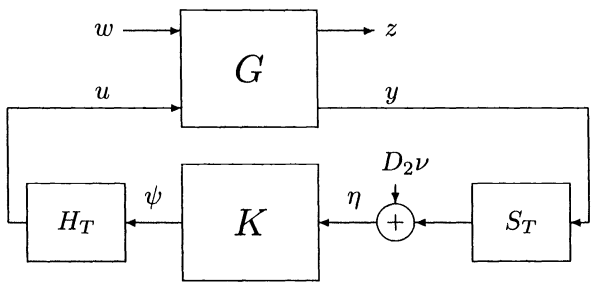


FIG. 1.

3. Sampled-data systems. In this section we consider a general linear interconnection of a FDLTI continuous-time plant and a FDLSI discrete-time controller by sample and hold devices. We show that such an interconnection is a special case of linear system with jumps. In particular, we show that such an interconnection has a state-space representation similar to (1).

Consider the sampled-data feedback system in Fig. 1. Here G is a FDLTI causal continuous-time plant, K is a FDLSI causal discrete-time controller, w is the exogenous input, u is the control input, z is the controlled output, and y is the measurement output. The block labeled S_T represents the sampling operator with time period T defined as follows:

$$S_T : \mathcal{C} \to \mathcal{S} : y \mapsto S_T y : (S_T y)(k) = y(kT).$$

The system block denoted by H_T represents the (zero-order) hold operator with time period T:

$$H_T : \mathcal{S} \to \mathcal{PC} : \psi \mapsto H_T\psi : (H_T\psi)(t) = \psi(k), \qquad kT < t \leq (k+1)T.$$

Consider the transfer function representation of G:

$$z = G_{11}w + G_{12}u, \qquad y = G_{21}w + G_{22}u.$$

We will assume throughout this paper that G_{22} is strictly proper. This ensures the well-posedness of the feedback system. In Fig. 1, note that S_T acts on the measurement output y. For this to make sense, y must be continuous. To ensure this, it is sufficient to assume that G_{21} is strictly proper. For the sake of simplicity in notation, we will assume that G_{11} is also strictly proper. However, this technical assumption can be easily removed.

Let

$$G: \quad \begin{aligned} \dot{x} &= Fx + E_1 w + E_2 u; \qquad x(0) = 0, \\ z &= H_1 x + D_{12} u, \\ y &= H_2 x; \end{aligned} \tag{2}$$

$$K: \quad \begin{aligned} \xi(k+1) &= \Phi\xi(k) + \Gamma\eta(k); \qquad \xi(0) = 0, \\ \psi(k) &= \Theta\xi(k) + \Upsilon\eta(k) \end{aligned} \tag{3}$$

be the state-space representations of the systems in Fig. 1. The input to the controller is corrupted by discrete-time noise and is given by

$$\eta(k) = y(kT) + D_2 w_d(k) = H_2 x(kT) + D_2 w_d(k).$$

The control input to the plant is constant between two sampling instants and is given by

$$u(t) = (H_T\psi)(t) = \psi(k), \qquad kT < t \leq (k+1)T.$$

The sampled-data feedback system in Fig. 1 is called internally asymptotically stable if the associated unforced shift-invariant discrete-time system with the state

$$\begin{pmatrix} x(k) \\ \xi(k) \end{pmatrix} := \begin{pmatrix} x(kT) \\ \xi(k) \end{pmatrix}$$

is asymptotically stable [11].

It has been shown in [15] that

$$x_{sd}(t) := \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}$$

represents the state of the closed loop hybrid sampled-data system where

$$\begin{aligned} x_1(t) &:= x(t) \quad \forall t, \\ x_2(t) &:= \psi(k), \qquad kT < t \leq (k+1)T, \\ x_3(t) &:= \xi(k+1), \qquad kT < t \leq (k+1)T. \end{aligned}$$

With this representation of the state of the closed-loop system in Fig. 1, we can obtain the state space representation of the hybrid system as

$$\Sigma_{sd}: \begin{cases} \dot{x}_{sd}(t) = Ax_{sd}(t) + Bw(t); \qquad x_{sd}(0) = 0, \quad t \neq kT, \\ x_{sd}(kT^+) = A_d x_{sd}(kT) + B_d w_d(k), \\ z(t) = Cx_{sd}(t), \end{cases} \tag{4}$$

where A, B, A_d, B_d, and H have the following representation:

$$A := \begin{pmatrix} F & E_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad B := \begin{pmatrix} E_1 \\ 0 \\ 0 \end{pmatrix},$$

$$A_d := \begin{pmatrix} I & 0 & 0 \\ \Upsilon H_2 & 0 & \Theta \\ \Gamma H_2 & 0 & \Phi \end{pmatrix}, \qquad B_d := \begin{pmatrix} 0 \\ \Upsilon D_2 \\ \Gamma D_2 \end{pmatrix},$$

$$C := (H_1 \;\; D_{12} \;\; 0).$$

Thus, the sampled-data system can be expressed (in state space form) as a linear system with finite jumps Σ_{sd}. Note that we do not have any discrete-time output in (4). However, if we want to include some function of the state x_{sd} or the input at the sampling instants in the cost function, then we will have a discrete-time output, z_d. It can be shown that the sampled-data feedback system in Fig. 1 is internally asymptotically stable if and only if the linear system with jumps Σ_{sd} described by (4) is stable.

4. Problem formulation. Consider the linear system with jumps Σ in (1). It is easy to see that Σ generates an input/output map $\mathcal{T}$ which is a causal linear operator

$$\mathcal{T} : \mathcal{L}_2[0,\tau] \oplus \ell_2[0,k] \to \mathcal{L}_2[0,\tau] \oplus \ell_2[0,k] : w \oplus w_d \mapsto z \oplus z_d, \tag{5}$$

where k is the largest integer such that $kT \leq \tau$. If Σ is stable, then

$$\mathcal{T} : \mathcal{L}_2[0,\infty) \oplus \ell_2 \to \mathcal{L}_2[0,\infty) \oplus \ell_2 : w \oplus w_d \mapsto z \oplus z_d \tag{6}$$

can be shown to be a causal bounded linear operator.

For the system Σ in (1), define the worst case performance measure $J(\tau)$ as

$$J(\tau) := \sup \left\{ \left[\frac{\|z\|^2_{[0,\tau]} + \|z_d\|^2_{[0,k]}}{\|w\|^2_{[0,\tau]} + \|w_d\|^2_{[0,k]}} \right]^{1/2} \right\}, \tag{7}$$

where the supremum is taken over all $w \in \mathcal{L}_2[0,\tau], w_d \in \ell_2[0,k]$ such that $\|w\|^2_{[0,\tau]} + \|w_d\|^2_{[0,k]} \neq 0$. Here, k is the largest integer such that $kT \leq \tau$.

The numerator and the denominator in the performance measure (7) should be thought of as "mixed" $\mathcal{L}_2/\ell_2$ norms on the inputs and the outputs of the system. Essentially, the performance measure $J(\tau)$ is the worst case ratio of the output energy to the input energy. Thus, $J(\tau)$ can be viewed as a worst case gain of the system Σ. In the infinite-horizon case, when the system Σ is stable, we will denote the performance measure as $J(\infty)$. This performance measure has been motivated by the $\mathcal{H}_\infty$ norm for standard linear systems [13], [21], [32]. Recall that the $\mathcal{H}_\infty$ norm of a stable FDLTI system is equal to its $\mathcal{L}_2$-induced norm. Thus, the performance measure defined above is a generalization of the usual $\mathcal{H}_\infty$ norm. Indeed, $J(\infty)$ reduces to the $\mathcal{H}_\infty$ norm if we specialize Σ to standard continuous-time or discrete-time systems. Let $\mathcal{T}$ (as described in (5)) be the linear operator associated with the system Σ. Then it can be easily verified that the induced operator norm of $\mathcal{T}$ is given by $J(\tau)$. Similarly, $J(\infty)$ is the induced operator norm of $\mathcal{T}$ in (6).

PROBLEM STATEMENT. *Given a real number $\gamma > 0$, give necessary and sufficient conditions such that $J(\infty) < \gamma$.*

5. Characterization of the $\mathcal{L}_2$-induced norm. We will first present the main result of this paper and then give some auxiliary results and corollaries that go along with this main result. We will present our analysis results for linear systems with jumps. Since sampled-data systems are a special case of linear systems with jumps, it follows that the results in this section can be easily specialized to the case of sampled-data systems.

5.1. Main result. Consider the system Σ in (1) over the time interval $t \in [0, \infty)$ and the associated performance measure $J(\infty)$. In this section we will assume that the system Σ is stable.

For $t \in [0, T]$, let

$$\Pi(t) := \begin{pmatrix} \Pi_{11}(t) & \Pi_{12}(t) \\ \Pi_{21}(t) & \Pi_{22}(t) \end{pmatrix} := \exp\left[\begin{pmatrix} A & \gamma^{-2}BB' \\ -C'C & -A' \end{pmatrix} t\right] \tag{8}$$

and

$$\Lambda(t) := \begin{pmatrix} \Lambda_{11}(t) & \Lambda_{12}(t) \\ \Lambda_{21}(t) & \Lambda_{22}(t) \end{pmatrix} := \exp\left[-\begin{pmatrix} -A' & -\gamma^{-2}C'C \\ BB' & A \end{pmatrix} t\right]. \tag{9}$$

We now give the main result for the infinite horizon case.

THEOREM 5.1. *Consider the linear system with jumps in* (1) *over the time interval* $t \in [0, \infty)$. *Let the system given in* (1) *be stable. Let* $\gamma > 0$ *be a real number. Then the following statements are equivalent.*

(i) $J(\infty) < \gamma$.

(ii) *There exists a symmetric matrix* $\tilde{P}$ *such that* $\left(I - \gamma^{-2}B_d'\tilde{P}B_d\right) > 0$ *and*

$$H_1\begin{pmatrix} I \\ \tilde{P} \end{pmatrix} = H_2\begin{pmatrix} I \\ \tilde{P} \end{pmatrix} F_P, \tag{10}$$

where

$$H_1 := \begin{pmatrix} A_d & 0 \\ -C_d'C_d & I \end{pmatrix}\begin{pmatrix} \Pi_{11}(T) & \Pi_{12}(T) \\ \Pi_{21}(T) & \Pi_{22}(T) \end{pmatrix}, \qquad H_2 := \begin{pmatrix} I & -\gamma^{-2}B_dB_d' \\ 0 & A_d' \end{pmatrix} \tag{11}$$

and

$$F_P := \left[I - \gamma^{-2}B_dB_d'\tilde{P}\right]^{-1} A_d\left(\Pi_{11}(T) + \Pi_{12}(T)\tilde{P}\right)$$

has all its eigenvalues within the open unit disk (discrete-time stable matrix). Moreover, $\left(\Pi_{11}(t) + \Pi_{12}(t)\tilde{P}\right)$ *is invertible for all* $t \in [0, T]$.

(iii) *There exists a symmetric matrix* $\tilde{Q}$ *such that* $\left(I - \gamma^{-2}C_d\tilde{Q}C_d'\right) > 0$ *and*

$$G_1\begin{pmatrix} I \\ \tilde{Q} \end{pmatrix} = G_2\begin{pmatrix} I \\ \tilde{Q} \end{pmatrix} F_Q, \tag{12}$$

where

$$G_1 := \begin{pmatrix} A_d' & 0 \\ -B_dB_d' & I \end{pmatrix}\begin{pmatrix} \Lambda_{11}(T) & \Lambda_{12}(T) \\ \Lambda_{21}(T) & \Lambda_{22}(T) \end{pmatrix}, \qquad G_2 := \begin{pmatrix} I & -\gamma^{-2}C_d'C_d \\ 0 & A_d \end{pmatrix} \tag{13}$$

and

$$F_Q := \left[I - \gamma^{-2} C_d' C_d \tilde{Q}\right]^{-1} A_d' \left(\Lambda_{11}(T) + \Lambda_{12}(T)\tilde{Q}\right)$$

has all its eigenvalues within the open unit disk (discrete-time stable matrix). Moreover, $\left(\Lambda_{11}(t) + \Lambda_{12}(t)\tilde{Q}\right)$ *is invertible for all* $t \in [0, T]$.

Note that there are two parts in the statement of (ii). In the first part of statement (ii), we claim that $\tilde{P}$ must satisfy the following properties:

1. $\left(I - \gamma^{-2} B_d' \tilde{P} B_d\right) > 0$;
2. $H_1 \begin{pmatrix} I \\ \tilde{P} \end{pmatrix} = H_2 \begin{pmatrix} I \\ \tilde{P} \end{pmatrix} F_P$;
3. F_P is stable.

In the context of sampled-data systems, there is an intutive explanation for item 1 given above. Sampled-data systems can be viewed as linear finite-dimensional discrete-time systems with infinite-dimensional inputs and outputs [2], [5], [23]. It is well known that such a positive definite condition appears in the standard $\mathcal{H}_\infty$ analysis problem for discrete-time systems [16], [31]. So it is only natural that such a condition also appear in the $\mathcal{H}_\infty$ analysis problem for sampled-data systems.

Regarding item 2 above, with some tedious algebra it can be shown that the matrices H_1 and H_2 form a symplectic pair, i.e., $H_1 J H_1' = H_2 J H_2'$ where

$$J := \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Thus, along with item 3 above, it is clear that $\tilde{P}$ is the unique stabilizing solution to the discrete-time algebraic Riccati equation (10) [26], [35], [16].

In the second part of statement (ii) we claim that $\tilde{P}$ should be such that $(\Pi_{11}(t) + \Pi_{12}(t)\tilde{P})$ is invertible for all $t \in [0, T]$. This condition is equivalent to a certain $\mathcal{L}_2$-induced norm of the system Σ being less than γ over one period [21]. Again, in the context of sampled-data systems, this condition has an intuitive explanation. Note that the sampled-data feedback system runs open loop (without the controller) in between sampling instants. For the induced norm of the feedback system to be less than γ over $[0, \infty)$, it is obvious that the induced norm should necessarily be less than γ within a sampling period. Now observe that the matrix function $\Pi(t)$ in (8) depends only on the parameters of the open-loop plant G and this explains the invertibility condition in (ii). This invertibility condition is related to the invertibility of a certain infinite-dimensional linear operator that appears in other works on sampled-data systems [5], [2], [18]. Numerically, this invertibility condition can be checked either by doing a search over one period or by checking the existence and boundedness of the solution to a standard Riccati differential equation (14) over one period.

A systematic way of checking if $J(\infty) < \gamma$ could be as follows. We can first check the existence of a solution $\tilde{P}$ to items 1–3 in the first part of statement (ii). If such a (stabilizing) solution exists, then it is unique. Now we can check if the solution $\tilde{P}$ satisfies the invertibility condition in the second part of statement (ii). *Thus, the invertibility condition needs to be checked only after the computation of* $\tilde{P}$.

Remarks similar to the ones made above can also be made regarding statement (iii) in the theorem. So (10) and (12) represent generalized eigenvalue problems, and one can solve for $\tilde{P}$ and $\tilde{Q}$ using numerical linear algebra methods.

Computation of $J(\infty)$.

Assume that a tolerance level $\epsilon > 0$ is given.

Step 1. Set $\gamma_h = L$, where L is a sufficiently large real number and $\gamma_l = 0$.

Step 2. Set $\gamma = (\gamma_h + \gamma_l)/2$.

Step 3. Check if (10) has a symmetric solution satisfying the conditions stated in (ii) of Theorem 5.1.

Step 4a. If a solution exists and

If $(\gamma_h - \gamma_l) < \epsilon$, then STOP.

Else, set $\gamma_h = \gamma$ and go to Step 2.

Step 4b. If a solution does not exist, then set $\gamma_l = \gamma$ and go to Step 2.

The above iterative procedure gives a simple way of calculating the induced norm of system Σ in (1) to a desired accuracy.

5.2. Finite- and infinite-horizon results. In the previous section we gave a characterization of the $\mathcal{L}_2$-induced norm of the system Σ in the infinite-horizon case. As stated in the Introduction, this result is obtained by analyzing the finite- and infinite-horizon cases. Therefore, we will next give the corresponding result in the finite-horizon case. As we can see, this result is a natural generalization of the known results on the finite-horizon $\mathcal{L}_2$-induced norm of standard linear systems (see, for example, [7], [21], [31]).

THEOREM 5.2. *Consider the linear system with jumps in* (1) *over a finite-time interval* $[0, \tau]$. *Let* k *be the largest integer such that* $0 \leq kT \leq \tau$. *Let* $\gamma > 0$ *be a real number. Then the following statements are equivalent.*

(i) $J(\tau) < \gamma$.

(ii) *There exists a symmetric piecewise differentiable matrix function* $P(t)$, $t \in [0, \tau]$ *such that* $(I - \gamma^{-2} B_d' P(iT^+) B_d) > 0$ *for all* $i \in \{0, 1, \ldots, k\}$ *and*

$$-\dot{P}(t) = A' P(t) + P(t)A + \gamma^{-2} P(t) B B' P(t) + C' C; \qquad t \neq iT, \tag{14}$$

$$\begin{aligned} P(iT) = {} & A_d' P(iT^+) A_d + C_d' C_d \\ & + A_d' P(iT^+) B_d \left(\gamma^2 I - B_d' P(iT^+) B_d\right)^{-1} B_d' P(iT^+) A_d, \end{aligned} \tag{15}$$

$$P(\tau^+) = 0. \tag{16}$$

(iii) *There exists a symmetric piecewise differentiable matrix function* $Q(t)$, $t \in [0, \tau]$ *such that* $(I - \gamma^{-2} C_d Q(iT^-) C_d') > 0$ *for all* $i \in \{0, 1, \ldots k\}$ *and*

$$\dot{Q}(t) = AQ(t) + Q(t)A' + \gamma^{-2} Q(t) C' C Q(t) + BB'; \qquad t \neq iT, \tag{17}$$

$$\begin{aligned} Q(iT) = {} & A_d Q(iT^-) A_d' + B_d B_d' \\ & + A_d Q(iT^-) C_d' \left(\gamma^2 I - C_d Q(iT^-) C_d'\right)^{-1} C_d Q(iT^-) A_d', \end{aligned} \tag{18}$$

$$Q(0^-) = 0. \tag{19}$$

Thus, to check if the performance measure $J(\tau)$ is less than the prespecified level γ, we must check the existence of a symmetric solution to a matrix Riccati differential equation with finite jumps. This solution reflects the hybrid nature of our system and the performance measure. The Riccati differential equation for P (14)–(16) is solved backwards in time. For the sake of simplicity, assume that the terminal time instant $\tau \in (kT, (k+1)T)$ for some positive integer k. We first integrate the differential equation (14) with the terminal condition $P(\tau^+) = 0$ up to the time instant $(kT + \epsilon)$,

where $\epsilon > 0$ is a sufficiently small real number. At the jump point kT, we use the difference equation (15) with the value of $P(kT+\epsilon)$ to obtain the initial value to the Riccati differential equation in the next interval. This procedure is repeated up to $t = 0$. If the solution $P(t)$ blows up for some $t \in [0, \tau]$, then clearly the solution to (14)–(16) does not exist and therefore $J(\tau) \geq \gamma$. Similarly, (17)–(18) is solved forward in time with initial condition $Q(0^-) = 0$ to obtain $Q(t), t \in [0, \tau]$. If $Q(t)$ blows up for some $t \in [0, \tau]$, then it again indicates that $J(\tau) \geq \gamma$. This leads to an obvious bisection method as in the previous subsection for computing $J(\tau)$ to any desired accuracy.

We will now present a result that is analogous to Theorem 5.2 in the infinite-horizon case. This result, along with Theorem 5.2, will be useful in establishing Theorem 5.1. Again, this result is a natural generalization of the corresponding results for standard continuous- and discrete-time linear systems.

Given a piecewise-continuous matrix function $P(t), t \in [0, \infty)$, define the linear system with jumps:

$$\Sigma_P : \begin{cases} \dot{x}(t) = (A + \gamma^{-2} B B' P(t)) x(t); & x(0) = 0; \qquad t \neq iT, \\ x(iT^+) = (I - \gamma^{-2} B_d B_d' P(iT^+))^{-1} A_d x(iT). \end{cases} \tag{20}$$

Similarly, given a piecewise-continuous matrix function $Q(t), t \in [0, \infty)$, define

$$\Sigma_Q : \begin{cases} \dot{x}(t) = (A + \gamma^{-2} Q(t) C' C) x(t); & x(0) = 0; \qquad t \neq iT, \\ x(iT) = A_d (I - \gamma^{-2} Q(iT^-) C_d' C_d)^{-1} x(iT^-). \end{cases} \tag{21}$$

Now we state the extension of Theorem 5.2 for the infinite-horizon case.

THEOREM 5.3. *Consider the linear system with jumps in* (1) *over the time interval* $t \in [0, \infty)$. *Let the system given in* (1) *be stable. Let* $\gamma > 0$ *be a real number. Then the following statements are equivalent.*

(i) $J(\infty) < \gamma$.

(ii) *There exists a bounded symmetric piecewise differentiable matrix function* $P(t)$, $t \in [0, \infty)$ *such that* $(I - \gamma^{-2} B_d' P(iT^+) B_d) > 0$ *for all* $i \in \{1, 2, \ldots\}$, $(I - \gamma^{-2} B_d' P(iT^+) B_d)^{-1}$ *is a bounded sequence,* $P(t)$ *satisfies* (14), (15) *for all* $t \in [0, \infty)$, *and the system* Σ_P *is stable.*

(iii) *There exists a bounded symmetric piecewise differentiable matrix function* $Q(t)$, $t \in [0, \infty)$ *such that* $(I - \gamma^{-2} C_d Q(iT^-) C_d') > 0$ *for all* $i \in \{1, 2, \ldots\}$, $(I - \gamma^{-2} C_d Q(iT^-) C_d')^{-1}$ *is a bounded sequence,* $Q(t)$ *satisfies* (17), (18) *for all* $t \in [0, \infty)$ *with* $Q(0^-) = 0$, *and the system* Σ_Q *is stable.*

Note that both $P(t)$ and $Q(t)$ in Theorem 5.3 are given by a differential Riccati equation in the interval between consecutive jumps, and at the beginning of every interval the initial condition to the Riccati differential equation is evaluated using a difference Riccati equation. The equation for $P(t)$ has no terminal condition and is obtained by taking limit of the finite-horizon solution (14), (15) as $\tau \to \infty$.

Since we are dealing with linear time-invariant systems with finite discrete jumps at periodic intervals, we should expect some sort of stationarity property in the Riccati equations in the infinite horizon case. *Intuitively, this stationarity property leads to the results in Theorem* 5.1 *from Theorem* 5.3. As we show in §6, $P(t)$ (respectively, $Q(t)$) in Theorem 5.3 is intrinsically related to $\tilde{P}$ (respectively, $\tilde{Q}$) in Theorem 5.1. In particular, due to the periodic nature of the underlying system, $P(t)$ is periodic (with a period T) and $\tilde{P} = P(iT^+)$ for $i \in \{0, 1, 2, \ldots\}$. The invertibility condition on

$\left(\Pi_{11}(t) + \Pi_{12}(t)\tilde{P}\right)$ in Theorem 5.1 guarantees the existence of $P(t)$ in between sampling instants. (Note that an analogous invertibility condition is absent in Theorem 5.3 since the existence of $P(t)$ is required for all t.)

Since continuous-time and discrete-time FDLTI systems are special cases of a linear system with jumps, it should come as no surprise that the standard analysis conditions for continuous-time and discrete-time FDLTI systems [13], [31] fall out naturally from Theorems 5.3 and 5.1.

COROLLARY 5.4. *Consider the linear time-invariant continuous-time asymptotically stable system*

$$\Sigma_c : \begin{cases} \dot{x}(t) = Ax(t) + Bw(t); & x(0) = 0, \\ z(t) = Cx(t). \end{cases} \tag{22}$$

Then,

$$J(\infty) = \sup_{0=w\in\mathcal{L}_2[0,\infty)} \frac{\|z\|_2}{\|w\|_2} =: \|T_{zw}\|_\infty.$$

Let $\gamma > 0$. Then, the following statements are equivalent.

(i) $\|T_{zw}\|_\infty < \gamma$.

(ii) *There exists a unique symmetric matrix P such that*

$$A'P + PA + C'C + \gamma^{-2}PBB'P = 0, \tag{23}$$

$(A + \gamma^{-2}BB'P)$ is asymptotically stable.

(iii) *There exists a unique symmetric matrix Q such that*

$$AQ + QA' + BB' + \gamma^{-2}PC'CP = 0, \tag{24}$$

$(A + \gamma^{-2}QC'C)$ is asymptotically stable.

COROLLARY 5.5. *Consider the linear shift-invariant discrete-time asymptotically stable system*

$$\Sigma_d : \begin{cases} x(k+1) = A_d x(k) + B_d w_d(k); & x(0) = 0, \\ z_d(k) = C_d x(k). \end{cases} \tag{25}$$

Then,

$$J(\infty) = \sup_{0=w_d\in\ell_2} \frac{\|z_d\|_2}{\|w_d\|_2} =: \|T_{z_d w_d}\|_\infty.$$

Let $\gamma > 0$. Then, the following statements are equivalent.

(i) $\|T_{z_d w_d}\|_\infty < \gamma$.

(ii) *There exists a unique symmetric matrix P such that $(I - \gamma^{-2}B_d'PB_d) > 0$,*

$$A_d'PA_d - P + C_d'C_d + A_d'PB_d(\gamma^2 I - B_d'PB_d)^{-1}B_d'PA_d = 0, \tag{26}$$

and $(A_d + B_d(\gamma^2 I - B_d'PB_d)^{-1}B_d'PA_d)$ is asymptotically stable.

(iii) *There exists a unique symmetric matrix Q such that $(I - \gamma^{-2}C_dQC_d') > 0$,*

$$A_dQA_d' - Q + B_dB_d' + A_dQC_d'(\gamma^2 I - C_dQC_d')^{-1}C_dQA_d' = 0, \tag{27}$$

and $(A_d + A_dQC_d'(\gamma^2 I - C_dQC_d')^{-1}C_d)$ is asymptotically stable.

6. Proofs. This section is divided into two parts. In the first part, we will give a proof for the finite-horizon result, and in the second part we will give a proof for the infinite-horizon result.

Consider the system (1) over the time interval $t \in [0, \tau]$. Suppose $\tau = kT$ for some positive integer k and there exists a piecewise differentiable matrix function $P(t)$ satisfying (14)–(16). Applying the terminal condition $P(\tau^+) = P(kT^+) = 0$ to (15) we get

$$P(kT) = P(\tau) = C_d' C_d.$$

Then, it is clear that condition (ii) in Theorem 5.2 is equivalent to the following condition.

There exists a symmetric piecewise differentiable matrix function $P(t)$, $t \in [0, \tau]$ such that $(I - \gamma^{-2} B_d' P(iT^+) B_d) > 0$ for all $i \in \{0, 1, \ldots, k\}$ and

$$-\dot{P}(t) = A' P(t) + P(t) A + \gamma^{-2} P(t) B B' P(t) + C' C; \qquad t \neq iT, \tag{28}$$

$$\begin{aligned} P(iT) = {} & A_d' P(iT^+) A_d + C_d' C_d \\ & + A_d' P(iT^+) B_d \left(\gamma^2 I - B_d' P(iT^+) B_d\right)^{-1} B_d' P(iT^+) A_d, \end{aligned} \tag{29}$$

$$P(\tau) = C_d' C_d. \tag{30}$$

Also note that if $\tau = kT$, it follows that

$$x'(\tau) P(\tau) x(\tau) = x'(\tau) C_d' C_d x(\tau) = z_d'(kT) z_d(kT). \tag{31}$$

If the terminal time instant lies between two consecutive jump instants, i.e., $\tau \in (kT, (k+1)T)$ for some nonnegative integer k (k depends on τ), then it is clear that $P(\tau) = P(\tau^+) = 0$.

We state a lemma next that will be used repeatedly in the proofs. This lemma can be established using routine "completion of squares" and "dynamic programming" arguments [36], [21]. For the sake of brevity, we do not give the proof of the lemma here; the details of this proof can be found in [30].

LEMMA 6.1. *Consider the system given in* (1). *Let* $J(\tau) < \gamma$. *Suppose there exists a symmetric piecewise differentiable matrix function* $P(t)$ *satisfying* (14)–(16) *in the interval* $[\sigma, \tau]$ *for some* $\sigma \geq 0$.

(i) *Then for* $t_0 \in [\sigma, \tau]$,

$$\begin{aligned} V(t_0, x_0) &:= \inf_{w, w_d} \left[\gamma^2 \|w\|^2_{[t_0,\tau]} - \|z\|^2_{[t_0,\tau]} + \gamma^2 \|w_d\|^2_{[m,l]} - \|z_d\|^2_{[m,l]} \,|x(t_0) = x_0\right] \\ &= -x_0' P(t_0) x_0, \end{aligned} \tag{32}$$

where m *is the smallest integer satisfying* $0 \leq \sigma \leq t_0 \leq mT$ *and*

$$l = \begin{cases} k & \text{if } \tau \in (kT, (k+1)T), \\ k+1 & \text{if } \tau = (k+1)T \end{cases}$$

for some nonnegative integer k. *The infimum in* (32) *is achieved by*

$$w(t) = \gamma^{-2} B' P(t) x(t) \quad \forall\, t \neq iT, \qquad t \in [t_0, \tau], \tag{33}$$

$$w_d(iT) = (\gamma^2 I - B_d' P(iT^+) B_d)^{-1} B_d' P(iT^+) A_d x(iT) \quad \forall\, i \in \{m, m+1, \ldots, l\}. \tag{34}$$

(ii) *Furthermore,*

$$V(t_0,0) := \inf_{w,w_d} \left[\gamma^2\|w\|^2_{[t_0,\tau]} - \|z\|^2_{[t_0,\tau]} + \gamma^2\|w_d\|^2_{[m,l]} - \|z_d\|^2_{[m,l]} \,|x(t_0) = 0\right] = 0 \tag{35}$$

and the infimum in (35) *is obtained with* $w(t) = 0, t \in [t_0,\tau]$ *and* $w_d(iT) = 0, i \in \{m,\dots,l\}$.

6.1. Finite-horizon case.

Proof of Theorem 5.2. We will first show (i) $\Leftrightarrow$ (ii) and then show (i) $\Leftrightarrow$ (iii) by duality.

Proof of (i) $\Leftarrow$ (ii) *in Theorem* 5.2. Let $P(t), t \in [0,\tau]$ be a piecewise differentiable matrix function satisfying (14)–(16). Let k be the largest integer such that $kT < \tau$. (To avoid triviality, assume $\tau > 0$.) Differentiating the function $x(t)P(t)x(t)$ and then integrating it from iT^+ to s (where $s := \min((i+1)T,\tau)$) we obtain

$$x'(t)P(t)x(t)|^s_{iT^+} = -\|z\|^2_{(iT,s]} + \gamma^2\|w\|^2_{(iT,s]} - \left\|\left(\gamma w - \frac{1}{\gamma}B'Px\right)\right\|^2_{(iT,s]}. \tag{36}$$

Similarly, at $t = iT$, we have

$$\begin{aligned} &x'(iT^+)P(iT^+)x(iT^+) - x'(iT)P(iT)x(iT) \\ &\quad = -z_d'(iT)z_d(iT) + \gamma^2 w_d'(iT)w_d(iT) - v'(iT)M_i v(iT), \end{aligned} \tag{37}$$

where $M_i := (\gamma^2 I - B_d'P(iT^+)B_d)$, and $v(iT) := w_d(iT) - M_i^{-1}B_d'P(iT^+)A_d x(iT)$.

Set

$$\hat{w}(t) := \gamma w(t) - \frac{1}{\gamma}B'P(t)x(t) \quad \forall\, t \in [0,\tau] \tag{38}$$

and

$$\hat{w}_d(iT) := M_i^{1/2}\left(w_d(iT) - M_i^{-1}B_d'P(iT^+)A_d x(iT)\right) \quad \forall i \in \{0,1,\dots,k\}. \tag{39}$$

Consider the system Σ_w :

$$\begin{aligned} \dot{x}(t) &= \left[A + \gamma^{-2}BB'P(t)\right]x(t) + \frac{1}{\gamma}B\hat{w}(t); \qquad x(0) = 0; \qquad t \neq iT, \\ x(iT^+) &= \left[A_d + B_d M_i^{-1}B_d'P(iT^+)A_d\right]x(iT) + B_d M_i^{-1/2}\hat{w}_d, \\ w(t) &= \frac{1}{\gamma}\hat{w}(t) + \gamma^{-2}B'P(t)x(t), \\ w_d(iT) &= M_i^{-1/2}\hat{w}_d(iT) + M_i^{-1}B_d'P(iT^+)A_d x(iT). \end{aligned} \tag{40}$$

Let $\mathcal{T}_w$ be the input/output operator generated by Σ_w, which can be described as

$$\mathcal{T}_w : \mathcal{L}_2[0,\tau] \oplus \ell_2[0,k] \to \mathcal{L}_2[0,\tau] \oplus \ell_2[0,k] : \hat{w} \oplus \hat{w}_d \mapsto w \oplus w_d. \tag{41}$$

As τ is finite, it is easy to see that the operator $\mathcal{T}_w$ is bounded, and there exists a real number $c > 0$ such that

$$\|w\|^2_{[0,\tau]} + \|w_d\|^2_{[0,k]} \le c\left(\|\hat{w}\|^2_{[0,\tau]} + \|\hat{w}_d\|^2_{[0,k]}\right). \tag{42}$$

If $\tau \neq (k+1)T$, then adding (36) and (37) from $i=0$ to $i=k$, we obtain

$$x'(\tau)P(\tau)x(\tau) - x'(0)P(0)x(0) = -\|z\|^2_{[0,\tau]} + \gamma^2\|w\|^2_{[0,\tau]} - \|\hat{w}\|^2_{[0,\tau]} - \|z_d\|^2_{[0,k]} + \gamma^2\|w_d\|^2_{[0,k]} - \|\hat{w}_d\|^2_{[0,k]}.$$

With $x(0)=0, P(\tau)=0$ and (42), it follows from the above equation that

$$0 \leq -\|z\|^2_{[0,\tau]} - \|z_d\|^2_{[0,k]} + (\gamma^2 - 1/c)\left(\|w\|^2_{[0,\tau]} + \|w_d\|^2_{[0,k]}\right).$$

This is true for all inputs $w, w_d \in \mathcal{L}_2[0,\tau] \oplus \ell_2[0,k]$ and the corresponding outputs $z, z_d \in \mathcal{L}_2[0,\tau] \oplus \ell_2[0,k]$, which proves that $J(\tau) \leq \sqrt{\gamma^2 - 1/c} < \gamma$ for the case $\tau \neq (k+1)T$.

If $\tau = (k+1)T$, then with $x(0)=0$, it follows that

$$\begin{aligned} 0 &\leq -\|z\|^2_{[0,\tau]} - \|z_d\|^2_{[0,k]} + (\gamma^2 - 1/c)\left(\|w\|^2_{[0,\tau]} + \|w_d\|^2_{[0,k]}\right) - x'(\tau)P(\tau)x(\tau) \\ &= -\|z\|^2_{[0,\tau]} - \|z_d\|^2_{[0,k+1]} + (\gamma^2 - 1/c)\left(\|w\|^2_{[0,\tau]} + \|w_d\|^2_{[0,k]}\right). \end{aligned}$$

Here, the last equality follows from (31). This establishes $J(\tau) < \gamma$ for the case $\tau = (k+1)T$.

Proof of (i) $\Rightarrow$ (ii) *in Theorem* 5.2. Now given that $J(\tau) < \gamma$, we must show that there exists a symmetric, piecewise-differentiable matrix function $P(t)$ satisfying (14)–(16). A brief outline of this proof is as follows. Let $\tau \in (kT, (k+1)T]$, where k is some integer. Note that $P(t)$ is a solution to a differential equation with jumps moving backward in time with a boundary condition at τ. So we first establish the existence of $P(t)$ in the interval $(kT, \tau]$, i.e., the last interval. Then, using the jump equation of the state in (1) we show the existence of $P(kT)$ satisfying the jump equation (15). Next, we show the existence of $P(t)$ in the interval $((k-1)T, kT)$. The existence of $P(t)$ for all $t \in [0,\tau]$ is established by repeating these arguments.

Step 1. For the case $\tau \neq (k+1)T$, the existence of $P(t)$ satisfying (14) in the interval $(kT,\tau]$ follows from [25, Thm. 2.3]. For the case $\tau = (k+1)T$, the numerator of $J(\tau)$ includes a term $z_d'((k+1)T)z_d((k+1)T)$. Clearly,

$$z_d'((k+1)T)z_d((k+1)T) = x_d'((k+1)T)C_d'C_dx_d((k+1)T).$$

Since $\tau = (k+1)T$, this term can be viewed as a penalty on the final state in the cost function $J(\tau)$. Such a cost function has been previously considered in [21]. Indeed, the existence of $P(t)$ satisfying (28) in the interval $(kT,\tau]$ follows from [21, Thm. 2.2] in this case. We will assume hereafter in this proof that $\tau \neq (k+1)T$. It is very easy to extend the proof to the case $\tau = (k+1)T$.

Since the solution to (14) exists in $(kT,\tau]$, it follows from Lemma 6.1 that

$$\inf_w \left[\gamma^2\|w\|^2_{(kT,\tau]} - \|z\|^2_{(kT,\tau]} \,\middle|\, x(kT^+) = x_0\right] = -x_0'P(kT^+)x_0. \tag{43}$$

It is easy to verify that

$$\begin{aligned} &\gamma^2 w_d'(kT)w_d(kT) - z_d'(kT)z_d(kT) - x'(kT^+)P(kT^+)x(kT^+) \\ &= \begin{pmatrix} w_d'(kT) & x'(kT) \end{pmatrix} \begin{pmatrix} M_k & -B_d'P(kT^+)A_d \\ -A_d'P(kT^+)B_d & -C_d'C_d - A_d'P(kT^+)A_d \end{pmatrix} \begin{pmatrix} w_d(kT) \\ x(kT) \end{pmatrix}, \end{aligned}$$

where $M_k := \gamma^2 I - B_d' P(kT^+) B_d$.

Now assume that $w(t) = 0$ for all $t \in [0, kT]$, $w_d(iT) = 0$ for all $i \in \{0, 1, \ldots, (k-1)\}$, and $w_d(kT) \neq 0$. Choose $w(t), t \in (kT, \tau]$ to be the input that achieves the infimum in the first part of Lemma 6.1. Since $x(0) = 0$, it follows that $x(kT) = 0$ and $z_d(kT) = 0$. Using the definition of M_k and the system equations, we obtain

$$\gamma^2 w_d'(kT) w_d(kT) - z_d'(kT) z_d(kT) - x'(kT^+) P(kT^+) x(kT^+) = w_d'(kT) M_k w_d(kT).$$

Using (43), Lemma 6.1, and the fact that $J(\tau) < \gamma$, it follows that there exists an $\epsilon > 0$ such that

$$\begin{aligned} w_d'(kT) M_k w_d(kT) &= \gamma^2 w_d'(kT) w_d(kT) - z_d'(kT) z_d(kT) - x'(kT^+) P(kT^+) x(kT^+) \\ &= \left[\gamma^2 w_d'(kT) w_d(kT) - z_d'(kT) z_d(kT) + \gamma^2 \|w\|^2_{(kT,\tau]} - \|z\|^2_{(kT,\tau]} \,|\, x(kT) = x(t_0) = 0\right] \\ &\geq \epsilon \|w_d(kT)\|^2. \end{aligned}$$

Therefore

$$M_k = (\gamma^2 I - B_d' P(kT^+) B_d) \geq \epsilon I. \tag{44}$$

Now set $P(kT)$ to be

$$P(kT) := A_d' P(kT^+) A_d + C_d' C_d + A_d' P(kT^+) B_d \left(\gamma^2 I - B_d' P(kT^+) B_d\right)^{-1} B_d P(kT^+) A_d.$$

Thus, we have shown the existence of solution $P(t)$ to (14), (15) in the interval $t \in [kT, \tau]$.

Step 2. We now show the existence of solution to (14), (15) on the interval $((k-1)T, kT)$. Since the solution to (14), (15) exists on the interval $[kT, \tau]$, using Lemma 6.1 it follows that

$$\begin{aligned} &\inf_{w, w_d} \left[\gamma^2 \|w\|^2_{(kT,\tau]} - \|z\|^2_{(kT,\tau]} + \left[\gamma^2 w_d'(kT) w_d(kT) - z_d'(kT) z_d(kT)\right] \,|\, x(kT) = x_0\right] \\ &\quad = -x_0' P(kT) x_0. \end{aligned}$$

Let $w(t) = 0$, for all $t \in (0, (k-1)T]$ and $w_d(iT) = 0$ for all $i \in \{1, \ldots, (k-1)\}$. Since $x(0) = 0$, it follows that $x((k-1)T^+) = 0$. Then, using dynamic programming arguments,

$$\begin{aligned} &\inf_{w, w_d} \left[\gamma^2 \|w\|^2_{((k-1)T,\tau]} - \|z\|^2_{((k-1)T,\tau]} \right. \\ &\qquad \left. + \left[\gamma^2 w_d'(kT) w_d(kT) - z_d'(kT) z_d(kT)\right] \,|\, x((k-1)T^+) = 0\right] \\ &\quad = \inf_{w} \left[\gamma^2 \|w\|^2_{((k-1)T,kT]} - \|z\|^2_{((k-1)T,kT]} - x'(kT) P(kT) x(kT) \,|\, x((k-1)T^+) = 0\right]. \end{aligned}$$

Since $J(\tau) < \gamma$, it follows that there exists an $\epsilon > 0$ such that

$$\begin{aligned} &\inf_{w} \left[\gamma^2 \|w\|^2_{((k-1)T,kT]} - \|z\|^2_{((k-1)T,kT]} - x'(kT) P(kT) x(kT) \,|\, x((k-1)T^+) = 0\right] \\ &\qquad \geq \epsilon \|w\|^2_{((k-1)T,kT]}. \end{aligned}$$

Then,

$$\sup_{w} \left\{ \frac{\|z\|^2_{((k-1)T,kT]} + x'(kT) P(kT) x(kT)}{\|w\|^2_{((k-1)T,kT]}} : \|w\|_{((k-1)T,kT]} \neq 0 \right\} < \gamma^2.$$

This is a finite-horizon $\mathcal{H}_\infty$ norm analysis problem with a penalty on the final state. Using [21, Thm. 2.2] it follows that there exists a symmetric matrix function $P(t), t \in ((k-1)T, kT)$ that satisfies (14).

By repeating the above arguments, we can show the existence of $P(t)$ satisfying (14) and (15) in the interval $t \in [0, (k-1)T]$. This completes the proof of (i) $\Leftrightarrow$ (ii) in Theorem 5.2.

Proof of (i) $\Leftrightarrow$ (iii) *in Theorem* 5.2.

This follows using standard arguments using adjoints and is omitted. See [30] for details. □

6.2. Infinite-horizon case.

Proof of Theorems 5.1 *and* 5.3.

We will prove (i) $\Leftrightarrow$ (ii) in Theorems 5.1 and 5.3 next. To maintain clarity and flow in the proofs, we will first prove (i) $\Rightarrow$ (ii) in Theorem 5.3 and then the same implication will be proved for Theorem 5.1. We will then follow this up with the proof of (i) $\Leftarrow$ (ii) in Theorems 5.3 and 5.1.

Proof of (i) $\Rightarrow$ (ii) *in Theorem* 5.3. Let the solution to (14)–(15) at time t with the final time τ and final time condition $P(\tau^+) = 0$ be denoted by $P(t, \tau)$. Consider the system in (1). From the finite-horizon results we know that if $J(\tau) < \gamma$, then the solution to (14)–(16) $P(t, \tau)$ exists for all $t \leq \tau < \infty$. In the infinite-horizon case, it can be shown that if $J(\infty) < \gamma$, then [27], [30]

1. There exists a $\beta > 0$, such that $P(t, \tau) \leq \beta I$ for all $t \in [0, \tau]$ and all $\tau \in [0, \infty)$,
2. $P(t, \tau)$ is nondecreasing as a function of the final time $\tau \in [0, \infty)$.

From items (1) and (2) above, it follows that $P(t, \tau)$ is a bounded function of t and τ and is nondecreasing with respect to τ. Hence,

$$\lim_{\tau \to \infty} P(t, \tau) =: \bar{P}(t)$$

exists and is bounded on $[0, \infty)$. Since ϵ is independent of τ and k in (44), it follows that for all $i \in \{0, 1, 2, \ldots\}$,

$$(\gamma^2 I - B_d' \bar{P}(iT^+) B_d) \geq \epsilon I > 0.$$

Therefore, $(I - \gamma^{-2} B_d' \bar{P}(iT^+) B_d)^{-1}$ is a bounded sequence. It now follows that (using arguments as in [19]) $\bar{P}(t)$ satisfies (14), (15) for all $t \in [0, \infty)$. We now show that Σ_P is stable.

Stability of Σ_P. Consider the system in (1). Since $J(\infty) < \gamma$, there exists an $\epsilon > 0$, such that $J(\infty) \leq \gamma - \epsilon$. Therefore, $J(\tau) \leq \gamma - \epsilon$ for all $\tau \in [0, \infty)$. Then, for any $t_0 \leq \tau$, with $x_0 := x(t_0) = 0$, it follows that

$$(45) \left[\|z\|^2_{[t_0,\tau]} + \|z_d\|^2_{[m,k]} - \gamma^2 (\|w\|^2_{[t_0,\tau]} + \|w_d\|^2_{[m,k]}) \right] \leq -\epsilon \left[\|w\|^2_{[t_0,\tau]} + \|w_d\|^2_{[m,k]} \right],$$

where m is the smallest integer such that $t_0 \leq mT$ and k is the largest integer such that $kT \leq \tau$. Given any inputs $w(t), w_d(iT)$ for $t, iT \in [t_0, \tau]$, and initial condition $x_0 \neq 0$, we can decompose the outputs $z(t)$ and $z_d(iT)$ as

$$z(t) = z_0(t) + z_i(t), \qquad z_d(iT) = z_{d0}(iT) + z_{di}(iT),$$

where z_0, z_{d0} are the homogeneous parts of the solution (depending only on x_0) and z_i, z_{di} are the forced parts (depending only on w, w_d.) Now, since the system (1) is

stable, it follows that there exists a constant $\alpha > 0$ (independent of t_0 and τ) such that

$$\sqrt{\left[\|z_0\|^2_{[t_0,\tau]} + \|z_{d0}\|^2_{[m,k]}\right]} < \alpha\|x_0\|. \tag{46}$$

Using arguments as in [27], it can be established that

(47)

$$\left[\|z\|^2_{[t_0,\tau]} + \|z_d\|^2_{[m,k]} - \gamma^2\|w\|^2_{[t_0,\tau]} - \gamma^2\|w_d\|^2_{[m,k]}\right]$$
$$\leq 2\alpha^2\|x_0\|^2 - \sqrt{\left[\|w\|^2_{[t_0,\tau]} + \|w_d\|^2_{[m,k]}\right]}\left[\epsilon\sqrt{\left[\|w\|^2_{[t_0,\tau]} + \|w_d\|^2_{[m,k]}\right]} - 4\alpha\gamma\|x_0\|\right].$$

Now consider the system Σ_P given in (20) where $P(t)$ is the solution $\bar{P}(t)$ to (14), (15). This system can be viewed as

$$\Sigma_{aux} : \begin{cases} \dot{x}(t) &= Ax(t) + Bw(t); \quad x(0) = 0; \quad t \neq iT, \\ x(iT^+) &= A_d x(iT) + B_d w_d(iT), \end{cases} \tag{48}$$

where the inputs are given by

$$w(t) = \gamma^{-2}B'\bar{P}(t)x(t)\ w_d(iT) = (\gamma^2 I - B_d'\bar{P}(iT^+)B_d)^{-1}B_d'\bar{P}(iT^+)A_d x(iT). \tag{49}$$

Observe that these inputs have the same form as the inputs that achieve the infimum in Lemma 6.1.

We first claim that there exists $\xi > 0$ such that given any initial time t_0 and initial condition x_0, we have

$$\left[\|w\|^2_{[t_0,\infty)} + \|w_d\|^2_2\right] \leq \xi\|x_0\|^2. \tag{50}$$

If not, given any $\eta > 0$ there exist x_0, t_0, τ such that

$$\left[\|w\|^2_{[t_0,\tau]} + \|w_d\|^2_{[m,k]}\right] > \eta^2\|x_0\|^2, \tag{51}$$

where m is the smallest integer such that $t_0 \leq mT$ and k is the largest integer such that $kT \leq \tau$. In particular, let

$$\eta \geq \frac{2\alpha\gamma}{\epsilon}\left[1 + \sqrt{\left(1 + \frac{\epsilon}{2\gamma^2}\right)}\right], \tag{52}$$

where $\epsilon, \alpha > 0$ are as defined before in (45) and (46).

From (48), (51), and (52) it follows that

$$\left[\|z\|^2_{[t_0,\tau]} + \|z_d\|^2_{[m,k]} - \gamma^2\|w\|^2_{[t_0,\tau]} - \gamma^2\|w_d\|^2_{[m,k]}\right] < \left[2\alpha^2 - \eta^2\left(\epsilon - \frac{4\alpha\gamma}{\eta}\right)\right]\|x_0\|^2$$
$$< 0.$$

However, the inputs to the system Σ_{aux} in (48) achieve the infimum in Lemma 6.1. It then follows that

$$x_0'P(t_0,\tau)x_0 = \left[\|z\|^2_{[t_0,\tau]} + \|z_d\|^2_{[m,k]} - \gamma^2\|w\|^2_{[t_0,\tau]} - \gamma^2\|w_d\|^2_{[m,k]}\right] < 0,$$

which contradicts the fact that $P(t_0, \tau) \geq 0$ for all $\tau \in [0, \infty)$ and all $t_0 \in [0, \tau]$. Thus, given any initial time t_0 and initial condition x_0, there exists $\xi > 0$ (independent of t_0) such that (50) is satisfied. Since the system Σ in (1) is stable, it follows for the system (48) there exists $\delta_1, \delta_2 > 0$ such that

$$\|x\|_{[t_0,\infty)} \leq \left[\delta_1 \|w\|_{[t_0,\infty)} + \delta_2 \|w_d\|_2\right]. \tag{53}$$

Since the inputs w and w_d are chosen as given by (49), it follows from (50) that

$$\|x\|_{[t_0,\infty)} \leq (\delta_1 + \delta_2)\sqrt{\xi}\|x_0\|$$

and the constants are independent of the initial time instant t_0. By [9, Thm. 3, p. 190] it follows that Σ_P is stable. This proves (i) $\Rightarrow$ (ii) in Theorem 5.3.

Proof of (i) $\Rightarrow$ (ii) *in Theorem* 5.1. Since the coefficients of the differential and difference equations (14)–(15) are constant, it follows that

$$P(t + T, \tau + T) = P(t, \tau).$$

Now, letting $\tau \to \infty$, it follows that $\bar{P}(t)$ is periodic with a period T. We can write $\bar{P}((i+1)T)$ in terms of $\bar{P}(iT^+)$ by solving (14) over one period with initial condition $\bar{P}(iT^+)$ as

$$\bar{P}((i+1)T) = (\Pi_{21}(T) + \Pi_{22}(T)\bar{P}(iT^+))(\Pi_{11}(T) + \Pi_{12}(T)\bar{P}(iT^+))^{-1}, \tag{54}$$

where $\Pi(t), t \in [0, T]$ is as defined in (8). Also, using (15) we can write $\bar{P}((i+1)T)$ in terms of $\bar{P}((i+1)T^+)$ as

$$\bar{P}((i+1)T) = A_d' \bar{P}((i+1)T^+)\left(I - \gamma^{-2} B_d B_d' \bar{P}((i+1)T^+)\right)^{-1} A_d + C_d' C_d. \tag{55}$$

Since $\bar{P}(t)$ is periodic, it follows that $\bar{P}(iT^+) = \bar{P}((i+1)T^+) =: \tilde{P}$. Then, using (54) and (55) it follows that

$$\begin{aligned}(\Pi_{21}(T) + \Pi_{22}(T)\tilde{P}) = A_d' \tilde{P}\left(I - \gamma^{-2} B_d B_d' \tilde{P}\right)^{-1} A_d(\Pi_{11}(T) + \Pi_{12}(T)\tilde{P}) \\ + C_d' C_d(\Pi_{11}(T) + \Pi_{12}(T)\tilde{P}).\end{aligned} \tag{56}$$

Thus, $\tilde{P}$ defined by (56) is the solution to (14), (15) at $t = iT^+$ for all $i \in \{0, 1, 2, \ldots\}$. Using the fact that $\Pi(T)$ defined in (8) is a symplectic matrix, after some tedious algebra, it can be verified that we can write (56) as

$$H_1 \begin{pmatrix} I \\ \tilde{P} \end{pmatrix} = H_2 \begin{pmatrix} I \\ \tilde{P} \end{pmatrix} F_P, \tag{57}$$

where H_1, H_2, and F_P are as defined in Theorem 5.1.

Consider the system Σ_P given in (20) where $P(t)$ is the solution $\bar{P}(t)$ to (14), (15). Since $\bar{P}(t)$ is periodic, it follows that Σ_P is a stable periodic system. The key observation is that F_P is the state transition matrix of Σ_P from time instant 0^+ to T^+. To verify this, we first need to compute the state transition matrix of

$$\dot{x}(t) = (A + \gamma^{-2} B B' \bar{P}(t))x(t) \tag{58}$$

from $t = 0^+$ to $t = T$. Consider the system of equations

$$\begin{pmatrix} \dot{x}(t) \\ \dot{p}(t) \end{pmatrix} = \begin{pmatrix} A & \gamma^{-2}BB' \\ -C'C & -A' \end{pmatrix} \begin{pmatrix} x(t) \\ p(t) \end{pmatrix}.$$

Then $\Pi(\hat{t})$ as defined in (8) is the state transition matrix of this system of equations from 0 to $\hat{t}$ for $\hat{t} \in [0, T]$. Define the change of variables

$$\begin{pmatrix} r(t) \\ s(t) \end{pmatrix} = \begin{pmatrix} I & 0 \\ -\bar{P}(t) & I \end{pmatrix} \begin{pmatrix} x(t) \\ p(t) \end{pmatrix}.$$

Then it is easy to verify that in the new variables

$$\begin{pmatrix} \dot{r}(t) \\ \dot{s}(t) \end{pmatrix} = \begin{pmatrix} A + \gamma^{-2}BB'\bar{P}(t) & \gamma^{-2}BB' \\ 0 & -\left(A + \gamma^{-2}BB'\bar{P}(t)\right)' \end{pmatrix} \begin{pmatrix} r(t) \\ s(t) \end{pmatrix}. \tag{59}$$

Let the state transition matrix of this system of differential equations be $\Pi_P(t, \tau)$. Then simple algebra shows that

$$\begin{aligned} \Pi_P(\hat{t}, 0) &= \begin{pmatrix} I & 0 \\ -\bar{P}(\hat{t}) & I \end{pmatrix} \Pi(\hat{t}) \begin{pmatrix} I & 0 \\ \bar{P}(0^+) & I \end{pmatrix} \\ &= \begin{pmatrix} \Pi_{11}(\hat{t}) + \Pi_{12}(\hat{t})\bar{P}(0^+) & * \\ 0 & * \end{pmatrix}, \end{aligned} \tag{60}$$

where $*$ represent quantities of no interest. Since (59) is upper triangular, it follows that the state transition matrix of the system in (58) from $t = 0^+$ to $t = \hat{t}$ is given by

$$\Phi_{A+\gamma^{-2}BB'\bar{P}(t)}(\hat{t}, 0^+) := \Pi_{11}(\hat{t}) + \Pi_{12}(\hat{t})\bar{P}(0^+) = \Pi_{11}(\hat{t}) + \Pi_{12}(\hat{t})\tilde{P}, \tag{61}$$

where the last equality follows from the fact that $\bar{P}(0^+) = \tilde{P}$. Moreover, since $(\Pi_{11}(\hat{t}) + \Pi_{12}(\hat{t})\tilde{P})$ is the transition matrix of the system in (58), it is invertible for all $\hat{t} \in [0, T]$. Now the state transition matrix of the system Σ_P from the time instant 0^+ to T^+ is given by

$$\begin{aligned} \Phi_P(T^+, 0^+) &:= (I - \gamma^{-2}B_dB_d'\tilde{P})^{-1}A_d\Phi_{A+\gamma^{-2}BB'\bar{P}(t)}(T, 0^+) \\ &= (I - \gamma^{-2}B_dB_d'\tilde{P})^{-1}A_d\left(\Pi_{11}(T) + \Pi_{12}(T)\tilde{P}\right) = F_P. \end{aligned}$$

Since the system Σ_P is exponentially stable and periodic, it follows that F_P is a stable matrix and hence all its eigenvalues are within the open unit disk. This concludes the proof of (i) $\Rightarrow$ (ii) in Theorem 5.1.

Proof of (ii) $\Rightarrow$ (i) *in Theorem* 5.3. The proof of this part is along the same lines as that of the proof of (ii) $\Rightarrow$ (i) in Theorem 5.2. Observe that since the system Σ given by (1) is stable, it follows that for all $w, w_d \in \mathcal{L}_2[0, \infty) \oplus \ell_2$,

$$\lim_{t \to \infty} x(t) = 0, \tag{62}$$

where $x(t)$ is the state of Σ at time $t \in [0, \infty)$. Also, since P is bounded, Σ_P is stable, and $(I - \gamma^{-2}B_d'P(iT^+)B_d)^{-1}$ is a bounded sequence, it follows that there exists a $c > 0$ such that for the system Σ_w in (40),

$$\|w\|_2^2 + \|w_d\|_2^2 \leq c\left[\|\hat{w}\|_2^2 + \|\hat{w}_d\|_2^2\right]. \tag{63}$$

Then, proceeding as in the proof of Theorem 5.2, we obtain

$$\left[\lim_{\tau\to\infty} x'(\tau)P(\tau)x(\tau)\right] - x'(0)P(0)x(0) = -\|z\|_2^2 + \gamma^2\|w\|_2^2 - \|\hat{w}\|_2^2 - \|z_d\|_2^2 + \gamma^2\|w_d\|_2^2 - \|\hat{w}_d\|_2^2.$$

Now using (62), (63), and $x(0) = 0$, it follows that

$$0 \le -\|z\|_2^2 - \|z_d\|_2^2 + (\gamma^2 - 1/c)\left(\|w\|_2^2 + \|w_d\|_2^2\right)$$

for all inputs $w, w_d \in \mathcal{L}_2[0,\infty] \oplus \ell_2$ and the corresponding outputs $z, z_d \in \mathcal{L}_2[0,\infty] \oplus \ell_2$, which proves that $J(\infty) < \gamma$.

Proof of (ii) $\Rightarrow$ (i) *in Theorem* 5.1. We first show that if a solution to (10) exists with $\left(\Pi_{11}(t) + \Pi_{12}(t)\tilde{P}\right)$ invertible for all $t \in [0,T]$, then a solution $P(t)$ to (14), (15) for $t \in [0,\infty)$ also exists and is bounded. Set $P(0^+) = \tilde{P}$. Define

$$P(0) := A_d'\tilde{P}A_d + C_d'C_d + A_d'\tilde{P}B_d\left(\gamma^2 I - B_d'\tilde{P}B_d\right)^{-1}B_d'\tilde{P}A_d.$$

Since $\left(\Pi_{11}(t) + \Pi_{12}(t)\tilde{P}\right)$ is invertible, we can define $P(t), t \in (0,T]$ as

$$P(t) := \left(\Pi_{21}(t) + \Pi_{22}(t)\tilde{P}\right)\left(\Pi_{11}(t) + \Pi_{12}(t)\tilde{P}\right)^{-1}.$$

It is easily verified that $P(t)$ satisfies (14) for $t \in (0,T]$. Moreover, extend $P(t)$ to $t \in [0,\infty)$ as a periodic function:

$$P(t + kT) = P(t), \quad t \in [0,T] \quad k \in \{0,1,2,\ldots\}.$$

It follows that $P(iT^+) = \tilde{P}$ for all $i \in \{0,1,\ldots\}$. This $P(t)$ satisfies (14), (15) and is bounded for all $t \in [0,\infty)$. Also, since $\left(I - \gamma^{-2}B_d'\tilde{P}B_d\right) > 0$, it follows that $\left(I - \gamma^{-2}B_d'P(iT^+)B_d\right) > 0$ for all $i \in \{0,1,\ldots\}$ and $\left(I - \gamma^{-2}B_d'P(iT^+)B_d\right)^{-1}$ is a bounded sequence. As shown above, F_P is the state transition matrix from kT^+ to $(k+1)T^+$ of the system Σ_P. If F_P is a stable matrix, then it follows that Σ_P is stable. So (14), (15) has a bounded solution over the time interval $[0,\infty)$ and the system Σ_P is stable. From the proof of (ii) $\Rightarrow$ (i) in Theorem 5.3, it follows that $J(\infty) < \gamma$.

The proof of (i) $\Leftrightarrow$ (iii) in Theorems 5.3 and 5.1 can be obtained by duality and arguments similar to those used in the proof of (i) $\Leftrightarrow$ (ii) in Theorems 5.3 and 5.1. For the sake of brevity, this proof and the proofs of the corollaries are omitted in this paper. The interested reader can find the details of this proof in [30]. □

7. Conclusion. We have defined an $\mathcal{H}_\infty$-like worst case performance measure for linear systems with finite jumps. We have given necessary and sufficient conditions for the performance measure to be less than a prespecified level. Applying these results to sampled-data systems, we have given an iterative method to compute the $\mathcal{L}_2$-induced norm of a sampled-data system. The results of this paper can be easily extended to the case of a sampled-data system with a generalized hold function.

Acknowledgments. The authors gratefully acknowledge Dr. Anton Stoorvogel for noticing an error in a preliminary draft of this paper.

REFERENCES

[1] B. D. O. ANDERSON, *An algebraic solution to the spectral factorization problem,* IEEE Trans. Automat. Control, 12 (1967), pp. 410–414.

[2] B. BAMIEH AND J. B. PEARSON, *A general framework for linear periodic systems with applications to $\mathcal{H}_\infty$ sampled-data control,* IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[3] ———, *The $\mathcal{H}_2$ problem for sampled-data systems,* Systems Control Lett., 19 (1992), pp. 1–12.

[4] B. BAMIEH, M. DAHLEH, AND J. B. PEARSON, *Minimization of the $\mathcal{L}_\infty$-Induced Norm for Sampled-Data Systems,* Report No. 9109, Dept. of Electrical and Computer Engineering, Rice University, Houston, TX, June 1991.

[5] B. BAMIEH, J. B. PEARSON, B. A. FRANCIS, AND A. TANNENBAUM, *A lifting technique for linear periodic systems with applications to sampled-data control,* Systems Control Lett., 17 (1991), pp. 79–88.

[6] T. BAŞAR, *Optimal $\mathcal{H}_\infty$ designs under sampled state measurements,* Systems Control Lett., 16 (1991), pp. 399–410.

[7] T. BAŞAR AND P. BERNHARD, *$\mathcal{H}_\infty$ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach,* Birkhäuser, Boston, MA, 1991.

[8] S. BOYD, V. BALAKRISHNAN, AND P. KABAMBA, *A bisection method for computing the $\mathcal{H}_\infty$ norm of a transfer matrix and related problems,* Math. Control, Signals, Systems, 2 (1989), pp. 207–219.

[9] R. BROCKETT, *Finite Dimensional Linear Systems* , John Wiley, New York, 1970.

[10] T. CHEN AND B. A. FRANCIS, *Input-output stability of sampled-data systems,* IEEE Trans. Automat. Control, 36 (1991), pp. 50–58.

[11] ———, *$\mathcal{H}_2$-optimal sampled-data control,* IEEE Trans. Automat. Control, 36 (1991), pp. 387–397.

[12] ———, *On the $\mathcal{L}_2$-induced norm of a sampled-data system,* Systems Control Lett., 15 (1990), pp. 211–219.

[13] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard $\mathcal{H}_\infty$ and $\mathcal{H}_2$ control problems,* IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[14] G. DULLERUD AND B. A. FRANCIS, *$\mathcal{L}_1$ performance in sampled-data systems,* IEEE Trans. Automat. Control, 37 (1992), pp. 436–446.

[15] B. A. FRANCIS AND T. T. GEORGIOU, *Stability theory for linear time-invariant plants with periodic digital controllers,* IEEE Trans. Automat. Control, 33 (1988), pp. 820–832.

[16] P. A. IGLESIAS AND K. GLOVER, *State space approach to discrete-time $\mathcal{H}_\infty$ control,* Internat. J. Control, 54 (1991), pp. 1031–1073.

[17] P. T. KABAMBA AND S. HARA, *On computing the induced norm of a sampled-data system,* in Proceedings of the American Control Conference, San Diego, CA, 1990, pp. 319–320.

[18] ———, *Worst Case Analysis and Design of Sampled-Data Control Systems,* preprint, 1991.

[19] R. E. KALMAN, *Contributions to the theory of optimal control,* Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[20] J. P. KELLER AND B. D. O. ANDERSON, *A new approach to discretization of continuous-time controllers,* IEEE Trans. Automat. Control, 37 (1992), pp. 214–223.

[21] P. P. KHARGONEKAR, K. M. NAGPAL, AND K. POOLLA, *$\mathcal{H}_\infty$ control with transients,* SIAM J. Control Optim., 29 (1991), pp. 1373–1393.

[22] P. P. KHARGONEKAR, K. POOLLA, AND A. TANNENBAUM, *Robust control of linear time-invariant plants using periodic compensation,* IEEE Trans. Automat. Control, 30 (1985), pp. 1088–1096.

[23] P. P. KHARGONEKAR AND N. SIVASHANKAR, *$\mathcal{H}_2$ optimal control for sampled-data systems,* Systems Control Lett., 17 (1991), pp. 425–436.

[24] V. LAKSHMIKANTHAM, D. D. BAINOV, AND P. S. SIMEONOV, *Theory of Impulsive Differential Equations,* World Scientific, Teaneck, NJ, 1989.

[25] D. J. N. LIMEBEER, B. D. O. ANDERSON, P. P. KHARGONEKAR, AND M. GREEN, *A game theoretic approach to $\mathcal{H}_\infty$ control for time-varying systems,* SIAM J. Control Optim., 30 (1992), pp. 262–283.

[26] T. PAPPAS, A. J. LAUB, AND N. R. SANDELL, JR., *On the numerical solution of the discrete-time algebraic Riccati equation,* IEEE Trans. Automat. Control, 25 (1980), pp. 631–641.

[27] R. RAVI, K. M. NAGPAL, AND P. P. KHARGONEKAR, *$\mathcal{H}_\infty$ control of linear time-varying systems: A state-space approach,* SIAM J. Control Optim., 29 (1991), pp. 1394–1414.

[28] N. SIVASHANKAR AND P. P. KHARGONEKAR, *Induced norms for sampled-data systems,* Automatica, 28 (1992), pp. 1267-1272. A conference version appeared in the Proceedings of the

American Control Conference, 1991, pp. 167–172.

[29] ———, *Robust stability and performance analysis of sampled-data systems,* IEEE Trans. Automat. Control, 38(1993), pp. 58–69. A conference version appeared in the Proceedings of the 30th Conference on Decision and Control, 1991, pp. 881–886.

[30] ———, *Characterization of the $\mathcal{L}_2$-Induced Norm for Linear Systems with Jumps with Applications to Sampled-Data Systems,* Internal Report, University of Michigan, Ann Arbor, MI, 1991.

[31] A. A. STOORVOGEL, *The $\mathcal{H}_\infty$ Control Problem: A State-Space Approach,* Prentice-Hall, Englewood Cliffs, NJ, 1992.

[32] W. SUN, K. NAGPAL, AND P. P. KHARGONEKAR, *$\mathcal{H}_\infty$ control and filtering with sampled measurements,* in Proceedings of the American Control Conference, Boston, MA, 1991, pp. 1652–1657; IEEE Trans. Automat. Control, 38 (1993), pp. 1162–1175.

[33] G. TADMOR, *Optimal $\mathcal{H}_\infty$ sampled-data control in continuous time systems*, in Proceedings of the American Control Conference, Boston, MA, 1991, pp. 1658–1663.

[34] H. T. TOIVONEN, *Sampled-data control of continuous-time system with an $\mathcal{H}_\infty$ optimality criterion,* Automatica, 28 (1992), pp. 45–54.

[35] P. VAN DOOREN, *A generalized eigenvalue approach for solving Riccati equations,* SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.

[36] J. C. WILLEMS, *Least-squares stationary optimal control and the algebraic Riccati equation,* IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[37] Y. YAMAMOTO, *New approach to sampled-data control systems-A function space method,* in Proceedings of the 29th Conference on Decision and Control, Honolulu, HI, 1990, pp. 1882–1887.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1151–1169, July 1994

015

THE EQUIVALENCE OF EXTREMALS IN DIFFERENT REPRESENTATIONS OF UNBOUNDED CONTROL PROBLEMS*

J. WARGA† AND Q. J. ZHU†

Abstract. Control problems defined by ordinary differential equations with right-hand sides that are unbounded functions of the control variables are considered. These problems can be reformulated in terms of bounded (relaxed or unrelaxed) differential inclusions by introducing a new independent variable (which is a function of the old state and control functions). These differential inclusions can have different "compact control" representations depending on both the choice of the new independent variable and on the different parametrizations of the set-valued right-hand sides.

The extremals of different (relaxed or unrelaxed) "compact control" representations of such unbounded problems are compared. It is proved that, for a representation that is Lipschitzian in the state variables, the extremals corresponding to different choices of the independent variable are in a one-to-one correspondence, with the corresponding state functions having the same images. If different representations that correspond to different choices of the independent variable and parametrization are compared, then the one-to-one correspondence applies to the sets of "Lojasiewicz extremals" (that is, state functions that remain extremal for every control that generates them) provided the representations are, except for a scalar factor, continuously differentiable in the state variables and satisfy certain "nondegeneracy" conditions. The latter results rely heavily on a theorem of S. Lojasiewicz, Jr., on the equivalence of extremals, which we generalize in certain respects.

Key words. controlled ordinary differential equations, unbounded controls, control representations of differential inclusions, rescaled independent variables, equivalent extremals

AMS subject classification. 49B10

1. Introduction.

A classical problem of the optimal control of autonomous differential equations is the search for the infimum of $h_0(y(1))$ subject to $h_1(y(1)) = 0$ and

$$y(t) = \int_0^t f(y(s), u(s))ds \quad \forall\, t \in [0,1], \tag{1.1}$$

where u is a measurable mapping of [0,1] into a topological space U and $f : \mathbf{R}^n \times U \mapsto \mathbf{R}^n$. We refer to a set $\mathcal{A}$ of solutions (y,u) of (1.1), or to the corresponding set of functions y, as *nearly optimal* if there exist positive numbers c' and c'' such that $\mathcal{A}$ contains all the solutions of (1.1) for which $h_0(y(1)) \leq c'$ and $\mid h_1(y(1)) \mid \leq c''$. If f is continuous, $f(\cdot, r)$ locally Lipschitzian uniformly for $r \in U$, U a compact metric space, (1.1) admits at least one solution for which $h_1(y(1)) = 0$, and some set of nearly optimal solutions y is bounded, then the corresponding relaxed differential inclusion

$$y'(t) \in \overline{co} f(y(t), U) \quad \text{a.e. in } [0,1], \qquad y(0) = 0$$

admits an optimal solution $\bar{y}$ and a *minimizing sequence* $((y_j, u_j))$ such that (y_j, u_j) satisfies (1.1) and

$$\lim_j (h_0, h_1)(y_j(1)) = (h_0, h_1)(\bar{y}(1)) = (h_0(\bar{y}(1)), 0).$$

*Received by the editors July 6, 1992; accepted for publication (in revised form) January 26, 1993.

†Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

If U is not compact and f is unbounded, as is usually the case in problems of the calculus of variations, then the existence of a relaxed solution can sometimes be assured if f satisfies certain "growth conditions." However, in the absence of such growth conditions, it may happen that a minimizing sequence $((y_j, u_j))$ yields functions y_j that converge, in some special sense, to a discontinuous function. Such cases were investigated by Neustadt [N], Rishel [R], and Schmaedeke [S] for certain impulsive controls u, and specifically when f is linear in certain unbounded scalar components of u with uniformly bounded L^1 norms while the other components of u have values confined to a compact set. A more general situation is one when there exists a number L such that, for all (y, u) in some nearly optimal set of solutions, we have

$$\int_0^1 \mid y'(t) \mid dt \leq L$$

and, as a consequence, $y([0,1]) \subset \text{int } V$, where V is the closed ball of center 0 and radius $L + 1$ in $\mathbf{R}^n$. If this is the case, the problem can be approached [W2], [W3, §VI.4] by treating (t, y) as a state function of a new independent variable θ. Using this approach, we select a continuous function $\varphi : V \times U \mapsto (0, \infty)$ such that, for some positive numbers c_1 and c_2,

$$c_1 \leq \varphi(v, r)^{-1}(\mid f(v, r) \mid +1) \leq c_2 \quad \forall\, v \in V, \quad r \in U.$$

We then choose $\theta = \int_0^t \varphi(y(s), u(s))ds$ as the new independent variable, and thus consider the system

$$d\tau/d\theta = \varphi(\eta(\theta), \tilde{u}(\theta))^{-1}, \qquad d\eta/d\theta = \varphi(\eta(\theta), \tilde{u}(\theta))^{-1} f(\eta(\theta), \tilde{u}(\theta)), \tag{1.2}$$

with the initial conditions $\tau(0) = 0, \eta(0) = 0$, on some "free" interval $[0, \alpha]$, where (τ, η) represents (t, y) as a function of $\theta, \tilde{u}(\theta) \in U$, and the condition $t \in [0, 1]$ is equivalent to $\tau(\alpha) = 1$. In system (1.2), the right-hand sides are bounded and the arguments of [W3, §VI.4] provide a way of constructing a minimizing sequence $((y_j, u_j))$ for (1.1) once an optimal relaxed solution of system (1.2) is available. The existence of such an optimal relaxed solution of system (1.2), with α in the interior of some compact interval I depending on L, c_1 and c_2, is assured so long as (1.1) admits at least one solution (y, u) with $h_1(y(1)) = 0$.

Thus system (1.2) and its relaxed version, with endpoint restrictions

$$\tau(\alpha) = 1, \qquad h_1(\eta(\alpha)) = 0, \tag{1.3}$$

become the object of our study. To simplify notation, we reformulate the corresponding control problem (with n, h_0, h_1, f, and V redefined accordingly) as one of searching for the infimum of $h_0(y(\alpha))$ subject to $h_1(y(\alpha)) = 0$ and

$$y(t) = \int_0^t \varphi(y(s), u(s))^{-1} f(y(s), u(s))ds \quad \forall\, t \in [0, \alpha]. \tag{1.4}$$

This formulation encompasses the "fixed time" problem defined by (1.1), in which case the redefined function f has one component equal to 1 (and corresponding to the equation $dt/dt = 1$). However, we may also consider a corresponding "free time" problem, in which the interval of integration is subject to choice (so that we no longer

require $\tau(\alpha) = 1$), and thus we make no assumption about a component of f being equal to 1 except in Theorem 3.1.

To search for a relaxed solution, it is natural to seek a representation of

$$F_\varphi(v) = \text{ closure } \{\varphi(v,r)^{-1} f(v,r) \mid r \in U\} \quad \forall\, v \in V$$

in the "control" form

$$F_\varphi(v) = f^\#(v, \Omega),$$

where Ω is a compact metric space, $f^\#$ is continuous, and $f^\#(\cdot, \omega)$ has a Lipschitz constant independent of ω. If we can find such a representation, then we are dealing with a conventional (possibly nonsmooth) optimal control problem in which the candidates for optimal solutions are found among the relaxed extremals of the problem.

Different choices of functions φ and $f^\#$ may yield different sets of extremals (each of which contains the optimal solutions). It is our purpose to discuss conditions under which extremals corresponding to such different choices are equivalent in the sense that the corresponding state functions transform into one another by a (nonlinear) rescaling of the independent variable. In §2, we consider the (smooth or nonsmooth) equation obtained from (1.4) by a mapping of a dense subset of a compact metric space onto the topological space U. We then show, in Theorem 2.5, that the sets of extremals corresponding to different choices of the function φ are in a one-to-one correspondence as described above. In Theorems 3.1 and 3.3 we consider not only different choices of φ but also different representations of $F_\varphi(v)$ in the control form $f^\#(v, \Omega)$. Our important tool in this endeavor is a (modified form of a) theorem of Lojasiewicz [L2] on the equivalence of "extremals(L)" corresponding to different C^1 representations of the same "nondegenerate" and bounded differential inclusion. This limits us, however, to representations $f^\#(v, \omega)$ that are (except for a scalar factor) C^1 in the first argument, a counterexample of Lojasiewicz [L1, Ex. 2, p. 8] showing that his result is invalid in the general nonsmooth case. Since another counterexample of Lojasiewicz [L2, Ex. 6, p. 251] demonstrates that his theorem is not valid in general for unbounded controls, we can also consider our Theorems 3.1 and 3.3 as variants of Lojasiewicz'es theorem applicable to a class of unbounded control problems that fall outside the scope of his theorem. We also derive Theorem 3.2, a variant of Lojasiewicz's "bounded control" theorem, and in certain respects a generalization of it. In §4, we present some comments and illustrative examples. Finally, in the Appendix, we adapt Lojasiewicz's arguments to prove the modified form of his theorem that is needed for our purposes.

2. A compactification of the parametrized problem. We henceforth make the following assumption.

Assumption 2.1. U is a topological space, $V \subset \mathbf{R}^n$, and $f : V \times U \mapsto \mathbf{R}^n$ continuous.

We first consider a representation of the set $F(v)$ introduced in [W3, p. 375–376]. We denote by $|\cdot|$ the euclidean norm.

DEFINITION 2.2.

2.2.1. *The set* Φ. We denote by Φ the set of all continuous functions $\varphi : V \times U \mapsto (0, \infty)$ such that

(a) the function $v \mapsto \varphi(v,r)^{-1}$ has a Lipschitz constant independent of r, and

(b) there exist positive constants c_φ and d_φ for which

$$c_\varphi \le \varphi(v,r)^{-1}(|\, f(v,r) \mid +1) \le d_\varphi \quad \forall\ v \in V, \quad r \in U.$$

2.2.2. *The set* $\mathcal{F}(\Omega, P, \varphi_0)$. Let $\varphi_0 \in \Phi$, Ω be a compact metric space, Ω', a dense subset of Ω, and $P : \Omega' \mapsto U$ a continuous surjection. We set, for all $v \in V$, $\omega \in \Omega'$ and $\varphi \in \Phi$,

$$f_\varphi(v,\omega) = \varphi(v, P(\omega))^{-1} f(v, P(\omega)), \qquad K_\varphi(v,\omega) = \varphi(v, P(\omega))/\varphi_0(v, P(\omega)).$$

We denote by $\mathcal{F}(\Omega, P, \varphi_0)$ the set of all $\varphi \in \Phi$ such that the function

$$(v,\omega) \mapsto (f_\varphi(v,\omega), K_\varphi(v,\omega)) : V \times \Omega' \mapsto \mathbf{R}^n \times \mathbf{R}$$

has a continuous extension to $V \times \Omega$ that is locally Lipschitzian as a function of v uniformly for $\omega \in \Omega$. It follows that, for every choice of $\varphi, \psi \in \mathcal{F}(\Omega, P, \varphi_0)$, the function

$$(v,\omega) \mapsto K_{\varphi,\psi}(v,\omega) := \varphi(v, P(\omega))/\psi(v, P(\omega)) : V \times \Omega' \mapsto (0,\infty)$$

has a continuous extension to $V \times \Omega$ that is bounded between two positive constants.

DEFINITION 2.3. *Derivate containers and generalized Jacobians.* Let $W \subset \mathbf{R}^p$, $w_0 \in \operatorname{int} W$ and $h : W \mapsto \mathbf{R}^k$ be locally Lipschitzian. We refer to a set $\Lambda h(w_0)$ as *a derivate container* of h at w_0 if there exists a sequence (h_j) of C^1 functions from a neighborhood $\mathcal{N}$ of w_0 to $\mathbf{R}^k$ converging uniformly to h on $\mathcal{N}$ and such that

$$\Lambda h(w_0) = \cap_{\epsilon>0} \text{ closure } \{h_j'(w) \mid \mid w - w_0 \mid < \epsilon,\ j > 1/\epsilon\}.$$

We denote by $\partial h(w_0)$ Clarke's *generalized Jacobian of* h at w_0 [C, p. 69] defined by

$$\partial h(w_0) := \cap_{\epsilon>0} \overline{co}\ \{h'(w) \mid h'(w) \text{ exists}, \mid w - w_0 \mid < \epsilon\}.$$

(It is well known [W6, Thm. 4, p. 549] that $\partial h(w_0)$ is the smallest convex derivate container of h at w_0 but examples can be given [W5, pp. 17–18], [W7, pp. 594–595] of nonconvex derivate containers $\Lambda h(w_0)$ that are proper subsets of $\partial h(w_0)$.) If h depends on an additional variable z, we denote by $\Lambda_1 h(w_0, z)$ or $\Lambda_v h(v, z) \mid_{v=w_0}$ ($\partial_1 h(w_0, z)$ or $\partial_v h(v, z) \mid_{v=w_0}$) a ("partial") derivate container (generalized Jacobian) of $h(\cdot, z)$ at w_0.

DEFINITION 2.4. *Extremals.* Let $W \subset \mathbf{R}^k$, Ω be a compact metric space, $g : W \times \Omega \mapsto \mathbf{R}^k$ continuous, $h_0 : W \mapsto \mathbf{R}$ and $h_1 : W \mapsto \mathbf{R}^m$ locally Lipschitzian, and the function $w \mapsto g(w,\omega)$ locally Lipschitzian uniformly for $\omega \in \Omega$. Let $\operatorname{rpm}(\Omega)$ be the set of Radon probability measures on Ω with the weak star topology of $C(\Omega)^*$, and let $\mathcal{S}(\Omega)$ (the set of relaxed controls corresponding to Ω) be the set of all measurable functions $\sigma : \mathbf{R} \mapsto \operatorname{rpm}(\Omega)$. For $\sigma \in \mathcal{S}(\Omega)$, we write

$$g(w, \sigma(t)) := \int g(w,\omega)\sigma(t)(d\omega),$$

and denote by the superscript ^{T} the transpose of a matrix or a column vector. As is well known, it suffices for control problems defined by (i) below to restrict ourselves to relaxed controls $\sigma \in \mathcal{S}'(\Omega)$, denoted by $[\rho, \theta, k]$, such that, for each t, $\sigma(t)$ is concentrated at k points $\rho_1(t), \ldots, \rho_k(t)$ with masses $\theta_1(t), \ldots, \theta_k(t)$. We henceforth use the term "relaxed control" to mean an element of $\mathcal{S}'(\Omega)$.

With the functions h_0 and h_1 fixed, we refer to a triplet

$$(y, \sigma, \alpha) \in C([0,\alpha], \mathbf{R}^n) \times \mathcal{S}'(\Omega) \times (0,\infty)$$

as a *g-extremal* if (y, σ, α) satisfies

(i) $$y(t) = \int_0^t g(y(s), \sigma(s))ds \quad \forall\, t \in [0, \alpha]$$

and there exist a nonzero vector $(\lambda_0, \lambda_1) \in [0, \infty) \times \mathbf{R}^m$, a derivate container

$$\Lambda(h_0, h_1)(y(\alpha))$$

of (h_0, h_1) at $y(\alpha)$, and an absolutely continuous function $p : [0, \alpha] \mapsto \mathbf{R}^k$ such that

(ii) $$p(\alpha)^T \in (\lambda_0, \lambda_1)^T \Lambda(h_0, h_1)(y(\alpha));$$

(iii) $$p'(t)^T \in -p(t)^T \partial_1 g(y(t), \sigma(t)) \text{ a.e. in } [0, \alpha];$$

(iv) $$H(t) := p(t)^T g(y(t), \sigma(t)) = \min_{\omega \in \Omega} p(t)^T g(y(t), \omega) \text{ a.e. in } [0, \alpha];$$

(v) $$H(t) = 0 \text{ a.e. in } [0, \alpha].$$

(If $v \mapsto g(v, \omega)$ is C^1 and g_v continuous then the constancy of $H(t)$ follows from relations (iii) and (iv) [W1, Thm. 5.1, p. 138] while the vanishing of $H(t)$ then follows from support conditions—a slightly generalized form of relation (ii)—for a problem with variable initial and end conditions (taking account of the fact that the time interval is "free"). In the more general case, however, relation (v) appears to be an independent consequence of optimality [W4, (2.2.4), p. 46], [C, p. 152].)

We refer to a homeomorphism $s : A \mapsto B$ as bi-Lipschitzian if both s and s^{-1} are Lipschitzian.

THEOREM 2.5. *Let $\Phi, \Omega, P,$ and $\mathcal{F}(\Omega, P, \varphi_0)$ be as in Definition 2.2, $\varphi, \psi \in \mathcal{F}(\Omega, P, \varphi_0)$, $\sigma = [\rho, \theta, k]$, and (y, σ, α) an f_φ-extremal corresponding to a choice of $\lambda_0, \lambda_1, \Lambda(h_0, h_1)(y(\alpha))$ and p. Then there exist a strictly increasing bi-Lipschitzian homeomorphism*

$$\tau \mapsto s(\varphi, \psi, y, \sigma, \alpha)(\tau) = s(\tau)$$

of $[0, s^{-1}(\alpha)]$ onto $[0, \alpha]$ and $\kappa = (\kappa_1, \ldots, \kappa_k)$ such that, for

$$\tilde{y} = y \circ s, \ \tilde{\sigma} = [\rho \circ s, \kappa, k], \ \tilde{\alpha} = s^{-1}(\alpha),$$

$(\tilde{y}, \tilde{\sigma}, \tilde{\alpha})$ is an f_ψ-extremal corresponding to the choice of

$$\lambda_0, \lambda_1, \Lambda(h_0, h_1)(\tilde{y}(\tilde{\alpha}))(= \Lambda(h_0, h_1)(y(\alpha))) \text{ and } \tilde{p} = p \circ s.$$

Furthermore, $s(\psi, \varphi, \tilde{y}, \tilde{\sigma}, \tilde{\alpha})$ is the inverse of $s(\varphi, \psi, y, \sigma, \alpha)$ and the inverse correspondence yields (y, σ, α) as corresponding to $(\tilde{y}, \tilde{\sigma}, \tilde{\alpha})$. Thus there exists a one-to-one correspondence between f_φ-extremals and f_ψ-extremals defined by a rescaling of the independent variable in y and ρ_j and a modification of the weights associated with ρ_j.

Proof. We observe that the functions f_φ, f_ψ and $K = K_{\psi,\varphi}$ satisfy the relation

$$f_\varphi(v, \omega) = K(v, \omega) f_\psi(v, \omega) \quad \forall\, v \in V, \quad \omega \in \Omega.$$

Now let

$$\nu(t) = 1/K(y(t), \sigma(t)) = 1/\Sigma_{j=1}^{k} \theta_j(t) K(y(t), \rho_j(t)) \quad \forall\, t \in [0, \alpha],$$

and let $\gamma(t) = \int_0^t \nu(\tau) d\tau$. Then γ is a strictly increasing Lipschitzian function on $[0, \alpha]$ and has a Lipschitzian inverse $s = \gamma^{-1}$. We observe that

$$s'(\gamma(t))\gamma'(t) = s'(\gamma(t))\nu(t) = 1 \text{ a.e. in } [0, \alpha];$$

hence $s'(\beta) = 1/\nu(s(\beta))$ almost everywhere in $[0, \gamma(\alpha)]$. Let

$$\kappa_j(\beta) = \nu(s(\beta))^{-1} \theta_j(s(\beta)) K(y(s(\beta)), \rho_j(s(\beta))) \quad \forall\, j = 1, \ldots, k.$$

Then

$$\begin{aligned}
\tilde{y}'(\beta) &= y'(s(\beta))/\nu(s(\beta)) = \nu(s(\beta))^{-1} \Sigma_{j=1}^{k} \theta_j(s(\beta)) f_\varphi(y(s(\beta)), \rho_j(s(\beta))) \\
&= \nu(s(\beta))^{-1} \Sigma_{j=1}^{k} \theta_j(s(\beta)) K(y(s(\beta)), \rho_j(s(\beta))) f_\psi(y(s(\beta)), \rho_j(s(\beta))) \\
&= \Sigma_{j=1}^{k} \kappa_j(\beta) f_\psi(y(s(\beta)), \rho_j(s(\beta))) = f_\psi(\tilde{y}(\beta), \tilde{\sigma}(\beta)) \text{ a.e. in } [0, \tilde{\alpha}].
\end{aligned}$$

Thus $(\tilde{y}, \tilde{\sigma}, \tilde{\alpha})$ satisfies (i) of Definition 2.4 for $g = f_\psi$.

Now let

$$\tilde{p}(\beta) = p(s(\beta)), \ \tilde{\rho}(\beta) = \rho(s(\beta)) \quad \forall\, \beta \in [0, \tilde{\alpha}].$$

Then

$$\tilde{p}(\tilde{\alpha})^T = p(\alpha)^T \in (\lambda_0, \lambda_1)^T \partial(h_0, h_1)(y(\alpha)) = (\lambda_0, \lambda_1)^T \partial(h_0, h_1)(\tilde{y}(\tilde{\alpha}))$$

and, for almost all $\beta \in [0, \tilde{\alpha}]$ and all $j = 1, \ldots, k$, we have

$$H(s(\beta)) = \tilde{p}(\beta)^T f_\varphi(\tilde{y}(\beta), \tilde{\rho}_j(\beta)) = \min_{\omega \in \Omega} \tilde{p}(\beta)^T f_\varphi(\tilde{y}(\beta), \omega) = 0;$$

hence

$$\tilde{H}(\beta) := \tilde{p}(\beta)^T f_\psi(\tilde{y}(\beta), \tilde{\rho}_j(\beta)) = \min_{\omega \in \Omega} \tilde{p}(\beta)^T f_\psi(\tilde{y}(\beta), \omega) = 0.$$

Finally, we observe that if, for $i = 1, \ldots, k$, M_i and N_i are real-valued and Lipschitzian near w and $M_i(w) = 0$ then

$$\partial(\Sigma_{i=1}^{k} M_i N_i)(w) = \partial_v [\Sigma_{i=1}^{k} N_i(w) M_i(v)]_{v=w}.$$

It follows, using the relations

$$f_\varphi = K f_\psi \text{ and } H(t) = p(t)^T (K f_\psi)(y(t), \rho_j(t)) = 0 \text{ a.e. in } [0, \alpha],$$

that

$$\begin{aligned}
&p(t)^T \partial_1 [\Sigma_{j=1}^{k} \theta_j(t) (K f_\psi)(y(t), \rho_j(t))] \\
&\quad = \partial_1 [\Sigma_{j=1}^{k} \theta_j(t) (K p(t)^T f_\psi)(y(t), \rho_j(t))] \\
&\quad = \partial_v [\Sigma_{j=1}^{k} \theta_j(t) K(y(t), \rho_j(t)) p(t)^T f_\psi(v, \rho_j(t))]_{v=y(t)} \\
&\quad = \partial_1 [\Sigma_{j=1}^{k} \nu(t) \kappa_j(s^{-1}(t)) p(t)^T f_\psi(v, \rho_j(t))]_{v=y(t)} \text{ a.e. in } [0, \alpha];
\end{aligned}$$

hence

$$\begin{aligned}\tilde{p}'(\beta)^T &= p'(s(\beta))^T s'(\beta) = \nu \circ s(\beta)^{-1} p'(s(\beta))^T \\ &\in -\nu \circ s(\beta)^{-1} p(s(\beta))^T \partial_1 (K f_\psi)(y \circ s(\beta), \sigma \circ s(\beta)) \\ &= -\nu \circ s(\beta)^{-1} p(s(\beta))^T \partial_1 [\Sigma_{j=1}^k \theta_j(s(\beta))(K f_\psi)(y(s(\beta)), \rho_j(s(\beta)))] \\ &= -\partial_1 [\Sigma_{j=1}^k \kappa_j(\beta) p(s(\beta))^T f_\psi(y(s(\beta)), \rho_j(s(\beta)))] \\ &= -\tilde{p}(\beta)^T \partial_1 f_\psi(\tilde{y}(\beta), \tilde{\sigma}(\beta)) \text{ a.e. in } [0, \tilde{\alpha}].\end{aligned}$$

This completes the proof that $(\tilde{y}, \tilde{\sigma}, \tilde{\alpha})$ is an f_ψ-extremal.

Now let us apply the same argument, with $\varphi, y, \sigma, \alpha, K_{\psi,\varphi}$ and $\psi, \tilde{y}, \tilde{\sigma}, \tilde{\alpha}, K_{\varphi,\psi}$ interchanged. Then $K_{\varphi,\psi} = 1/K$ and, setting

$$\tilde{\gamma}(\beta) = \int_0^\beta K(\tilde{y}(b), \tilde{\sigma}(b)) db, \ \beta = \gamma(t) = s^{-1}(t),$$

we obtain

$$\tilde{\gamma}'(\gamma(t)) = K(y(t), \sigma(t)) = 1/\gamma(t) \ \text{ a.e.;}$$

hence $\tilde{\gamma}'(\gamma(t)) = t$ almost everywhere and $\tilde{\gamma} = s$. We can also verify that the control corresponding to $\tilde{\sigma}$ is σ. This shows that the correspondence of the f_φ-extremals and the f_ψ-extremals is one-to-one, with $s(\psi, \varphi, \tilde{y}, \tilde{\sigma}, \tilde{\alpha})$ the inverse of $s(\varphi, \psi, y, \sigma, \alpha)$. □

3. Equivalence of Lojasiewicz extremals of different representations. For the sake of simplicity, we refer to a function $f\# : V \times \Omega \mapsto \mathbf{R}^k$ as C^1 in v if Ω is a compact metric space, $f\#$ continuous, $f\#(\cdot, \omega)$ a C^1 function for all $\omega \in \Omega$, and $D_1 f\#$ (the partial derivative with respect to the first argument of $f\#$) continuous. We say that $f\#$ is a C^1 representation of sets $G(v)$ for $v \in V$ if $f\#$ is C^1 in v and $f\#(v, \Omega) = G(v)$ for $v \in V$. We denote by $\mathrm{Arc}_{\mathrm{rel}}(g, \Omega)$ (respectively, $\mathrm{Arc}_{\mathrm{unrel}}(g, \Omega)$) the collection of all (y, α) such that, for some relaxed (respectively, unrelaxed) control σ,

$$y(t) = \int_0^t g(y(s), \sigma(s)) ds \ \ \forall\, t \in [0, \alpha]$$

g-extremal(L). Let $W \subset \mathbf{R}^k$, Ω be a compact metric space, $g : W \times \Omega \mapsto \mathbf{R}^k$ continuous and the function $w \mapsto g(w, \omega)$ locally Lipschitzian uniformly for $\omega \in \Omega$. We refer to a couple (y, α) as a relaxed (respectively, unrelaxed) *g-extremal* (L) (for a *Lojasiewicz extremal of g*) if $y \in \mathrm{Arc}_{\mathrm{rel}}(g, \Omega)$ (respectively, $\mathrm{Arc}_{\mathrm{unrel}}(g, \Omega)$) and, for every relaxed (respectively, unrelaxed) control σ that satisfies the equation

$$y(t) = \int_0^t g(y(s), \sigma(s)) ds \quad \forall\, t \in [0, \alpha],$$

there exists a nonzero absolutely continuous function $p : [0, \alpha] \mapsto \mathbf{R}^n$ such that

$$p'(t)^T \in -p(t)^T \partial_1 g(y(t), \sigma(t)) \text{ a.e. in } [0, \alpha];$$

$$H(t) := p(t)^T g(y(t), \sigma(t)) = \min_{\omega \in \Omega} p(t)^T g(y(t), \omega) \text{ a.e. in } [0, \alpha];$$

$$H(t) = 0 \text{ a.e. in } [0, \alpha].$$

(This is a necessary condition for $y(\alpha)$ to belong to the boundary of the attainable set $\{y(t) \mid (y,\alpha) \in Arc(g,\Omega), t \leq \alpha\}$, where Arc = Arc_{rel}, respectively, Arc = $\text{Arc}_{\text{unrel}}$.)

We apply the following modified form of a result of Lojasiewicz [L2, Thm. 3, p. 253].[1]

LOJASIEWICZ'S THEOREM. *Let g_1 and g_2 be C^1 representations of sets $G(v)$ such that, for all $v \in V$, the convex hulls of $G(v)$ have nonempty interiors. Then the (relaxed respectively unrelaxed) g_1-extremals(L) and g_2-extremals(L) coincide.*

DEFINITION. *The set X.* Let B be the open (euclidean) unit ball in $\mathbf{R}^n$ and $Z = \mathbf{R}^n \backslash \mathrm{B}$. We denote by X' the set of all C^1 functions $\chi : V \times Z \mapsto (0,\infty)$ such that each is bounded between two positive constants and $\lim_{|z|\to\infty} \chi(v,z)$ exists uniformly for all $v \in$ V. We denote by $[Z]$ the one-point compactification of Z (i.e., $Z \cup \{\infty\}$) and extend χ to $V \times [Z]$ as a continuous function by setting $\chi(v,\infty) = \lim_{|z|\to\infty} \chi(v,z)$. We denote by X the set of all such extensions of $\chi \in X'$.

THEOREM 3.1. *Let f have one component equal to* 1, $\varphi_\chi(v,r) = \chi(v, f(v,r)) \mid f(v,r) \mid$, *and Ξ be the set of all $\xi = (\chi, g, \Gamma)$ such that $\chi \in X$, Γ is a compact metric space, $g : V \times \Gamma \mapsto \mathbf{R}^n$ continuous, $g(\cdot,\gamma)$ locally Lipschitzian uniformly for $\gamma \in \Gamma$, the function $(v,\gamma) \mapsto \mid g(v,\gamma) \mid^{-1} g(v,\gamma)$ C^1 in v, and*

$$g(v,\Gamma) = \text{closure}\{\varphi_\chi(v,r)^{-1} f(v,r) \mid r \in U\} \quad \forall\, v \in V.$$

Assume that either

(a) *there exists some $\chi_0 \in X$ such that*

$$\text{int co}\, \{\varphi_{\chi_0}(v,r)^{-1} f(v,r) \mid r \in U\} \neq \emptyset \quad \forall\, v \in V,$$

or

(b) *V is compact, $f(v,U)$ is infinite, and $f(v,U)$ spans $\mathbf{R}^n$ for all $v \in V$.*

Then, for every choice of $\xi' = (\chi, g, \Gamma)$ and $\xi'' = (\psi, h, \Delta)$ in Ξ and of a relaxed (respectively, unrelaxed) g-extremal(L) (y,α), there exists a strictly increasing bi-Lipschitzian homeomorphism $s := s(\xi', \xi'', y, \alpha) : [0, s^{-1}(\alpha)] \mapsto [0,\alpha]$ such that $(\tilde{y}, \tilde{\alpha}) = (y \circ s, s^{-1}(\alpha))$ is a relaxed (respectively, unrelaxed) h-extremal(L), and this correspondence is one-to-one. If $\chi = \psi$ then every g-extremal(L) is also an h-extremal(L).

Theorem 3.2 below applies to bounded problems only and represents a variant of Lojasiewicz's Theorem. For the sake of greater simplicity, we have replaced the pertinent (and somewhat complicated) assumptions of Lojasiewicz's Theorem by the stronger assumption that the control variables have compact ranges. However, our arguments are independent of these assumptions except when they refer to a modified form of Lojasiewicz's Theorem. Thus Theorem 3.2 remains valid when the compactness assumption is replaced by the pertinent assumptions of Lojasiewicz's Theorem. On the other hand, Theorem 3.2 applies to some nonsmooth problems and somewhat weakens the assumption that int co $f(v,U) \neq \emptyset$.

THEOREM 3.2. *Let $V \subset \mathbf{R}^n$, U and U_1 be compact metric spaces, $f : V \times U \mapsto \mathbf{R}^n$ and $g : V \times U_1 \mapsto \mathbf{R}^n$ bounded and continuous, $f(v,U) = g(v,U_1)$ for all $v \in V$, and $f(\cdot,r)$, respectively, $g(\cdot,r_1)$ locally Lipschitzian uniformly for $r \in U$, respectively, $r_1 \in U_1$. Assume that either*

---

[1] Lojasiewicz has pointed out to us the following correction due to a typographical mistake in the proof of Theorem 3 of [L2]: on p. 254, line 6*, on the right-hand side of the equality sign, y should be replaced by $M(t)y$.

(a) *there exist* $c_2 > c_1 > 0$ *and a locally Lipschitzian function* $\chi : V \times \mathbf{R}^n \mapsto [c_1, c_2]$ *such that the functions*

$$f^{\#}(v,r) = \chi(v, f(v,r))^{-1} f(v,r), \qquad g^{\#}(v,r_1) = \chi(v, g(v,r_1))^{-1} g(v,r_1),$$

are C^1 *in* v *and* int co $f^{\#}(v,U) \neq \emptyset$ *for all* $v \in V$, *or*

(b) f *and* g *are* C^1 *in* v, V *is compact and, for all* $v \in V$, $f(v,U)$ *is infinite and* $f(v,U)$ *spans* $\mathbf{R}^n$.

Then every relaxed (respectively, unrelaxed) f*-extremal*(L) (y,α) *is also a relaxed (respectively, unrelaxed)* g*-extremal*(L).

Theorem 3.1 applies to functions f with one component equal to 1 and the functions $\varphi(v,r)$ must be chosen as functions of v and $f(v,r)$. We can dispense with these conditions if, for some fixed $\varphi_0(v,r)$, both $\varphi(v,r)^{-1} f(v,r)$ and $\varphi_0(v,r)^{-1} f(v,r)$ can be expressed in terms of the new control. The price that we must pay for this somewhat greater freedom is a weakening of our conclusions. The new Theorem 3.3 applies to the more restricted class of *extremals*$(L+)$ ("*strengthened Lojasiewicz extremals*").

g-extremal$(L+)$. We refer to a couple (y,α) as a relaxed (respectively, unrelaxed) g-extremal$(L+)$ if $(y \circ s, \tilde{\alpha})$ is a relaxed (respectively, unrelaxed) g-extremal(L) for every bi-Lipschitzian homeomorphism $s : [0,\tilde{\alpha}] \mapsto [0,\alpha]$ for which $(y \circ s, \tilde{\alpha}) \in \mathrm{Arc}_{\mathrm{rel}}(g,\Omega)$ (respectively, $(y \circ s, \tilde{\alpha}) \in \mathrm{Arc}_{\mathrm{unrel}}(g,\Omega)$). This definition describes a property that clearly characterizes every solution with its endpoint on the boundary of the relaxed (respectively, unrelaxed) attainable set.

THEOREM 3.3. *Let* Φ *be as in Definition* 2.2.1 *and* $\varphi_0 \in \Phi$ *such that*

$$\text{int co}\{\varphi_0(v,r)^{-1} f(v,r) \mid r \in U\} \neq \emptyset \quad \forall\, v \in V.$$

Let Ξ *be the set of all* $\xi = (\varphi, g, K, \Gamma)$ *such that* $\varphi \in \Phi$, Γ *is a compact metric space, the functions* $g : V \times \Gamma \mapsto \mathbf{R}^n$ *and* $K : V \times \Gamma \mapsto (0,\infty)$ *are* C^1 *in* v, *and*

$$g(v,\Gamma) = \text{closure}\{\varphi(v,r)^{-1} f(v,r) \mid r \in U\},$$

$$\{K(v,\gamma) g(v,\gamma) \mid \gamma \in \Gamma\} = \text{closure}\{\varphi_0(v,r)^{-1} f(v,r) \mid r \in U\}.$$

Then, for every choice of $\xi' = (\varphi, g, K, \Gamma)$ *and* $\xi'' = (\psi, h, L, \Delta)$ *in* Ξ *and of a relaxed (respectively, unrelaxed)* g*-extremal*$(L+)$ (y,α), *there exists a strictly increasing bi-Lipschitzian homeomorphism* $s := s(\xi', \xi'', y, \alpha) : [0, s^{-1}(\alpha)] \mapsto [0,\alpha]$ *such that* $(\tilde{y}, \tilde{\alpha}) = (y \circ s, s^{-1}(\alpha))$ *is a relaxed (respectively, unrelaxed)* h*-extremal*$(L+)$, *and this correspondence is one-to-one.*

In the proofs of Theorems 3.1–3.3 we use the term "g-extremal" as in Definition 2.4 but with the relation (ii) replaced by $p(\alpha) \neq 0$.

Proof of Theorem 3.1.

Step 1. First assume that condition (a) is satisfied. Since, by assumption, f has one component equal to 1, we can write it in the form $f = (1, f_2)$. Thus, if we set $f(v,U) = F(v)$ then $z = (1, z_2) \in Z = \mathbf{R}^n \backslash B$ for every $z \in F(v)$. Similarly, we write $g = (g_1, g_2)$. For $z = (z_1, z_2) \in Z$, let

$$p(z) = (1, z_2/z_1) \; if \; z_1 \neq 0, \quad p(z) = \infty \; if \; z_1 = 0,$$

and, for all $v \in V$ and $\gamma \in \Gamma$, let

$$K(v,\gamma) = \chi_0(v, p(g(v,\gamma)))/\chi(v, p(g(v,\gamma))), \qquad g^{\#}(v,\gamma) = K(v,\gamma)^{-1} g(v,\gamma).$$

We observe that

$$F_\chi(v) := \{\chi(v, f(v,r))^{-1} \mid f(v,r) \mid^{-1} f(v,r) \mid r \in U\}$$
$$= \{\chi(v,z)^{-1} \mid z \mid^{-1} z \mid z \in F(v)\}$$

and $g(v,\Gamma) = \overline{F_\chi(v)}$.

To each $\zeta \in F_\chi(v)$ there correspond some $z = (1, z_2) \in F(v)$, r in U and γ in some subset $\Gamma'(v)$ of Γ such that

$$\zeta = \chi(v,z)^{-1} \mid z \mid^{-1} z = g(v,\gamma) = (g_1, g_2)(v,\gamma), \qquad z = f(v,r),$$

and similarly for each element of $\{z, r, \zeta, \gamma\}$ there are other elements of this quadruplet satisfying the above relations. Thus $g_1(v,\gamma) \neq 0$ and $z_2 = g_2(v,\gamma)/g_1(v,\gamma)$. It follows that

$$g^\#(v,\gamma) = \frac{\chi(v,z)}{\chi_0(v,z)} \frac{1}{\chi(v,z) \mid z \mid} z = \varphi_{\chi_0}(v,r)^{-1} f(v,r)$$

and

$$\text{closure } g^\#(v,\Gamma'(v)) = \text{closure}\{\varphi_{\chi_0}(v,r)^{-1} f(v,r) \mid r \in U\} = \overline{F_{\chi_0}(v)}.$$

Now let $v \in V$ and $\gamma \in \Gamma$. Then there exist a sequence (r_i) in U and corresponding sequences (γ_i) and (z_i) such that

$$g(v,\gamma) = \lim_i g(v,\gamma_i) = \lim_i \chi(v, f(v,r_i))^{-1} \mid f(v,r_i) \mid^{-1} f(v,r_i).$$

Since $[Z]$ is metrizable and compact and both χ and χ_0 are continuous on $V \times [Z]$ and bounded away from 0, we may replace (r_i) by an appropriate subsequence so that the points $z_i = f(v,r_i)$ converge to a limit in $[Z]$ and

$$g^\#(v,\gamma) = K(v,\gamma)^{-1} g(v,\gamma) = \lim_i \chi_0(v, p(f(v,r_i)))^{-1} \mid f(v,r_i) \mid^{-1} f(v,r_i)$$
$$= \lim_i \chi_0(v,z_i)^{-1} \mid z_i \mid^{-1} z_i \in \text{closure } g^\#(v,\Gamma'(v)) = \overline{F_{\chi_0}(v)};$$

hence $g^\#(v,\Gamma) = \overline{F_{\chi_0}(v)}$.

The proof of Theorem 2.5, with f_φ, f_ψ replaced by $g, g^\#$, applies without change and we conclude that for every g-extremal (y,α) there exists a bi-Lipschitzian homeomorphism $s(\xi', y, \alpha) = s_1$ such that the couple $(y^\#, \alpha^\#) := (y \circ s_1, s_1^{-1}(\alpha))$ is a $g^\#$-extremal and conversely, with s_1 replaced by its inverse. We observe, furthermore, that the homeomorphism s_1, as defined in the proof of Theorem 2.5, is uniquely determined by χ and y. Indeed, s_1 is defined by $\int_0^t \nu(\tau)d\tau$, where $\nu(t) = 1/K(y(t), \sigma(t))$, and we have

$$K(y(t),\sigma(t)) = \chi_0(y(t), g(y(t),\sigma(t)))/\chi(y(t), g(y(t),\sigma(t)))$$
$$= \chi_0(y(t), y'(t))/\chi(y(t), y'(t)) \text{ a.e. in } [0,\alpha].$$

The same argument, applied to ξ'', yields $h^\# : V \times \Delta \mapsto \mathbf{R}^n$ such that

$$g^\#(v,\Gamma) = h^\#(v,\Delta) = \text{closure}\{\chi_0(v,z)^{-1} \mid z \mid^{-1} z \mid z \in F(v)\} = \overline{F_{\chi_0}(v)}.$$

Thus, by assumption (a), int co $g^\#(v,\Gamma) \neq \emptyset$ for all $v \in$ V. Now let $\xi''' = (1/\psi, h^\#, \Delta)$. Since the homeomorphisms $s(\cdot)$ are uniquely determined by the state functions and the functions χ, it follows, by Lojasiewicz's Theorem, that to every relaxed (respectively,

unrelaxed) g-extremal(L) (y, α) there corresponds a unique function $z = y \circ s$, with $s = s(\xi', y, \alpha) \circ s(\xi''', y^{\#}, \alpha^{\#})$, such that $(z, s^{-1}(\alpha))$ is a relaxed (respectively, unrelaxed) h-extremal(L). Furthermore, if $\chi = \psi$ then

$$s(\xi''', y^{\#}, \alpha^{\#}) = s(\xi', y, \alpha)^{-1},$$

and therefore $z = y$ and (y, α) is also an h-extremal(L).

Step 2. Now assume that condition (b) is satisfied. We show that condition (a) follows, thus completing the proof of the theorem.

Let $f^*(v, r) = \mid f(v, r) \mid^{-1} f(v, r)$. Then, for each $v \in V$, there exist $r_v^0, \ldots, r_v^n \in U$ such that the set $H(v) = \{f^*(v, r_v^0), \ldots, f^*(v, r_v^n)\}$ spans $\mathbf{R}^n$ and, if int co $f^*(v, U) \neq \emptyset$, then also int co $H(v) \neq \emptyset$. It follows that each $v \in V$ has a neighborhood $\mathcal{N}(v)$ in V such that, for all $w \in \mathcal{N}(v), \mathcal{H}(w, v) = \{f^*(w, r_v^0), \ldots, f^*(w, r_v^n)\}$ spans $\mathbf{R}^n$ and, if int co $f^*(w, U) \neq \emptyset$, then also int co $\mathcal{H}(w, v) \neq \emptyset$. The compact set V can be covered by a finite collection of such neighborhoods, say by $\mathcal{N}(v_1), \ldots, \mathcal{N}(v_k)$. Let the set

$$\{r_{v_j}^i \mid i = 0, \ldots, n; j = 1, \ldots, k\}$$

be enumerated as $\{\rho_1, \ldots, \rho_s\}$. Then, for every choice of $v \in V$, the set

$$\mathcal{K}(v) = \{f^*(v, \rho_1), \ldots, f^*(v, \rho_s)\}$$

spans $\mathbf{R}^n$ and, if int co $f^*(v, U) \neq \emptyset$, then also int co $\mathcal{K}(v) \neq \emptyset$.

Let $B(z, a) (\bar{B}(z, a))$ denote the open (closed) ball of center z and radius a in $\mathbf{R}^n$. For all $v \in V$ and $i = 1, \ldots, s$, let $z_i(v) = f(v, \rho_i)$. We first show that there exists $\beta_0 \in (0, 1]$ such that

$$f(v, U) \setminus \bigcup_{i=1}^{s} B(z_i(v), \beta_0) \neq \emptyset \quad \forall\, v \in V.$$

Indeed, otherwise there exists a sequence (v_n) in the compact set V converging to some w and such that

$$(*) \qquad f(v_n, U) \subset \bigcup_{i=1}^{s} B(z_i(v_n), 1/n).$$

Since $f(w, U)$ is infinite, there exist $r_w \in U$ and $\alpha > 0$ such that

$$\mid f(w, r_w) - z_i(w) \mid > \alpha \quad \forall\, i = 1, \ldots, s.$$

It follows, by the continuity of f and z_i, that

$$\mid f(v_n, r_w) - z_i(v_n) \mid > \alpha/2 \quad \forall\, i = 1, \ldots, s$$

for sufficiently large n, thus contradicting (*).

Now let

$$A(v) = \bigcup_{i=1}^{s} \bar{B}(z_i(v), \beta_0/2), \qquad B(v) = \mathbf{R}^n \setminus \bigcup_{i=1}^{s} B(z_i(v), \beta_0).$$

Then $A(v)$ and $B(v)$ are closed, the set-valued mappings $v \mapsto A(v)$ and $v \mapsto B(v)$ continuous (in the Hausdorff metric), and

$$A(v) \cap B(v) = \emptyset, \quad F(v) \cap A(v) \neq \emptyset, \quad F(v) \cap B(v) \neq \emptyset.$$

Therefore the graphs of $A(\cdot)$ and $B(\cdot)$ are closed and disjoint in $V \times \mathbf{R}^n$ and there exists a continuous function $\chi^{\#} : V \times \mathbf{R}^n \mapsto [1,2]$ such that $\chi^{\#}(v, A(v)) = 1$ and $\chi^{\#}(v, B(v)) = 2$ for all $v \in \mathrm{V}$.

If int co $f^*(v,U) = \emptyset$ then there exist $\lambda(v) \in \mathbf{R}^n \backslash \{0\}$ and $c(v) \in \mathbf{R} \backslash \{0\}$ such that $\lambda(v)^T f^*(v,U) = \{c(v)\}$ and $\lambda(v)$ is unique up to a constant nonzero multiplier (because $f^*(v,U)$ spans $\mathbf{R}^n$). Thus $2\lambda(v)^T f^*(v,r) = 2c(v) \neq c(v)$ for all $r \in \mathrm{U}$. It follows that

$$(**) \qquad \text{int co}\{\chi^{\#}(v,z)^{-1} \mid z \mid^{-1} z \mid z \in F(v)\} \neq \emptyset\} \quad \forall\, v \in V.$$

Since the set $\bigcup_{v \in V} \bigcup_{i=1}^{s} B(z_i(v), \beta_0)$ is bounded and $\chi^{\#}(v,z) = 2$ outside that set, we can approximate $\chi^{\#}$ uniformly and arbitrarily closely with a C^1 function converging to 2 as $\mid z \mid \to \infty$, and therefore relation (**) will remain valid if $\chi^{\#}$ is replaced by an appropriate function $\chi_0 \in \mathrm{X}$. □

Proof of Theorem 3.2. First assume that condition (a) is satisfied. Let

$$\varphi(v,r) = \chi(v, f(v,r)), L(v,\gamma) = \chi(v, g(v,\gamma)), \qquad g^{\#}(v,\gamma) = L(v,\gamma)^{-1} g(v,\gamma).$$

Then

$$g^{\#}(v, U_1) = \{\varphi(v,r)^{-1} f(v,r) \mid r \in U\}.$$

The proof of Theorem 2.5, with f_φ, f_ψ, K replaced by $g, g^{\#}, L$ applies without change and we conclude that, for every g-extremal (y, σ, α), there exists a bi-Lipschitzian homeomorphism $s(y,\sigma,\alpha) = s$ such that the triplet $(\tilde{y}, \tilde{\sigma}, \tilde{\alpha}) := (y \circ s, \sigma \circ s, s^{-1}(\alpha))$ is a $g^{\#}$-extremal and conversely, with s replaced by its inverse. We observe, furthermore, that the homeomorphism s, as defined in the proof of Theorem 2.5, is uniquely determined by χ and y. Indeed, s is defined by $\int_0^t \nu(\tau) d\tau$, where $\nu(t) = 1/L(y(t), \sigma(t))$, and we have

$$L(y(t), \sigma(t)) = \chi(y(t), g(y(t), \sigma(t))) = \chi(y(t), y'(t)) \quad \text{a.e. in } [0, \alpha].$$

This shows that if (y, α) is a g-extremal(L) then $(\tilde{y}, \tilde{\alpha})$ is a $g^{\#}$-extremal(L), and conversely.

The same argument, applied to f, yields $f^{\#} : V \times U \mapsto \mathbf{R}^n$ such that

$$g^{\#}(v, U_1) = f^{\#}(v, U) = \{\varphi(v,r)^{-1} f(v,r) \mid r \in U\}$$

and

$$f^{\#}(v,r) = \chi(v, f(v,r))^{-1} f(v,r) \quad \forall\, r \in U, v \in V.$$

Thus, by assumption (a), int co $g^{\#}(v, U_1) \neq \emptyset$ for all $v \in \mathrm{V}$. It follows, by Lojasiewicz's Theorem, that every relaxed (respectively, unrelaxed) $g^{\#}$-extremal(L) $(\tilde{y}, \tilde{\alpha})$ is also an $f^{\#}$-extremal(L). Now let $\tilde{\sigma}_1$ be a control such that

$$\tilde{y}'(\beta) = \int_0^{\beta} f^{\#}(\tilde{y}(b), \tilde{\sigma}_1(b)) db \quad \forall\, \beta \in [0, \tilde{\alpha}].$$

Then our previous argument shows that the $f^{\#}$-extremal $(\tilde{y}, \tilde{\sigma}_1, \tilde{\alpha})$ corresponds to an f-extremal (y, σ_1, α), where (y, α) is the same g-extremal(L) that corresponds to the $g^{\#}$-extremal(L) $(\tilde{y}, \tilde{\alpha})$. This shows that every g-extremal(L) is also an f-extremal(L).

Since we have shown in Step 2 of the proof of Theorem 3.1 that assumption (b) implies assumption (a), our proof is complete. □

Proof of Theorem 3.3. By the definition of Φ (Definition 2.2.1), the function K is bounded between two positive constants. Let

$$g^{\#}(v, \gamma) = K(v, \gamma) g(v, \gamma) \quad \forall\, v \in V, \gamma \in \Gamma.$$

The proof then proceeds essentially as in the last two paragraphs of Step 1 of the proof of Theorem 3.1. However, since the homeomorphism $s(\cdot)$ may now depend on both y and the control σ, we must restrict ourselves to extremals($L+$). □

4. Comments and examples. In the modified version of Lojasiewicz's Theorem in §3, we assert that if the (relaxed or unrelaxed) g_1-extremal(L) is a "free time" extremal, i.e., the corresponding Hamiltonian $H(t) = 0$ almost everywhere then the same is the case for the g_2-extremal(L). The proof of this assertion, in the Appendix, is an adaptation of Lojasiewicz's original proof.

Another pertinent theorem of Lojasiewicz [L2, Thm. 2, p. 246] asserts that the assumption about co $G(v)$ can be replaced by the assumption that the function

$$(v, p) \mapsto \mathcal{H}(v, p) := \min_{r \in U} p^T f(v, r) : V \times \mathbf{R}^n \mapsto \mathbf{R}$$

is differentiable at $(y(t), q)$ for almost all $t \in [0, \alpha]$ and for all q such that $\mathcal{H}(y(t), q) = q^T y'(t)$, where (y, α) is the g_1-extremal(L).

We should mention that there are several counterexamples showing that Lojasiewicz's Theorem is not valid for ordinary extremals as distinct from extremals(L) (e.g., if the control problem admits a state function $y(\cdot)$ that is generated by two controls, one of which is extremal and the other nonextremal). However, we know of no counterexamples showing that both of the above conditions (about co $G(v)$ or the differentiability of $\mathcal{H}$) cannot be dispensed with. While our Theorem 3.2 shows that these "nondegeneracy" assumptions as well as the C^1 requirement may be somewhat weakened, it is still an open question whether Lojasiewicz's Theorem remains valid for C^1 problems without any assumptions concerning the convex hull of $G(v)$.

Our Theorem 3.3 applies to extremals($L+$), a more restricted class than the extremals(L) that appear in Lojasiewicz's Theorem and in Theorems 3.1 and 3.2. The following example of Lojasiewicz [L3] shows that, even for simple problems, an extremal(L) need not be an extremal($L+$). Consider the control problem defined by

$$y_1' = y_2, \quad y_2' = -y_1 + u, \quad \text{with } u(t) \in [0, 1].$$

Then $(y, \alpha) = (0, \alpha)$ is an extremal(L) for $\alpha \leq \pi$ but it is not extremal (in the Pontryagin sense) for $\alpha > \pi$. However, while our proof only enables us to draw conclusions concerning extremals($L+$), we are unable at present to answer the question whether Theorem 3.3 is valid for extremals(L).

Our first example describes a class of simple problems to which Theorem 2.5 applies.

Example 4.1. Consider the following control problem with linear controls:

$$y'(t) = g(y(t))u(t), \qquad u(t) \in U,$$

where $g(v) = (g_{ij}(v))$ is a $n \times m$ matrix with locally Lipschitz continuous elements g_{ij} and U is an arbitrary subset of $\mathbf{R}^m$. Let S^m be the boundary of the unit ball in $\mathbf{R}^{m+1}$, and let $Q : R^m \mapsto S^m$ be defined by

$$Q(r) = (1+ \mid r \mid^2)^{-1/2}(r, 1).$$

Then, for $s = (s_1, \ldots, s_{m+1})$, $Q^{-1}(s) := P(s) = (s_1, \ldots, s_m)s_{m+1}^{-1}$. Let $\Omega' = Q(U)$ and $\Omega = \text{closure}(\Omega')$. Then (Ω, Ω', P) is a metric compactification of U. For any $j = 1, 2, \ldots$, set $\varphi_j(r) = (1+ \mid r \mid^j)^{1/j}$, and observe that

$$\varphi_j(P(s))^{-1}g(y(t))P(s) = g(y(t))(s_1, \ldots, s_m)/\left(s_{m+1}^j + \left(\sum_{i=1}^m s_i^2\right)^{j/2}\right)^{1/j}$$

is continuous on Ω. Let $k \in \{1, 2, \ldots\}$ and $\varphi_0 = \varphi_k$. Then $\varphi_j \in \mathcal{F}(\Omega, P, \varphi_0)$ for all $j = 1, 2, \ldots$ and Theorem 2.5 is applicable for every choice of locally Lipschitzian h_0 and h_1. Thus the (relaxed or unrelaxed) extremals of the bounded control problems corresponding to different choices of $\varphi = \varphi_j$ are equivalent modulo a rescaling of the independent variable.

The next example illustrates the use of Theorem 3.1.

Example 4.2. For all $v = (v_1, v_2) \in \mathbf{R}^2$, let

$$f(v, r) = (1, r) \quad \forall\, r \in U = (0, \infty),$$

$$\chi_1(v, z) = 1,\ \chi_2(v, z) = (1, 1)^T z/ \mid z \mid \quad \forall\, z \in \mathbf{R}^2, \mid z \mid \geq 1.$$

Consider the systems

$$y' = g(y, \gamma) = (\mid \cos(\gamma) \mid, \mid \sin(\gamma) \mid), \qquad \gamma(t) \in \Gamma = [-\pi, \pi] \tag{4.2.1}$$

and

$$y' = h(y, \delta) = (\cos^2(y_2 + \delta), \sin^2(y_2 + \delta)), \qquad \delta(t) \in \Delta = [-\pi, \pi]. \tag{4.2.2}$$

Then, in the notation of Theorem 3.1,

$$g(v, \Gamma) = \text{closure}\{\varphi_{\chi_1}(v, r)^{-1} f(v, r) \mid r \in U\},$$

$$h(v, \Delta) = \text{closure}\{\varphi_{\chi_2}(v, r)^{-1} f(v, r) \mid r \in U\}.$$

We observe that int co $g(v, \Gamma) \neq \emptyset$. Thus Theorem 3.1 implies that systems (4.2.1) and (4.2.2) have sets of extremals(L) that are equivalent modulo a rescaling of the independent variable.

While the conditions of Theorems 2.5 and 3.1 are often easier to check than those of Theorem 3.3, the latter (with its weaker conclusions) can apply in some problems where the former theorems fail. The following is an example.

Example 4.3. Consider the system

$$y_1' = 1, \qquad y_2' = u, \qquad u(t) \in (0, \infty). \tag{4.3.1}$$

Let

$$\varphi(r) = (1 + r^2)^{1/2}(2 + \sin r^{-1}), \qquad \psi(r) = (1 + r^2)^{1/2}(2 + \sin r).$$

Theorem 2.5 does not apply here because, no matter what the choice of φ_0 is, this theorem requires that the function

$$\varphi(r)/\psi(r) = (2 + \sin r^{-1})/(2 + \sin r),$$

with $r = P(\omega)$, can be extended continuously from Ω' to Ω, which is impossible.

Now consider the following two different bounded reparametrizations of system (4.3.1):

$$y' = \omega = g_\varphi(\omega), \qquad \omega(t) \in \Omega_\varphi = \text{ closure } \{\varphi(r)^{-1}(1, r) \mid r \in (0, +\infty)\} \tag{4.3.2}$$

and

$$y' = \omega = g_\psi(\omega), \qquad \omega(t) \in \Omega_\psi = \text{closure}\{\psi(r)^{-1}(1, r) \mid r \in (0, +\infty)\}. \tag{4.3.3}$$

Theorem 3.1 does not apply in this case either because, in the setting of Theorem 3.1, the functions φ and ψ above correspond to

$$\chi_\varphi = 2 + \sin z_2^{-1}, \qquad \chi_\psi = 2 + \sin z_2,$$

and neither χ_φ nor χ_ψ has a continuous extension to $[Z]$. However, by choosing

$$\varphi_0(v, r) = (1 + r^2)^{1/2}, \quad K_\varphi(v, \omega) = \mid g_\varphi(\omega) \mid^{-1}, \quad K_\psi(v, \omega) = \mid g_\psi(\omega) \mid^{-1},$$

we can apply Theorem 3.3 to conclude that the sets of extremals$(L+)$ of (4.3.2) and (4.3.3) are equivalent up to a rescaling. We observe that the set Ω_φ is the union of the image of the function

$$r \mapsto (1 + r^2)^{-1/2}(2 + \sin r^{-1})^{-1}(r, 1) : (0, \infty) \mapsto \mathbf{R}^2,$$

of the singleton $\{(\frac{1}{2}, 0)\}$ and of the straight line segment $\{0\} \times [\frac{1}{3}, 1]$ while the set

$$\{K(v, \omega)g(v, \omega) \mid \omega \in \Omega_\varphi\}$$

is the unit quarter-circle in the first quadrant. A similar description applies to Ω_ψ.

Our last example illustrates an application of Theorem 3.2.

Example 4.4. Let $V \subset \mathbf{R}^2, w : V \mapsto [1, 2]$ be locally Lipschitzian but not differentiable, and $g_1, g_2 : V \mapsto \mathbf{R}^2$ be C^1 and such that the set $\{g_1(v), g_2(v)\}$ is linearly independent for all $v \in$ V. Let

$$f(v, r) = w(v)(g_1(v) + r[g_2(v) - g_1(v)]) \quad \forall r \in U = [0, 1].$$

Consider the control problems defined by

$$y'(t) = f(y(t), u(t)), \qquad u(t) \in U, \tag{4.4.1}$$

and

$$y'(t) = (y_1, y_2)'(t) = f(y(t), \sin^2(y_1(t) + \gamma(t))), \qquad \gamma(t) \in \Gamma = [0, \pi]. \tag{4.4.2}$$

These two control problems are both representations of the same differential inclusion $y'(t) \in G(y(t))$ almost everywhere. As an immediate consequence of Theorem 3.2 (e.g.,

with $\chi(v,z) =\mid z \mid$), they have the same set of extremals(L). However, Lojasiewicz's theorems do not apply here because systems (4.4.1) and (4.4.2) are not C^1 in the state variables, int co $G(v) = \emptyset$ for all v, and the function

$$(v,p) \mapsto \mathcal{H}(v,p) = \min_{r\in U} p^T f(v,r) = w(v)\min\{p^T g_1(v),\ p^T g_2(v)\}$$

is not differentiable for (p,v) such that $p^T g_1(v) - p^T g_2(v) = 0$ and wherever $w(\cdot)$ is not differentiable.

5. Appendix. The proofs of Lojasiewicz's Theorem and of our version of it require a specialized separation lemma. For the sake of completeness, we include a proof of this lemma.

LEMMA 5.1 (Lojasiewicz's separation lemma). *Let $X_1,\dots,X_k$ be compact and convex subsets of $\mathbf{R}^n$. Then $\cap_{i=1}^k X_i = \emptyset$ if and only if there exist half-spaces $H_1,\dots, H_{k-1}$ such that*

$$X_1 \subset H_1,\ H_i \cap X_{i+1} \subset H_{i+1} \quad \forall i = 1,2,\dots,k-2, \quad H_{k-1}\cap X_k = \emptyset.$$

Proof. The "if" part is obvious. We now assume that $\cap_{i=1}^k X_i = \emptyset$ and prove the "only if" part. Let

$$M = \max\{\mid x \mid \mid x \in \cup_{i=1}^k X_i\},\ K = \{(x,x,\dots,x) \in \mathbf{R}^{kn} \mid \mid x \mid \leq M\}$$

$$X = X_1 \times X_2 \times \cdots \times X_k.$$

Then K and X are convex and compact subsets of $\mathbf{R}^{kn}$ and $\cap_{i=1}^k X_i = \emptyset$ if and only if $X \cap K = \emptyset$. By the classical convex separation theorem, there exist n_i, α and a half-space

$$H = \left\{(x_1,x_2,\dots,x_k) \in \mathbf{R}^{kn} \mid \sum_{i=1}^k n_i \cdot x_i \leq \alpha\right\}$$

containing X and disjoint from K. Set $\beta_i = \max\{n_i \cdot x_i \mid x_i \in X_i\}$. Let

$$H_j = \left\{x \in \mathbf{R}^n \mid (n_1 + \ldots + n_j)\cdot x \leq \alpha - \sum_{i=j+1}^k \beta_i\right\}.$$

We claim that $H_1,\dots,H_{k-1}$ satisfy the statement of the lemma. Indeed, $x \in X_1$ implies $n_1 \cdot x \leq \beta_1$; hence $x \in H_1$. Furthermore, $x \in H_i \cap X_{i+1}$ for $i \in \{1,\cdots,k-2\}$ implies that $(n_1 + \cdots + n_i)\cdot x \leq \alpha - \sum_{j=i+1}^k \beta_j$ and $n_{i+1}\cdot x \leq \beta_{i+1}$. Therefore $(n_1 + \cdots + n_{i+1})\cdot x \leq \alpha - \sum_{j=i+2}^k \beta_j$, i.e., $x \in H_{i+1}$.

It remains to show that $H_{k-1} \cap X_k = \emptyset$. Assume that this is not the case and let $x \in H_{k-1}\cap X_k$. Then we have $(n_1 + \cdots + n_{k-1})\cdot x \leq \alpha - \beta_k$ and $n_k \cdot x \leq \beta_k$, i.e., $(n_1 + \cdots + n_k)\cdot x \leq \alpha$, which implies $(x,x,\dots,x) \in K \cap H$, a contradiction. □

Proof of the modified form of Lojasiewicz's Theorem. Let $g : V \times \Omega \mapsto \mathbf{R}^n$ be C^1 in v. Then we can extend g to $V \times$ rpm(Ω) by setting $g(v,\nu) := \int g(v,\omega)\nu(d\omega)$ for all $\nu \in$ rpm(Ω) and, as extended, g remains C^1 in v, with

$$g(v, \text{rpm}(\Omega)) = co \quad g(v,\Omega).$$

Thus it suffices to prove the theorem for original (i.e., unrelaxed) g-extremals(L) (the relaxed g-extremals(L) becoming the unrelaxed g-extremals(L) when U is replaced by rpm(U)).

Let $g_1 : V \times U \mapsto \mathbf{R}^n$ and $g_2 : V \times \Omega \mapsto \mathbf{R}^n$, and let (y, α) be a g_1-extremal(L) and σ an unrelaxed control such that

$$y(t) = \int_0^t g_2(y(s), \sigma(s))ds \quad \forall\, t \in [0, \alpha].$$

Define $M(t)$ to be the solution of the Cauchy problem

$$M'(t) = D_1 g_2(y(t), \sigma(t))M(t) \text{ a.e. in } [0, \alpha], \qquad M(\alpha) = I.$$

We can verify that, if $p_0 \in \mathbf{R}^n$ and $p(t)^T = p_0^T M(t)^{-1}$, then

$$p'(t)^T = -p(t)^T D_1 g_2(y(t), \sigma(t)).$$

Therefore it suffices to show that there exists $p_0 \in \mathbf{R}^n \backslash \{0\}$ such that

$$p_0 M(t)^{-1}(z - y'(t)) \geq 0 \quad \forall\, z \in G(y(t)), \qquad p_0 M(t)^{-1} y'(t) = 0.$$

Let $e : [0, \alpha] \mapsto \mathbf{R}^n$ be a measurable function and, for $t \in [0, \alpha]$, set

$$E(t) = \{(s, u) \in [0, \alpha] \times U \mid g_1(y(t) + s \cdot e(t), u) = g_2(y(t) + s \cdot e(t), \sigma(t))\}$$

and

$$E_0(t) = \{0\} \times U \cap \text{closure}\{E(t) \backslash (\{0\} \times U)\}.$$

Since g_1 and g_2 are representations of the sets $G(v)$ for all v, it follows that, for every choice of $t, s \in (0, \alpha]$, there exists $u \in U$ such that $(s, u) \in E(t)$. Thus $E_0(t) \neq \emptyset$ and, by well-known measurability theorems (see, e.g., [W3, I.7.6 and I.7.7, pp. 150–151]), $E(\cdot)$ and $E_0(\cdot)$ are measurable set-valued mappings with closed values and there exists a measurable selection $\tau(\cdot)$ of $E_0(\cdot)$. Denote by u_e the second coordinate of τ. Then u_e is measurable and $u_e(t) \in U$ almost everywhere in $[0, \alpha]$. Moreover, for almost all $t \in [0, \alpha]$ and all $k = 1, 2, \ldots$, we can find $s_k > 0$ and $u_k \in U$ satisfying

$$g_1(y(t) + s_k \cdot e(t), u_k) = g_2(y(t) + s_k \cdot e(t), \sigma(t)), \tag{5.1}$$

$$\lim_{k \to \infty} (s_k, u_k) = (0, u_e(t)).$$

Fix $t \in [0, \alpha]$ such that $y'(t) = g_2(y(t), \sigma(t))$ and (5.1) holds. Then

$$\begin{aligned} g_1(y(t) + s_k \cdot e(t), u_k) &= g_1(y(t), u_k) + s_k D_1 g_1(y(t), u_k)e(t) + o(s_k) \\ &= g_2(y(t), \sigma(t)) + s_k D_1 g_2(y(t), \sigma(t))e(t) + o(s_k) \\ &= g_2(y(t) + s_k \cdot e(t), \sigma(t)). \end{aligned}$$

Therefore

$$\lim_{k \to \infty} s_k^{-1}(g_1(y(t), u_k) - g_2(y(t), \sigma(t))) \tag{5.2}$$

$$= -[D_1 g_1(y(t), u_e(t)) - D_1 g_2(y(t), \sigma(t))]e(t).$$

Relation (5.1) implies that $g_1(y(t), u_e(t)) = g_2(y(t), \sigma(t))) = y'(t)$ for almost all $t \in [0, \alpha]$; hence u_e is a control corresponding to our g_1-extremal(L) (y, α) and therefore there exists a corresponding arc p_e such that

$$p_e(t)^T y'(t) = p_e(t)^T g_2(y(t), \sigma(t)) \leq p_e(t)^T g_1(y(t), u_k);$$

hence, by (5.2),

$$0 \leq -p_e(t)^T[D_1 g_1(y(t), u_e(t)) - D_1 g_2(y(t), \sigma(t))]e(t) \\ = [p_e'(t)^T + p_e(t)^T D_1 g_2(y(t), \sigma(t))]e(t). \tag{5.3}$$

Let $q_e(t)^T = p_e(t)^T M(t)$. Then (5.3) yields

$$q_e'(t)^T M(t)^{-1} e(t) \geq 0. \tag{5.4}$$

Next consider, for each t, the nonempty cone

$$P(t) = \{p \in \mathbf{R}^n \mid p^T M(t)^{-1}(z - y'(t)) \geq 0 \quad \forall z \in G(y(t)), \qquad p^T M(t)^{-1} y'(t) = 0\}.$$

Let B be the closed unit ball in $\mathbf{R}^n$ and ∂B its boundary. It follows from Lusin's theorem that the measurable compact-valued mapping $t \mapsto Q(t) := P(t) \cap \partial B$ admits a subset I of $[0, \alpha]$ of full measure such that $y'(t) \in G(y(t))$ for all $t \in I$ and $Q(\cdot)$ is approximately continuous at each $t \in I$. Therefore any continuous function $p : [0, \alpha] \mapsto \mathbf{R}^n$ such that $p(t) \in P(t)$ almost everywhere also satisfies $p(t) \in P(t)$ for all $t \in I$.

We next prove that

$$\bigcap_{t \in [0,\alpha]} P(t) \neq \{0\}. \tag{5.5}$$

First observe that $D_1 g_1(y(t), U)$ is bounded and there exists, therefore, a constant r such that every adjoint vector function $p(\cdot)$ of the g_1-extremal (y, α), with $\mid p(\alpha) \mid = 1$, satisfies $r < \mid p(t)^T M(t) \mid < 1/r$. Assume that (5.5) does not hold, i.e., $\cap_{t \in [0,\alpha]} P(t) = \{0\}$. Then $\cap_{t \in [0,\alpha]} P(t) \cap \partial B = \emptyset$ and, since $P(t) \cap \partial B$ are compact, it follows that there exist $k \geq 2$ and $t_1, \ldots, t_k \in I$ with $0 = t_1 < t_2 < \cdots < t_k \leq \alpha$ such that $\cap_{1 \leq i \leq k} P(t_i) \cap \partial B = \emptyset$; hence $\cap_{1 \leq i \leq k} P(t_i) = \{0\}$. Let X_i be the convex hull of the set of all $x \in P(t_i)$ satisfying

$$r \leq \mid x \mid \leq 1/r.$$

Observe that each $P(t_i)$ is a proper cone (i.e., $P(t_i)$ does not contain a straight line) because co $G(y(t))$ has a nonempty interior. Therefore $0 \notin X_i$ and $\cap_{1 \leq i \leq k} X_i = \emptyset$. By Lemma 5.1, this last relation implies that there exist closed half-spaces $H_1, \ldots H_{k-1}$ such that

$$X_1 \subset H_1, \ H_i \cap X_{i+1} \subset H_{i+1} \quad \forall\, i = 1, 2, \ldots, k-2, \qquad H_{k-1} \cap X_k = \emptyset. \tag{5.6}$$

Let n_i be the outward unit normal to H_i, i.e., $\mid n_i \mid = 1$ and, for some c_i, $H_i = \{p \in \mathbf{R}^n \mid p \cdot n_i \geq c_i\}$. Define $e : [0, \alpha] \mapsto \mathbf{R}^n$ as follows: $e(t) = M(t) n_i$ when $t \in [t_i, t_{i+1})$ and $i = 1, 2, \cdots k-1$. Consider corresponding p_e and q_e as defined in the previous paragraph. Then

$$q_e(t_i) \in X_i \text{ for } i = 1, 2, \ldots, k. \tag{5.7}$$

By (5.4), we have $q_e'(t) \cdot n_i \geq 0$ almost everywhere in $[t_i, t_{i+1}]$. Therefore $q_e(t_{i+1}) \cdot n_i \geq q_e(t_i) \cdot n_i$. Thus we have the following implication:

$$q_e(t_i) \in H_i \ \text{ implies } \ q_e(t_{i+1}) \in H_i. \tag{5.8}$$

Relations (5.6)–(5.8) yield $q_e(t_k) \in H_{k-1} \cap X_k = \emptyset$, which is absurd. Thus there exists a nonzero $p_0 \in P(t)$ such that

$$p_0 M(t)^{-1}(z - y'(t)) \geq 0 \quad \forall z \in G(y(t)), \quad p_0 M(t)^{-1} y'(t) = 0.$$

This completes the proof. □

Acknowledgment. We thank S. Lojasiewicz, Jr. for helpful comments on an early version of this paper.

REFERENCES

[C] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[L1] S. LOJASIEWICZ JR., *Invariance of Extremals*, preprint, 1988.

[L2] ———, *Invariance of Extremals, Nonlinear Controllability and Optimal Control*, H. Sussmann, ed. (Series in Pure and Applied Math., Vol. 133), Marcel Dekker, 1990.

[L3] ———, Private communication, 1992.

[N] L. W. NEUSTADT, *A general theory of minimum-fuel space trajectories*, J. SIAM, Series A, Control 3 (1965), pp. 317–356.

[R] R. W. RISHEL, *An extended principle for control systems whose control laws contain measures*, J. SIAM, Series A, Control 3 (1965), pp. 191–205.

[S] W. W. SCHMAEDEKE, *Optimal control theory for nonlinear vector differential equations containing measures*, J. SIAM, Series A, Control 3 (1965), pp. 231–280.

[W1] J. WARGA, *Necessary conditions for minimum in relaxed variational problems*, J. Math. Anal. Applic. 4 (1962), pp. 129–145.

[W2] ———, *Variational problems with unbounded controls*, J. SIAM, Series A, Control 3 (1965), pp. 424–438.

[W3] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[W4] ———, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.

[W5] ———, *Derivate containers, inverse functions, and controllability*, in Calculus of Variations and Control Theory, D.L. Russell, ed., Academic Press, New York, 1976.

[W6] ———, *Fat homeomorphisms and unbounded derivate containers*, J. Math. Anal. Appl. 81 (1981), pp. 545–560; 90 (1982), pp. 582–583.

[W7] ———, *Homeomorphisms and local C1 approximations*, J. Nonlinear Anal., TMA 12 (1988), pp. 593–597.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1170–1185, July 1994

016

CONTROLLABILITY OF A SYSTEM OF TWO SYMMETRIC RIGID BODIES IN THREE SPACE*

MICHAEL J. ENOS†

Abstract. Consider a mechanical system consisting of two completely symmetric, three-dimensional rigid bodies, each with inertia tensor the identity matrix and mass center at $0 \in \boldsymbol{R}^3$. Imposing the constraint that this system have zero total angular momentum, the angular velocities of these bodies are negatives of one another, and the transfer of this system from one position to another is nonholonomic. While Chow's theorem establishes the fixed-endpoint controllability of this system, this result does not explicitly exhibit any motions between a given set of endpoints. This paper shows how to explicitly construct simple motions of this system on $[0, 1]$ with arbitrary endpoints in $SO(3)^2$. In particular, using normalized quaternions to describe rotations in terms of elementary functions, a continuous motion with the given endpoints is constructed; it consists of at most three successive motions during each of which the bodies rotate on fixed axes.

Key words. controllability, system of rigid bodies, nonholonomic mechanics, rotation group

AMS subject classifications. 49, 70, 93

1. Introduction.

Recent papers ([5], [8]) have studied the free motion of a mechanical system consisting of two identical, three-dimensional rigid bodies attached at an ideal spherical joint. Here we consider a simple realization of this system; in particular, we assume that the mass centers of the bodies are coincident and that each body has a completely symmetric mass distribution about its mass center. With this latter assumption we may assume that the inertia tensor of each body is the identity matrix E.

This system will be denoted $(\mathcal{B}_1, \mathcal{B}_2)$ and positions of the $\mathcal{B}_i$ will be identified with matrices $A_i \in SO(3)$ in the standard way.

As a physical example of this system, we might consider a satellite with spherical, concentric inner and outer shells connected by a Cardan suspension of negligible mass, the inner shell being more dense than the outer one (see Fig. 1). This example differs from the "classical" dual-spin satellite in the nature of the coupling of the bodies; in classical models (see [3],[6]), the inner body is constrained to rotate about an axis fixed in the outer one, while here the bodies have full $SO(3)$ freedom in their relative orientations as α, β and γ are varied.

As in the papers mentioned above, we will consider motions of this system in the absence of external forces, and we will focus on those with zero angular momentum; our assumptions then guarantee that

$$\omega_1 + \omega_2 = 0, \tag{1.1}$$

where ω_i is the angular velocity of $\mathcal{B}_i$, $i = 1, 2$, so the bodies instantaneously rotate with opposite angular velocities. On the other hand, we will consider a larger class of motions than the free motions and the construction of paths between given endpoints. We can study these motions by considering the control system with state (A_1, A_2) and

* Received by the editors February 25, 1991; accepted for publication (in revised form) January 4, 1993. This research was partially supported by the Ministry of Colleges and Universities of Ontario and the Natural Sciences and Engineering Research Council of Canada.

† The Fields Institute for Research in Mathematical Sciences, Waterloo, Ontario, Canada N2L 5Z5.

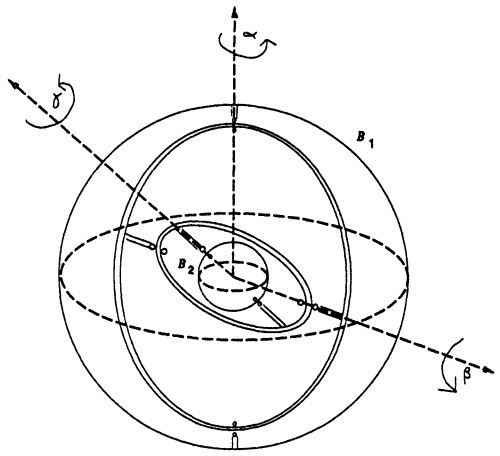


FIG. 1

control $\omega \stackrel{\text{def}}{=} \omega_1$, for which the state satisfies the differential equation

$$\left(\dot{A}_1, \dot{A}_2\right) = (\hat{\omega} A_1, -\hat{\omega} A_2), \tag{1.2}$$

where $\hat{\omega} \in so(3)$ is the skew-symmetric matrix of the cross product operator $v \mapsto \omega \times v$ on $\mathbb{R}^3$.

The free motions of this system are trivial to obtain; indeed, during a free, zero-angular-momentum motion the bodies rotate about a fixed axis in opposite directions at constant angular rates. The collection of all zero angular momentum motions is much larger; for instance, for any motion A' of the first body and initial condition $(A_1, A_2)(0)$ of the whole system, there is a corresponding zero angular momentum motion (A_1, A_2) with this initial condition for which $A_1 = A'_1$.

From a mechanical standpoint, we can view those motions for which ω is non-constant as arising from internal torques exerted by the $\mathcal{B}_i$ on each other (indeed, by Euler's equations, a torque of $2\dot{\omega}_1$ exerted by $\mathcal{B}_2$ on $\mathcal{B}_1$ produces such a motion). In the example in Fig. 1, these torques could be affected by motors of negligible mass that control the angles α, β, and γ.

Let $\mathcal{S}$ denote the collection of continuous solutions of (1.2) on $[0, 1]$ with piecewise constant controls ω that are right continuous and have only finitely many discontinuities. This paper concerns the controllability of this system for given endpoints $(A_1, A_2)(0) = (A_1^{(0)}, A_2^{(0)})$ and $(A_1, A_2)(1) = (A_1^{(1)}, A_2^{(1)})$ with motions in $\mathcal{S}$. Physically speaking, such motions are impossible since they require impulsive, Dirac-type torques at the discontinuities; on the other hand, we can always reparametrize a motion in $\mathcal{S}$ (in such a way that the controls vanish at the discontinuities) to obtain a smooth, physically possible motion.

Although zero-total-angular-momentum motions of the bodies are highly constrained, the constraint is on the velocities and the system is nonholonomic. For instance, if the first body rotates through an angle of $\pi/2$ about the fixed vector $\mathbf{i}$, the net rotation of the second body is through $\pi/2$ on $-\mathbf{i}$; on the other hand, if the system undergoes a series of free motions during which the first body rotates through an angle of $\pi/2$ on the vectors $\mathbf{j}$, $\mathbf{i}$, then $-\mathbf{k}$, its net rotation is again through $\pi/2$ on $\mathbf{i}$ but we obtain a different net rotation of the second body.

In fact, it is not hard to see that this system is controllable. Let $M = SO(3)^2$; using the right invariance of the system (2), we can identify each tangent space T_pM with $so(3)^2$, or, equivalently, the set of all pairs (ω_1, ω_2) of angular velocity vectors given in space coordinates. Using the Riemannian metric $\langle(\omega_1,\omega_2),(\omega_1',\omega_2')\rangle = 1/2(\langle\omega_1,\omega_1'\rangle + \langle\omega_2,\omega_2'\rangle)$ on TM, let H denote the distribution on TM spanned by vector fields of the form $(\omega, -\omega)$ and V the distribution spanned by those of the form (ω,ω). The velocity along any zero angular momentum motion must lie in H and the Lie bracket $[(\omega,-\omega),(\omega',-\omega')]$ of any two such vector fields is $(\omega\times\omega', \omega\times\omega') \in V$. Since it is easily seen that $TM = H \oplus V$ and any vector field in V can be obtained as a bracket of vector fields in H, it follows that the brackets of vector fields in H generate the entire tangent bundle TM, so the system is controllable by Chow's theorem. Since the results of Chow's theorem also imply controllability with piecewise-constant controls, the system is controllable with motions in $\mathcal{S}$.

While Chow's theorem establishes controllability, it gives no immediate information about those motions in S that have a given set of endpoints. Our main objective in this paper is the explicit construction of a simple class of motions in $\mathcal{S}$ with given endpoints. In particular, we will show that for any choice of endpoints in M, there is a motion in $\mathcal{S}$ with these endpoints and for which the control has at most two discontinuities, and further produce this motion explicitly in terms of elementary functions.

It is surprising that beyond the use of Chow's theorem to establish controllability, very little work has been done on the explicit construction of motions of nonholonomic systems with given endpoints. A notable exception is [7], where the controllability with certain constrained motions of a two-body, freely rotating system (generally referred to as "the falling cat") is considered. While that example involves a system with less material symmetry, the constraints on the coupling of the bodies are so restrictive that the motions of this system are essentially those of a single rigid body. The anholonomy in our present example is much more pronounced.

In §2 we show that any motion in $\mathcal{S}$ is completely determined by a finite collection of rotations and that the construction of a motion in $\mathcal{S}$ with given endpoints is equivalent to obtaining suitable factorizations for each of a pair of rotations.

For general choices of endpoints, the construction of a motion in $\mathcal{S}$ with those endpoints involves eight different rotations, and the use of matrix coordinates for rotations becomes cumbersome. On the other hand, when we instead use normalized quaternions to parametrize $SO(3)$, the computations become simpler; moreover, with this approach, many features of the underlying geometry in this problem become apparent. In §3 we recall the algebraic properties of the group of normalized quaternions and describe a useful geometric interpretation of quaternions due to Rodrigues. We restate the factorization of the preceding section in these new coordinates.

In §4, we describe those endpoints that can be joined with a motion with constant control and show that in general a control with at least two discontinuities is required for a given set of endpoints. We also state our main result here, that at most two discontinuities in the controls are required for a given set of endpoints.

Section 5 gives a factorization that implies controllability with one discontinuity on the control when certain conditions are present on the endpoints. In §6, we show that for any set of endpoints, we can, by making a slight adjustment (which amounts to adding a discontinuity in the control), reduce the situation to that in §5.

Hence, combining these results, we obtain our main result in §7. At each step we provide explicit formulas for constructing these motions.

2. Controllability with motions in $\mathcal{S}$. As described in the Introduction, we will consider the control system $((A_1, A_2), \omega)$ with state (A_1, A_2) and control ω taking values in $SO(3)^2$ and $\mathbb{R}^3$, respectively, and satisfying the differential equation

$$\left(\dot{A}_1, \dot{A}_2\right) = (\hat{\omega} A_1, -\hat{\omega} A_2)\,. \tag{2.1}$$

We will study the collection $\mathcal{S}$, those continuous motions (A_1, A_2) on $[0, 1]$ with right continuous, piecewise constant controls ω having only finitely many discontinuities.

If $(A_1, A_2) \in \mathcal{S}$, it is clear that we can write the control ω corresponding to this motion as

$$\omega = \sum_{j=1}^{N-1} \omega^j \chi_{[t_{j-1}, t_j)} + \omega^N \chi_{[t_{N-1}, t_N]}, \tag{2.2}$$

where $N < \infty$,

$$0 = t_0 < t_1 < \cdots < t_N = 1$$

is some partition of $[0, 1]$, and the ω^j are vectors in $\mathbb{R}^3$.

When $(A_1, A_2) \in \mathcal{S}$, it is easy to solve (1) on any interval of continuity $[t_{j-1}, t_j)$ of ω by exponentiating $\hat{\omega}$; in particular, we have

$$(A_1, A_2)(t) = \left(D_j(t) A_1(t_{j-1}), D_j^t(t) A_2(t_{j-1})\right), \qquad t \in [t_{j-1}, t_j],$$

where

$$D_j(t) = \exp\left((t - t_{j-1})\hat{\omega}^j\right). \tag{2.3}$$

It is easy to see that $D_j(t)$ is the matrix of the rotation through the angle $(t - t_{j-1})|\omega^j|$ about the constant vector $\omega^j/|\omega^j|$ when $|\omega^j| \neq 0$ and $D_j(t)$ is the identity rotation when $\omega^j = 0$.

Applying this fact successively on the intervals $[t_{j-1}, t_j)$ and defining the A_i so as to be continuous on $[0, 1]$, we have the following lemma.

LEMMA 2.1. *Any motion in $\mathcal{S}$ is of the form*

$$\left(A_1, A_2\right)(t) = \left(D_j(t) D_{j-1}(t_{j-1}) \cdots D_1(t_1)\, A_1(0), D_j^t(t) D_{j-1}^t(t_{j-1}) \cdots D_1^t(t_1)\, A_2(0)\right), \qquad t \in [t_{j-1}, t_j)$$

for some partition $\{t_j\}_{j=1}^N$ of $[0, 1]$ and set of vectors $\{\omega^j\}_{j=1}^N$, where the D_j are as in (2.3).

Setting $t = 1$ in this formula and letting $G_j \overset{\text{def}}{=} D_j(t_j), 1 \leq j \leq N$, we have the following characterization of those motions in $\mathcal{S}$ with a given set of endpoints.

COROLLARY 2.2. *Let $A_i^{(k)} \in SO(3)$, $i = 1, 2$ and $k = 0, 1$. Any motion $(A_1, A_2)_* \in \mathcal{S}$ which satisfies the endpoint conditions*

$$(A_1, A_2)_*(0) = \left(A_1^{(0)}, A_2^{(0)}\right) \quad \text{and} \quad (A_1, A_2)_*(1) = \left(A_1^{(1)}, A_2^{(1)}\right)$$

has the form in Lemma 2.1, *and the G_j satisfy*

$$\left(A_1^{(1)} A_1^{(0)t}, A_2^{(1)} A_2^{(0)t}\right) = \left(G_N \cdots G_1, G_N^t \cdots G_1^t\right). \tag{2.4}$$

□

This reduces the problem of constructing a motion in $\mathcal{S}$ with given endpoints to that of obtaining a suitable factorization of a given pair of rotations. In particular, we want to factor the *net rotations* $A_i^{(1)} A_i^{(0)t}$ that the bodies undergo during the desired fixed-endpoint motion.

Remark 2.3. Given rotations G_j with the properties in Corollary 2.2, it is easy to construct a motion in $\mathcal{S}$: Let ϕ^j and $\mathbf{n}^j$ denote the angle and axis of the rotation G_j, let $t_j = j/N$, and define

$$\omega^j \stackrel{\text{def}}{=} \frac{\phi^j}{t_j - t_{j-1}} \mathbf{n}^j,$$

$1 \leq j \leq N$. Then the formula in Lemma 2.1, with $D_j(t)$ as in (2.3), is a motion in $\mathcal{S}$ with the correct endpoints.

3. Quaternions. Let $C \in SO(3)$. The action of C on $\mathbb{R}^3$ is that of rotation through an angle ϕ about a unit vector $\mathbf{n}$. Assuming that a positive angle ϕ rotates vectors in a counterclockwise direction when viewed from the tip of $\mathbf{n}$, we will write $R(\phi\mathbf{n})$ for this rotation. It is clear that any rotation can be represented as $R(\phi\mathbf{n})$ for some $\mathbf{n}$ and some unique $\phi \in [0, \pi]$. Formulas for ϕ and $\mathbf{n}$ in terms of the matrix entries of C can be found in the Appendix.

For any rotation $R(\phi\mathbf{n})$, put

$$(\lambda, \Lambda) \stackrel{\text{def}}{=} \left(\cos\frac{\phi}{2}, \sin\frac{\phi}{2}\mathbf{n}\right). \tag{3.1}$$

We say that λ and Λ are *Euler–Rodrigues parameters* (ERPs) for $R(\phi\mathbf{n})$. It is clear that any (λ, Λ) so defined satisfies

$$\lambda^2 + |\Lambda|^2 = 1, \tag{3.2}$$

and so may be naturally identified with a point on the three sphere S^3.

Algebraically, those (λ, Λ) that satisfy (3.2) form the group of normalized quaternions, using quaternion multiplication:

$$(\lambda_1, \Lambda_1)(\lambda_2, \Lambda_2) = (\lambda_1\lambda_2 - \langle \Lambda_1, \Lambda_2 \rangle, \lambda_1\Lambda_2 + \lambda_2\Lambda_1 + \Lambda_1 \times \Lambda_2); \tag{3.3}$$

the group identity is $(1, 0)$, and group elements are inverted using the formula

$$(\lambda, \Lambda)(\lambda, -\Lambda) = (\lambda, -\Lambda)(\lambda, \Lambda) = (1, 0). \tag{3.4}$$

Let $\hat{R} : S^3 \mapsto SO(3)$ be the map taking a quaternion to the rotation it represents, i.e., $\hat{R}(\cos\frac{\phi}{2}, \sin\frac{\phi}{2}\mathbf{n}) = R(\phi\mathbf{n})$. $\hat{R}$ is a covering map and group homomorphism; moreover it is a two-to-one map, since

$$\hat{R}(\lambda_1, \Lambda_1) = \hat{R}(\lambda_2, \Lambda_2) \Longleftrightarrow (\lambda_1, \Lambda_1) = \pm(\lambda_2, \Lambda_2) \tag{3.5}$$

(this is a nice illustration of the fact that $SO(3)$ is the quotient of S^3 by the antipodal map).

Finally, if $(\lambda, \Lambda) \in S^3$ and $u \in \mathbb{R}^3$, we have

$$(0, \hat{R}(\lambda, \Lambda)u) = (\lambda, \Lambda)(0, u)(\lambda, -\Lambda);$$

using this formula, we can recover the matrix $\hat{R}(\lambda, \Lambda)$ from (λ, Λ) by conjugating $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ with (λ, Λ) to obtain the column vectors of this matrix.

Remark 3.1. Using the fact that $\hat{R}$ is a group homomorphism and (3.5), it is clear that the factorization (2.4) of Corollary 2.2 is equivalent to

$$(\gamma_N, \Gamma_N) \cdots (\gamma_1, \Gamma_1) = \pm (\lambda_1, \Lambda_1)$$

and

$$(\gamma_N, -\Gamma_N) \cdots (\gamma_1, -\Gamma_1) = \pm (\lambda_2, \Lambda_2), \tag{3.6}$$

where the (λ_i, Λ_i) are ERPs for the $A_i^{(1)} A_i^{(0)t}$ and the (γ_j, Γ_j) are ERPs for the G_j.

We will find quaternions more convenient for the computations that follow and study the factorization (3.6) in what follows.

By (3.5), it is clear that any rotation has ERPs (λ, Λ) with $\lambda \geq 0$; with this choice we have $\lambda = \cos \frac{\phi}{2}$, where ϕ is the "standard" angle of the rotation, i.e., that in $[0, \pi]$. In what follows we will assume the λ_i in Remark 3.1 are nonegative.

4. Controllability with constant and single-switch controls and statement of Theorem 4.2. In constructing motions in $\mathcal{S}$ with a given set of endpoints, an obvious first question to ask is how many discontinuities in the controls (equivalently, how many factors in the products in (2.4) and (3.6)) are required for a motion with a given set of endpoints. It should come as no surprise that only very special choices of endpoints may be connected with a path with constant control. In fact, this is only possible when the net rotations of the bodies between the endpoints are inverses of each other.

LEMMA 4.1. *Let* $A_i^{(k)} \in SO(3)$ *and let* (λ_i, Λ_i) *be as in Remark* 3.1, $i = 1, 2$ *and* $k = 0, 1$. *There is a path with these endpoints and constant control if and only if*

$$(\lambda_1, -\Lambda_1) = (\lambda_2, \Lambda_2) \quad \text{if } \lambda_1 \neq 0$$

or

$$\lambda_2 = 0 \quad \text{and} \quad \Lambda_2 = \pm \Lambda_1 \quad \text{if } \lambda_1 = 0.$$

In either case we can take $N = 1$ *and* $(\gamma_1, \Gamma_1) = (\lambda_1, -\Lambda_1)$ *in Remark* 3.1.

Proof. Controllability with a constant control means we can take $N = 1$ in Corollary 2.2, i.e., using Remark 3.1, there exists (γ, Γ) such that

$$\pm (\lambda_1, \Lambda_1) = (\gamma, \Gamma) = \pm (\lambda_2, -\Lambda_2).$$

If $\lambda_1 \neq 0$, our assumption that each λ_i is positive implies $\Lambda_2 = -\Lambda_1$, and if $\lambda_1 = 0$, this only requires $\Lambda_1 = \pm \Lambda_2$. □

As a second try, we can consider those motions in $\mathcal{S}$ with controls having one discontinuity. Taking $N = 2$ in (3.6), we require (γ_i, Γ_i), $i = 1, 2$, such that

$$(\gamma_2, \Gamma_2)(\gamma_1, \Gamma_1) = \pm (\lambda_1, \Lambda_1) \quad \text{and} \quad (\gamma_2, -\Gamma_2)(\gamma_1, -\Gamma_1) = \pm (\lambda_2, \Lambda_2).$$

Making use of (3.3), this shows that we must have

$$\pm \lambda_1 = \gamma_2 \gamma_1 - \langle \Gamma_2, \Gamma_1 \rangle = \pm \lambda_2,$$

and since the λ_i have the same sign by assumption, this means $\lambda_1 = \lambda_2$. In other words, by (3.1), the net rotations of the bodies must have the same angle of rotation.

Since this is not generally true, we must consider those controls with at least two switches.

As it turns out, two is always enough; in what follows, we will establish the following theorem.

THEOREM 4.2. *For any choice of the* (λ_i, Λ_i), *there exist* (γ_j, Γ_j), $1 \leq j \leq 3$, *such that* $\gamma_j \geq 0$,

$$(\gamma_3, \Gamma_3)(\gamma_2, \Gamma_2)(\gamma_1, \Gamma_1) = (\lambda_1, \Lambda_1), \quad \textit{and} \quad (\gamma_3, -\Gamma_3)(\gamma_2, -\Gamma_2)(\gamma_1, -\Gamma_1) = (\lambda_2, \Lambda_2).$$

In the process of proving this result, we will explicitly construct and geometrically describe the (γ_j, Γ_j). We will also derive formulas for the (γ_j, Γ_j) in terms of elementary functions of the (λ_i, Λ_i).

Remark 4.3. Note, by the discussion at the beginning of §3, that the angles ϕ^j and axes $\mathbf{n}^j$ of the rotations $\hat{R}(\gamma_j, \Gamma_j)$ are immediately available from the (γ_j, Γ_j); in particular, we have $\phi^j = 2\cos^{-1}\gamma_j \in [0, \pi]$ and $\mathbf{n}^j = \Gamma_j/|\Gamma_j|$ $(\Gamma_j \neq 0)$. We can then obtain the desired motion as in Remark 2.3.

5. A result on motions with a single switch. As noted above, the factorization in Remark 3.1 with $N = 2$ requires $\lambda_1 = \lambda_2$. Here we describe a condition that is also sufficient for the factorization, and derive formulas for the (γ_j, Γ_j) in this case.

We will combine this result with one from the next section to prove Theorem 4.2.

PROPOSITION 5.1. *Suppose* (δ_i, Δ_i) $(i = 1, 2)$ *are such that* $\delta_1 = \delta_2 \in [0, 1)$ *and the* Δ_i *are linearly independent. Then there exist* (γ_j, Γ_j), $j = 2, 3$, *such that each* γ_j *is nonnegative,*

$$(\gamma_3, \Gamma_3)\,(\gamma_2, \Gamma_2) = (\delta_1, -\Delta_1) \quad \textit{and} \quad (\gamma_3, -\Gamma_3)\,(\gamma_2, -\Gamma_2) = (\delta_2, -\Delta_2)\,.$$

Explicitly,

$$(\gamma_2, \Gamma_2) = \left(\frac{|\Delta_2 - \Delta_1|}{2\,|\Delta_1|} \cos\theta_2, \sin\theta_2 \, \frac{\Delta_2 - \Delta_1}{|\Delta_2 - \Delta_1|} + \cos\theta_2 \, \frac{|\Delta_1 + \Delta_2|}{2\,|\Delta_1|\,|\Delta_2 \times \Delta_1|} \Delta_2 \times \Delta_1 \right)$$

and

$$(\gamma_3, \Gamma_3) = \left(\frac{|\Delta_2 - \Delta_1|}{2\,|\Delta_1|} \cos\theta_3, \sin\theta_3 \, \frac{\Delta_2 - \Delta_1}{|\Delta_2 - \Delta_1|} - \cos\theta_3 \, \frac{|\Delta_1 + \Delta_2|}{2\,|\Delta_1|\,|\Delta_2 \times \Delta_1|} \Delta_2 \times \Delta_1 \right),$$

where $\theta_2, \theta_3 \in \left[0, \frac{\pi}{2}\right]$ *is any pair of angles satisfying* $\theta_2 + \theta_3 = \frac{\phi}{2}$ *and* ϕ *is the unique angle in* $(0, \pi]$ *such that* $\delta_1 = \cos\frac{\phi}{2}$.

Proof. Let $\phi \stackrel{\text{def}}{=} 2\cos^{-1}\delta_1 \in (0, \pi]$, and note that by assumption, $|\Delta_i| = \sin\frac{\phi}{2} \in (0, 1]$ for each i. Since the Δ_i are linearly independent,

$$u \stackrel{\text{def}}{=} -\frac{\Delta_1 + \Delta_2}{2} \quad \text{and} \quad v \stackrel{\text{def}}{=} \frac{\Delta_2 - \Delta_1}{2} \tag{5.1}$$

are nonzero; moreover, it is easily seen that

$$\langle u, v \rangle = 0 \quad \text{and} \quad |u|^2 + |v|^2 = |\Delta_1|^2 = \sin^2\frac{\phi}{2}. \tag{5.2}$$

The construction will be more transparent under a rotation of axes. Recall that if $D \in SO(3)$ and $q \in \mathbb{R}^3$, then the coordinates of the vector q observed from the

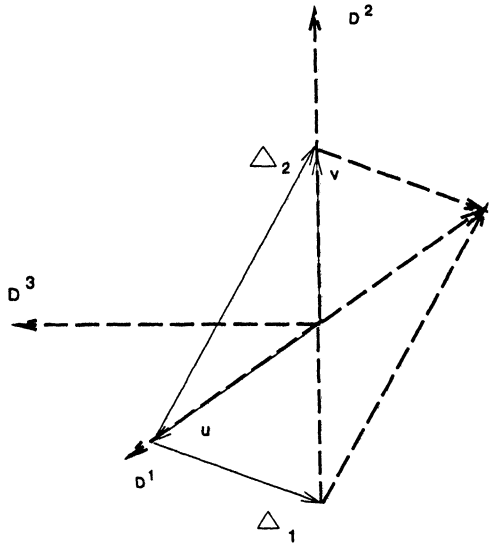


FIG. 2

coordinate axes $\{D^k\}_{k=1}^3$ are given by $D^t q$. Now, if we define the matrix D by

$$D \stackrel{\text{def}}{=} \left(\frac{u}{|u|} \;\vdots\; \frac{v}{|v|} \;\vdots\; \frac{u}{|u|} \times \frac{v}{|v|} \right), \tag{5.3}$$

then $D \in SO(3)$, $D^t u = |u|\,\mathbf{i}$, and $D^t v = |v|\,\mathbf{j}$ (see Fig. 2).

Pick $\theta_j \in \left[0, \frac{\pi}{2}\right]$, $j = 2, 3$, to satisfy $\theta_2 + \theta_3 = \frac{\phi}{2}$ and define

(5.4)

$$\gamma_j \stackrel{\text{def}}{=} \frac{|v| \cos \theta_j}{\sin \frac{\phi}{2}}, \quad j = 2, 3, \quad \overline{\Gamma}_2 \stackrel{\text{def}}{=} \begin{pmatrix} 0 \\ \sin \theta_2 \\ \frac{|u| \cos \theta_2}{\sin \frac{\phi}{2}} \end{pmatrix}, \quad \text{and} \quad \overline{\Gamma}_3 \stackrel{\text{def}}{=} \begin{pmatrix} 0 \\ \sin \theta_3 \\ \frac{-|u| \cos \theta_3}{\sin \frac{\phi}{2}} \end{pmatrix}.$$

Then

$$\left|\overline{\Gamma}_j\right|^2 + \gamma_j^2 = \frac{(|u|^2 + |v|^2) \cos^2 \theta_j}{\sin^2 \frac{\phi}{2}} + \sin^2 \theta_j = \cos^2 \theta_j + \sin^2 \theta_j = 1 \quad (j = 2, 3),$$

$$\gamma_2 \gamma_3 - \left\langle \overline{\Gamma}_2, \overline{\Gamma}_3 \right\rangle = \frac{|v|^2 + |u|^2}{\sin^2 \frac{\phi}{2}} \cos \theta_2 \cos \theta_3 - \sin \theta_2 \sin \theta_3 = \cos(\theta_2 + \theta_3) = \delta_1 = \delta_2,$$

$$\begin{aligned} \gamma_2 \overline{\Gamma}_3 + \gamma_3 \overline{\Gamma}_2 &= \frac{|v| \cos \theta_2}{\sin \frac{\phi}{2}} \begin{pmatrix} 0 \\ \sin \theta_3 \\ \frac{-|u| \cos \theta_3}{\sin \frac{\phi}{2}} \end{pmatrix} + \frac{|v| \cos \theta_3}{\sin \frac{\phi}{2}} \begin{pmatrix} 0 \\ \sin \theta_2 \\ \frac{|u| \cos \theta_2}{\sin \frac{\phi}{2}} \end{pmatrix} \\ &= \frac{|v| \sin(\theta_2 + \theta_3)}{\sin \frac{\phi}{2}} \mathbf{j} = |v|\,\mathbf{j} = D^t v, \end{aligned}$$

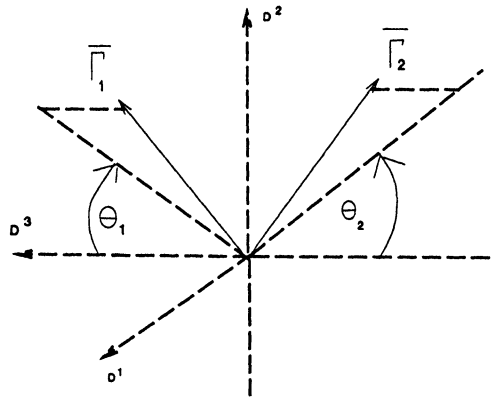


FIG. 3

and

$$\overline{\Gamma}_3 \times \overline{\Gamma}_2 = \begin{pmatrix} 0 \\ \sin\theta_3 \\ \frac{-|u|\cos\theta_3}{\sin\frac{\phi}{2}} \end{pmatrix} \times \begin{pmatrix} 0 \\ \sin\theta_2 \\ \frac{|u|\cos\theta_2}{\sin\frac{\phi}{2}} \end{pmatrix}$$

$$= \frac{|u|}{\sin\frac{\phi}{2}}(\sin\theta_2\cos\theta_3 + \sin\theta_3\cos\theta_2)\mathbf{i} = \frac{|u|\sin(\theta_2+\theta_3)}{\sin\frac{\phi}{2}}\mathbf{i} = D^t u.$$

Now let $\Gamma_j \overset{\text{def}}{=} D\overline{\Gamma}_j$, $j = 2, 3$. It is clear that the first two equations above hold when the $\overline{\Gamma}_j$ are replaced by the Γ_j, since D is orthonormal. It follows that the (γ_j, Γ_j) satisfy the normalization condition (3.2) and by (3.3) the scalar components of the products $(\gamma_3, \Gamma_3)(\gamma_2, \Gamma_2)$ and $(\gamma_3, -\Gamma_3)(\gamma_2, -\Gamma_2)$ are both $\delta_1 = \delta_2$.

Moreover, since $DD^t = E$ and $D(q \times r) = (Dq) \times (Dr)$ for any $q, r \in \mathbb{R}^3$, multiplying the second two equations above by D gives $\gamma_2\Gamma_3 + \gamma_3\Gamma_2 = v$ and $\Gamma_3 \times \Gamma_2 = u$. Adding and subtracting these equations and using (5.1) and (3.3) the vector components of $(\gamma_3, \Gamma_3)(\gamma_2, \Gamma_2)$ and $(\gamma_3, -\Gamma_3)(\gamma_2, -\Gamma_2)$ are $-\Delta_1$ and $-\Delta_2$, respectively.

By the definitions of D and the Γ_j, the formulas for the $\overline{\Gamma}_j$ in (5.4) express the Γ_j as linear combinations of unit vectors in the directions of u and $v \times u$; by (5.1), it is clear that we can obtain these vectors by normalizing $\Delta_2 - \Delta_1$ and $\Delta_2 \times \Delta_1$. The coefficients of these vectors in the formulas for the Γ_j, as well as the formulas for the γ_j, are clear from (5.4), (5.1), and the fact that $|\Delta_1| = \sin\frac{\phi}{2}$. □

Notice that the (γ_j, Γ_j) obtained above are not unique; rather, the set of pairs with the desired properties is a one-parameter family for which the vector components all lie in a given plane (see Fig. 3).

COROLLARY 5.2. *If, for a given choice of endpoints, the* (λ_i, Λ_i) *satisfy the hypotheses on the* Δ_i *in Proposition* 5.1, *then there is a motion in* $\mathcal{S}$ *with these endpoints and one discontinuity in its control.*

Proof. Take $(\delta_i, \Delta_i) = (\lambda_i, -\Lambda_i)$, $i = 1, 2$ in Proposition 5.1. The motion can then be constructed as outlined in Remark 4.3. □

When these hypotheses do not hold, we can apply Proposition 5.1 after some adjustments; we deal with the logistics of this in the next section.

6. The first rotation. We will devote this section to proving the following proposition.

PROPOSITION 6.1. *For any* (λ_i, Λ_i), $i = 1, 2$, *there exists* (γ_1, Γ_1) *such that* $(\delta_1, \Delta_1) \stackrel{\text{def}}{=} (\gamma_1, \Gamma_1)(\lambda_1, -\Lambda_1)$ *and* $(\delta_2, \Delta_2) \stackrel{\text{def}}{=} (\gamma_1, -\Gamma_1)(\lambda_2, -\Lambda_2)$ *are each the identity* $(1, 0)$ *or satisfy the hypotheses of Proposition* 5.1, *i.e.,* $\delta_1 = \delta_2 \in [0, 1)$ *and the* Δ_i *are linearly independent. Moreover,* $\gamma_1 \geq 0$.

For the remainder of this section, we will assume that some fixed choice of the (λ_i, Λ_i) has been made. First note that the condition $\delta_1 = \delta_2$ is not particularly restrictive.

LEMMA 6.2. *Let* $\mathbf{n}$ *be any unit vector in* $\mathbb{R}^3$. *There exists* $\phi \in [0, 2\pi]$ *such that* $\delta_1 = \delta_2$ *when* $(\gamma_1, \Gamma_1) = (\cos \frac{\phi}{2}, \sin \frac{\phi}{2}\mathbf{n})$. *In particular,*

$$\tan \frac{\phi}{2} = \frac{\lambda_2 - \lambda_1}{\langle \Lambda_1, \mathbf{n} \rangle + \langle \Lambda_2, \mathbf{n} \rangle}. \tag{6.1}$$

Proof. Note that for any $\mathbf{n}$ we have $\delta_1 = \cos \frac{\phi}{2}\lambda_1 + \sin \frac{\phi}{2}\langle \Lambda_1, \mathbf{n} \rangle$ and $\delta_2 = \cos \frac{\phi}{2}\lambda_2 - \sin \frac{\phi}{2}\langle \Lambda_2, \mathbf{n} \rangle$. Equating these expressions, dividing the resulting equation by $\cos \frac{\phi}{2}$, and solving for $\tan \frac{\phi}{2}$ gives the result. □

This result tells us that we can obtain (γ_1, Γ_1) explicitly in Proposition 5.1; however, we must choose $\mathbf{n}$ in such a way that $\phi \in [0, \pi]$, the δ_i are nonnegative, and the Δ_i are linearly independent when nonzero. The choice of $\mathbf{n}$ required for these conditions varies, depending on the endpoints.

As it turns out, there is always an appropriate choice of $\mathbf{n}$ for which $\mathbf{n}, \Lambda_1$, and Λ_2 are coplanar. For such $\mathbf{n}$, it is easiest to address these issues under an appropriate rotation of axes. In particular, the action of those quaternions with vector part parallel to $\mathbf{n}$ becomes quite transparent when we rotate axes so that $\mathbf{n}$ lies on a coordinate axis and $\mathbf{n}, \Lambda_1$, and Λ_2 lie in a coordinate plane: Let J be any $SO(3)$ matrix with first column vector parallel to $\mathbf{n}$ and third column vector orthogonal to each Λ_i; then $\mathbf{n}$ and the $\overline{\Lambda}_i \stackrel{\text{def}}{=} J^t\Lambda_i$ have these features.

Let $\overline{\Gamma}_1 \stackrel{\text{def}}{=} J^t\Gamma_1$ and $\overline{\Delta}_i \stackrel{\text{def}}{=} J^t\Delta_i$, $i = 1, 2$. It is easily checked using (3.3) that $(\delta_1, \overline{\Delta}_1) = (\gamma_1, \overline{\Gamma}_1)(\lambda_1, -\overline{\Lambda}_1)$ and similarly $(\delta_2, \overline{\Delta}_2) = (\gamma_1, -\overline{\Gamma}_1)(\lambda_2, -\overline{\Lambda}_2)$. Moreover, since $(\gamma_1, J^t\overline{\Gamma}_1) = (\cos \frac{\phi}{2}, \sin \frac{\phi}{2}\mathbf{i})$, another application of (3.3) shows that

$$\begin{pmatrix} \delta_1 \\ \overline{\Delta}_1 \end{pmatrix} = \begin{pmatrix} \left\langle \frac{\phi}{2} \right\rangle & 0 \\ 0 & \left\langle \frac{\phi}{2} \right\rangle \end{pmatrix} \begin{pmatrix} \lambda_1 \\ -\overline{\Lambda}_1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \delta_2 \\ \overline{\Delta}_2 \end{pmatrix} = \begin{pmatrix} \left\langle -\frac{\phi}{2} \right\rangle & 0 \\ 0 & \left\langle -\frac{\phi}{2} \right\rangle \end{pmatrix} \begin{pmatrix} \lambda_2 \\ -\overline{\Lambda}_2 \end{pmatrix},$$

where $\langle (\cdot) \rangle \in SO(2)$ is the matrix of the rotation by $(\cdot)$.

On the other hand, by the definition of J, for each i we have

$$\begin{pmatrix} \lambda_i \\ -\overline{\Lambda}_i \end{pmatrix} = \begin{pmatrix} \lambda_i \\ -\langle \Lambda_i, \mathbf{n} \rangle \\ \pm \left| \Lambda_i |_{\mathbf{n}^\perp} \right| \\ 0 \end{pmatrix} = \begin{pmatrix} \ell_i \begin{pmatrix} \cos k_i \\ \sin k_i \end{pmatrix} \\ \pm\sqrt{1 - \ell_i^2} \\ 0 \end{pmatrix} \tag{6.2}$$

for suitable $\ell_i \in [0, 1]$ and $k_i \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ (in particular, $\ell_i = \left(\lambda_i^2 + \langle \Lambda_i, \mathbf{n} \rangle^2\right)^{1/2}$ and $k_i = -\tan^{-1}\left[\frac{\langle \Lambda_i, \mathbf{n} \rangle}{\lambda_i}\right]$). This is illustrated in Fig. 4.

It follows that

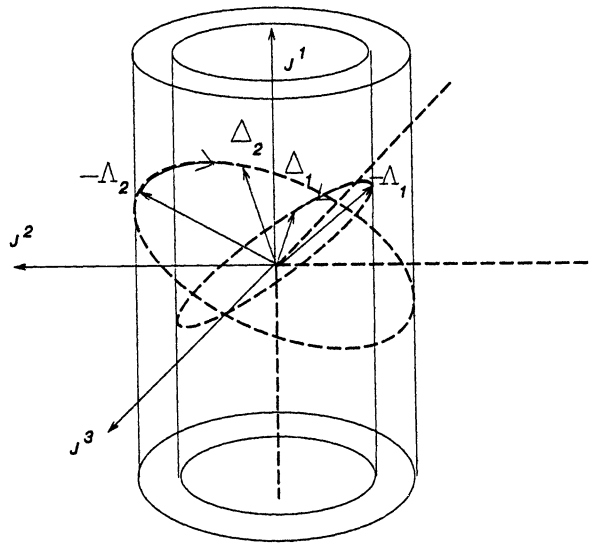


FIG. 4

(6.3)

$$\begin{pmatrix}\delta_1\\ \overline{\Delta}_1\end{pmatrix} = \begin{pmatrix} \ell_1 \begin{pmatrix} \cos\left(k_1+\frac{\phi}{2}\right)\\ \sin\left(k_1+\frac{\phi}{2}\right)\end{pmatrix} \\ \pm\sqrt{1-\ell_1^2}\begin{pmatrix}\cos\frac{\phi}{2}\\ \sin\frac{\phi}{2}\end{pmatrix}\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}\delta_2\\ \overline{\Delta}_2\end{pmatrix} = \begin{pmatrix} \ell_2 \begin{pmatrix} \cos\left(k_2-\frac{\phi}{2}\right)\\ \sin\left(k_2-\frac{\phi}{2}\right)\end{pmatrix} \\ \pm\sqrt{1-\ell_2^2}\begin{pmatrix}\cos\frac{\phi}{2}\\ -\sin\frac{\phi}{2}\end{pmatrix}\end{pmatrix}.$$

This shows that the rotation of axes defined by J "straightens out" the action of $(\gamma_1,\Gamma_1) = (\cos\frac{\phi}{2}, \sin\frac{\phi}{2}\mathbf{n})$ for a given choice of $\mathbf{n}$. Varying ϕ, we obtain the following geometric description of the $(\delta_i,\overline{\Delta}_i)$: The $\overline{\Delta}_i$ rotate about $\mathbf{i}$ on the surfaces of cylinders, their projections on $\mathbf{i}$ varying sinusoidally; correspondingly, the δ_i also vary sinusoidally (this is also a nice geometric interpretation of the exponential map on S^3).

This construction makes it straightforward to prove Proposition 6.1; before doing this we make one more observation.

LEMMA 6.3. *If for some choice of* $\mathbf{n}$ *and some* $\phi' \in [0,\pi]$ *we have* $\operatorname{sgn}(\ell_1\cos(k_1+\frac{\phi'}{2}) - \ell_2\cos(k_2-\frac{\phi'}{2})) = \operatorname{sgn}(\lambda_2-\lambda_1)$, *and if each* $\ell_i \in [0,1)$, *then there is a solution* ϕ *of* (6.1) *for which the conclusions of Proposition* 6.1 *hold with* $(\gamma_1,\Gamma_1) = (\cos\frac{\phi}{2}, \sin\frac{\phi}{2}\mathbf{n})$.

Proof. Since $\lambda_i = \ell_i\cos k_i$ for each i and $\ell_1\cos(k_1+\frac{\phi}{2}) - \ell_2\cos(k_2-\frac{\phi}{2})$ is continuous in ϕ, the first assumption, (6.3), and the intermediate value theorem guarantee that there is a solution ϕ of (6.1) in $(0,\phi')$; by Lemma 6.2, this means that $\delta_1 = \delta_2$ for this choice of ϕ. Moreover, since neither ℓ_i is 1, (6.3) shows that the projections of the $\overline{\Delta}_i$ in $\mathbf{i}^\perp$ are nonzero, and since $\phi \in (0,\pi)$, these vectors are linearly independent, i.e., the $\overline{\Delta}_i$ are linearly independent. It is clear that $\delta_1 > 0$. □

The proof of Proposition 6.1 proceeds in five steps. We begin with the "generic" case.

LEMMA 6.4. *If* $\lambda_1 \neq \lambda_2$, $\lambda_i \in [0,1)$, $i = 1,2$, *and* Λ_1 *is a positive scalar multiple of* Λ_2 *when these vectors are linearly dependent, choose* $\mathbf{n}$ *to satisfy*

$$-\langle \Lambda_i, \mathbf{n}\rangle \quad \text{and} \quad \left|\Lambda_i|_{\mathbf{n}^\perp}\right| > 0, \qquad i = 1,2 \quad \text{if} \quad \lambda_1 > \lambda_2;$$

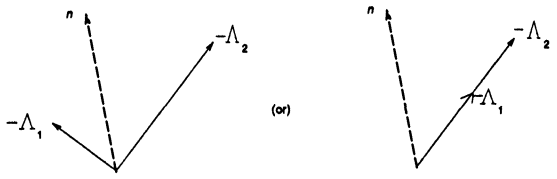


FIG. 5

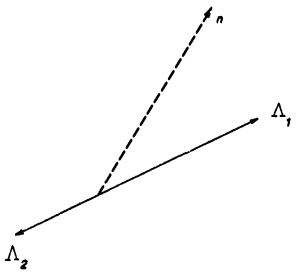


FIG. 6

$$\left\langle \Lambda_i, \mathbf{n} \right\rangle \quad \text{and} \quad \left| \Lambda_i|_{\mathbf{n}^\perp} \right| > 0, \qquad i = 1, 2 \quad \text{if} \quad \lambda_1 < \lambda_2.$$

Then the conclusions of Lemma 6.3 *hold.*

Proof. Suppose $\lambda_1 > \lambda_2$ and choose $\mathbf{n}$ as above (see Fig. 5).

By (6.2), $\ell_i \geq 0$ and $-\langle \Lambda_i, \mathbf{n} \rangle = -\langle \overline{\Lambda}_i, \mathbf{i} \rangle = \ell_i \sin k_i$ and $\left| \Lambda_i|_{\mathbf{n}^\perp} \right| = \left| \overline{\Lambda}_i|_{\mathbf{i}^\perp} \right| = \sqrt{1 - \ell_i^2}$ are both strictly positive. It follows that each $\ell_i \in (0, 1)$ and each $k_i \in (0, \frac{\pi}{2})$.

Let $\phi' \stackrel{\text{def}}{=} \pi - 2k_1$. Then $\ell_1 \cos(\frac{\phi'}{2} + k_1) = 0$ and $\ell_2 \cos(k_2 - \frac{\phi'}{2}) = \ell_2 \cos(k_1 + k_2 - \frac{\pi}{2}) > 0$, so that $\operatorname{sgn}(\lambda_2 - \lambda_1) = \operatorname{sgn}(\ell_1 \cos(k_1 + \frac{\phi}{2}) - \ell_2 \cos(k_2 - \frac{\phi}{2}))$.

The result follows from Lemma 6.3.

The proof is similar when $\lambda_2 > \lambda_1$; with the second choice above of $\mathbf{n}$, the k_i are negative and one takes $\phi' = \pi + 2k_2$. □

The next lemma treats a less generic case.

LEMMA 6.5. *When $\lambda_1 \neq \lambda_2$, each $\lambda_i < 1$, and the Λ_i are negative scalar multiples, let $\mathbf{n}$ be any vector satisfying*

$$\left\langle \Lambda_1, \mathbf{n} \right\rangle \quad \text{and} \quad \left| \Lambda_1|_{\mathbf{n}^\perp} \right| > 0.$$

Then the conclusions of Lemma 6.3 *hold.*

Proof. Choose $\mathbf{n}$ as stated (see Fig. 6).

Since the projection in $\mathbf{n}^\perp$ of Λ_1 is nonzero, $\ell_1 < 1$, and since $\langle \Lambda_1, \mathbf{n} \rangle$ is nonzero, $\ell_1 > 0$, by (6.2). Since the Λ_i are parallel, ℓ_2 is also in $(0, 1)$.

Equation (6.2) also shows that $\langle \Lambda_i, \mathbf{n} \rangle = -\ell_i \sin k_i$ for each i, so that $k_1 \in [-\frac{\pi}{2}, 0)$ and $k_2 \in (0, \frac{\pi}{2}]$ and consequently

$$\ell_1 \cos\left(\frac{\pi}{2} + k_1\right) = -\ell_1 \sin k_1 = \langle \Lambda_1, \mathbf{n} \rangle = \sqrt{1 - \lambda_1^2}\, \langle \mathbf{n}_1, \mathbf{n} \rangle$$

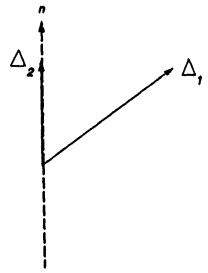


FIG. 7

and

$$\ell_2 \cos\left(k_2 - \frac{\pi}{2}\right) = \ell_2 \sin k_2 = -\langle \Lambda_2, \mathbf{n}\rangle = -\sqrt{1-\lambda_2^2}\,\langle \mathbf{n}_2, \mathbf{n}\rangle\,,$$

where $\mathbf{n}_i \overset{\text{def}}{=} \Lambda_i/|\Lambda_i|$, $i = 1, 2$. It follows that

$$\ell_1 \cos\left(k_1 + \frac{\pi}{2}\right) - \ell_2 \cos\left(k_2 - \frac{\pi}{2}\right) = \langle \mathbf{n}_1, \mathbf{n}\rangle \left(\sqrt{1-\lambda_1^2} - \sqrt{1-\lambda_2^2}\right),$$

since $\langle \mathbf{n}_1, \mathbf{n}\rangle = -\langle \mathbf{n}_2, \mathbf{n}\rangle$. Moreover, it is easy to see that $\operatorname{sgn}\sqrt{1-\lambda_1^2} - \sqrt{1-\lambda_2^2} = \operatorname{sgn}(\lambda_2 - \lambda_1)$, since by assumption each $\lambda_i \in [0, 1]$.

Since the hypotheses of Lemma 6.3 hold, the result now follows. □

The next result deals with another special case.

LEMMA 6.6. *If $\lambda_1 \neq \lambda_2$ and one of the λ_i is 1, choose $\mathbf{n}$ so that*

$$\left\langle \Lambda_1, \mathbf{n}\right\rangle \quad \textit{and} \quad \left|\Lambda_1|_{\mathbf{n}^\perp}\right| > 0 \quad \textit{if} \quad \lambda_2 = 1;$$

and

$$-\left\langle \Lambda_2, \mathbf{n}\right\rangle \quad \textit{and} \quad \left|\Lambda_2|_{\mathbf{n}^\perp}\right| > 0 \quad \textit{if} \quad \lambda_1 = 1.$$

Then there is a solution ϕ of (6.1) in $[0, \pi]$ for which Proposition 6.1 holds with $(\gamma_1, \Gamma_1) = (\cos\frac{\phi}{2}, \sin\frac{\phi}{2}\mathbf{n})$.

Proof. If $\lambda_2 = 1$, make the first choice above of $\mathbf{n}$.

Since $\lambda_2 = 1$, $\ell_2 = 1$ and $k_2 = 0$ by (6.2). Moreover, since $\left|\Lambda_1|_{\mathbf{n}^\perp}\right| > 0$, $\ell_1 \in (0, 1)$, and since $\ell_1 \sin k_1 = -\langle \Lambda_1, \mathbf{n}\rangle < 0$, $k_1 \in [-\frac{\pi}{2}, 0]$. Since $\ell_1 \cos(k_1 + \frac{\pi}{2}) > 0$ and $\ell_2 \cos(k_2 - \frac{\pi}{2}) = 0$, we have $\operatorname{sgn}(\lambda_2 - \lambda_1) = \operatorname{sgn}(\ell_1 \cos(k_1 + \frac{\pi}{2}) - \ell_2 \cos(k_2 - \frac{\pi}{2}))$, so, as in Lemma 6.3, there is a solution $\phi \in (0, \pi)$ of (6.1).

Moreover, by (6.3), we have $\overline{\Delta}_2 = \sin\frac{\phi}{2}\mathbf{i}$ since $\ell_2 = 1$, and the projection of $\overline{\Delta}_1$ in $\mathbf{i}^\perp$ has magnitude $\sqrt{1-\ell_1^2} > 0$, so the $\overline{\Delta}_i$ are linearly independent (see Fig. 7).

The proof is similar for the case $\lambda_1 = 1$. □

The last special case we deal with is the one in which the bodies have the same net rotations and the angles of net rotation are smaller than π.

LEMMA 6.7. *If $\Lambda_1 = \Lambda_2$ and $\lambda_1 = \lambda_2 \in (0, 1)$, let $\mathbf{n}$ be any unit vector in $\Lambda_1^\perp$, ϕ any angle in $(0, \pi)$. Then the conclusions of Lemma 6.3 hold for this choice of ϕ.*

Proof. With the stated choice of $\mathbf{n}$, we have $\langle \Lambda_i, \mathbf{n}\rangle = -\ell_i \sin k_i = 0$ and $\cos k_i = \lambda_i > 0$, by (6.2); therefore, $k_i = 0$ and $\ell_i \in (0, 1)$, $i = 1, 2$ (see Fig. 8). It follows that $\delta_1 = \ell_1 \cos(k_1 + \frac{\phi}{2}) = \ell_2 \cos(k_2 - \frac{\phi}{2}) = \delta_2$ for *any* ϕ. Using (6.3), it is clear that for any $\phi \in (0, \pi)$ the Δ_i are linearly independent. □

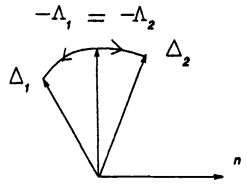


FIG. 8

Proof of Proposition 6.1. Suppose first that the Λ_i are linearly independent. Since $|\Lambda_i|^2 + {\lambda_i}^2 = 1$, we have $\lambda_i < 1$, $i = 1, 2$. If $\lambda_1 \neq \lambda_2$, the hypotheses of Lemma 6.4 hold, and if $\lambda_1 = \lambda_2$, it is clear that Proposition 6.1 holds with $(\gamma_1, \Gamma_1) = (1, 0)$.

Suppose next that the Λ_i are linearly dependent and $\lambda_1 \neq \lambda_2$. If one of the λ_i is 1, we are in the situation of Lemma 6.6. If neither of the λ_i is 1, the Λ_i are either positive scalar multiples, in which case we have the hypotheses of Lemma 6.4, or negative scalar multiples, in which case we are in the situation of Lemma 6.5, so in any case Proposition 6.1 is true.

Finally, suppose the Λ_i are linearly dependent and the λ_i are equal. If the Λ_i are equal with $\lambda_1 \in (0, 1)$, we have the hypotheses of Lemma 6.7. If the Λ_i are equal and $\lambda_1 = 0$, or if the Λ_i are nonzero and negatives of one another, we have the hypotheses of Lemma 4.1. The only remaining case is that when $\lambda_1 = \lambda_2 = 1$, in which case Proposition 6.1 is immediate.

7. Proof of Theorem 4.2. Theorem 4.2 is now easy to prove based on Lemma 4.1, and Propositions 5.1 and 6.1. Given a choice of the net rotations of the bodies, with quaternion representations (λ_i, Λ_i) as in Remark 3.1, it is clear by Lemma 4.1 that we can take $(\gamma_2, \Gamma_2) = (\gamma_3, \Gamma_3) = (1, 0)$ in Theorem 4.2 when the (λ_i, Λ_i) are inverses of each other, and if the (λ_i, Λ_i) satisfy the hypotheses of Proposition 6.1, we can take $(\gamma_1, \Gamma_1) = (1, 0)$ in Theorem 4.2. If neither of these conditions holds, Proposition 6.1 implies that for some (γ_1, Γ_1), $(\delta_1, \Delta_1) = (\gamma_1, \Gamma_1)(\lambda_1, -\Lambda_1)$ and $(\delta_2, \Delta_2) = (\gamma_1, -\Gamma_1)(\lambda_2, -\Lambda_2)$ satisfy the hypotheses of Proposition 5.1. Applying Proposition 5.1, we have (γ_j, Γ_j), $j = 2, 3$, such that

$$(\gamma_3, \Gamma_3)(\gamma_2, \Gamma_2)(\gamma_1, \Gamma_1)(\lambda_1, -\Lambda_1) = (1, 0) \Longleftrightarrow (\gamma_3, \Gamma_3)(\gamma_2, \Gamma_2)(\gamma_1, \Gamma_1) = (\lambda_1, \Lambda_1)$$

and

$$\begin{aligned}(\gamma_3, -\Gamma_3)(\gamma_2, -\Gamma_2)(\gamma_1, -\Gamma_1)(\lambda_2, -\Lambda_2) = (1, 0)\\ \Longleftrightarrow (\gamma_3, -\Gamma_3)(\gamma_2, -\Gamma_2)(\gamma_1, -\Gamma_1) = (\lambda_2, \Lambda_2)\end{aligned}$$

by (3.4). □

In proving this result, we have shown how to construct motions in $\mathcal{S}$ for a given set of endpoints and have given explicit formulas at each step; following the proof of Proposition 6.1, we can always identify which of the zero-, two-, or one-switch situations we are in, and Lemma 4.1, Proposition 5.1, and Lemmas 6.4–6.6 give the formulas for the rotations involved. Once the quaternions are obtained, the actual construction of motions proceeds as in Remarks 4.3 and 2.3.

8. Concluding remarks. The motions obtained here provide simple and intuitive ways of moving between points in $SO(3)^2$ with the imposed constraints. Indeed, a rotation at constant angular velocity is the most fundamental and easily understood motion of a rigid body. By the results of §§5 and 6, the triples of rotations involved are far from being uniquely determined.

In our development we relied on the fact that these motions could be determined algebraically, working on S^3. We can regard the continuous behavior of such a trajectory as being determined by a "shell" of rotations.

Papers treating optimal control problems with this same two body system and other, more complicated, multibody systems have been completed by the author and others since the original submission of this paper. They describe alternative motions of nonholonomic systems like the one described here.

Finally, we mention here that the author has noted generalizations of the results presented here to two body systems with less material symmetry using piecewise constant controls given by the angular momentum of one of the bodies. We hope to report on this at a later date.

9. Appendix: Angle and axis of a rotation. Any $M \in SO(3)$ is the matrix of a rotation through an angle $\phi \in [0, \pi]$ about a unit vector $\mathbf{n} \in \mathbb{R}^3$. The following standard formulas express ϕ and $\mathbf{n}$ in terms of the entries M^{ij} of M when $\sin\phi \neq 0$:

$$\cos\phi = \frac{\mathrm{Tr}(M) - 1}{2},$$

$$\mathbf{n} = \frac{1}{2\sin\phi} \begin{pmatrix} M^{23} - M^{32} \\ M^{31} - M^{13} \\ M^{12} - M^{21} \end{pmatrix}.$$

Derivations of these formulas can be found in, for instance, [1]. As noted in this and other references, the axis $\mathbf{n}$ of the rotation with angle of rotation 0 can be selected arbitrarily. On the other hand, the formula above for $\mathbf{n}$ also becomes undefined when $\phi = \pi$, and the axis of rotation is certainly not indeterminate in this case.

We know of no reference that gives a formula for $\mathbf{n}$ when $\phi = \pi$. We provide one here; for details consult [4]. If $\phi = \pi$, it can be shown that either there is a standard unit vector E^i which is fixed by this rotation, or there are standard unit vectors E^j and E^k that are linearly independent from M^j and M^k, respectively. In the first case, we have

$$\mathbf{n} = E^i,$$

and in the second case we have

$$\mathbf{n} = \pm \frac{(E^j - M^j) \times (E^k - M^k)}{|(E^j - M^j) \times (E^k - M^k)|}.$$

REFERENCES

[1] S. L. ALTMANN, *Rotations, Quaternions, and Double Groups*, Oxford University Press, New York, 1986.

[2] V. I. ARNOL'D, *Mathematical Methods of Classical Mechanics*, 2nd ed., K. Vogtman and A. Weinstein, trans., Springer–Verlag, New York, 1978.

[3] P. E. CROUCH, *Spacecraft attitude control and stabilization: Applications of geometric control theory to rigid body models*, IEEE Trans. Automat. Control, 29 (1986), pp. 321–331.

[4] M. J. ENOS, *Angular Momentum Optimization of Rigid Body Trajectories*, Ph.D. Thesis, Department of Mathematics, Syracuse University, Syracuse, NY, 1990.

[5] R. GROSSMAN, P. S. KRISHNAPRASAD, AND J. E. MARSDEN, *The dynamics of two coupled three-dimensional rigid bodies*, in Dynamical Systems Approaches to Nonlinear Problems in Systems and Circuits, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988.

[6] P. S. KRISHNAPRASAD, *Lie–Poisson structures, dual-spin spacecraft, and asymptotic stability*, Nonlinear Anal. Theory, Methods, Appl., 9 (1985), pp. 1011–1035.

[7] T. R. KANE AND M. P. SCHER, *A dynamical explanation of the falling cat phenemenon*. Internat. J. Solids Structures, 5 (1969), pp. 663–670.

[8] G. W. PATRICK, *The dynamics of two coupled rigid bodies in three space*, in Contemporary Mathematics 97, American Mathematical Society, Providence RI, 1989.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, pp. 1186–1193, July 1994

017

OPTIMAL ANGULAR VELOCITY TRACKING WITH FIXED-ENDPOINT RIGID BODY MOTIONS*

MICHAEL J. ENOS†

Abstract. The problem of finding a fixed-endpoint motion of a rigid body in three space with angular velocity close to a given, arbitrary vector function ω is considered. In particular, if u is the angular velocity of the body in space coordinates, minimizers of $\left\| |u-\omega| \right\|_p$ on an admissible class consisting of smooth rigid body motions on $[0,1]$ with prescribed endpoints are sought for $1 \leq p \leq \infty$. It is shown that, when working in an appropriate moving frame, each of these problems can be formulated as an autonomous problem that can be solved completely in closed form. While this moving frame must in general be obtained numerically, it can be obtained in advance, independently of the solutions; hence all of the extremals for this problem are identified and existence and uniqueness results are obtained.

Key words. rigid body dynamics, nonautonomous optimal control problems, rotation group

AMS subject classifications. 49, 70, 93

1. Introduction.

There has been a renewed interest in recent years in the study of rigid body dynamics from a control point of view. In fact, there are many natural optimal control problems with rigid body motions that have, surprisingly, remained untouched. In spite of much recent development of theoretical machinery for control problems on manifolds, there is (in our opinion) a shortage of examples of nontrivial, specific problems of this type that can be solved completely. Here we describe one such problem that may be regarded as a natural generalization of the classical geodesic problem on $SO(3)$. It is a special case of a more general problem studied in [5].

Consider the motions of any rigid body; ignoring the motion of its center of mass, the configuration space of the body is the rotation group $SO(3)$. Along any smooth motion $A(t)$, we have $\dot{A} = \hat{u}A$ for some unique, skew-symmetric operator $\hat{u}$ (in particular, $\hat{u} = \dot{A}A^t$). Using $SO(3)$ matrices to describe the orientation of the body in space coordinates, $A(t)$ is an $SO(3)$-valued matrix function and $\hat{u}(t)$ is the matrix of the cross product operator on $\mathbb{R}^3$, $(\cdot) \mapsto u \times (\cdot)$. Consequently, the column vectors of A, and the points in the rigid body, instantaneously rotate about u at the angular rate $|u|$, i.e., the vector u is the *angular velocity* of the body in space coordinates.

The transfer to a given state or stabilization about some position of the angular velocity or angular momentum of a rigid body, *without regard to the initial and terminal orientations of the body*, has been studied in several papers (for instance [4], [6], and [3]). In this paper we are concerned with finding a *fixed-endpoint motion* of a rigid body on $[0,1]$ with angular velocity $u(t)$ that is as close as possible, on the average, to a given, *arbitrary* vector $\omega(t)$. Note that imposing the fixed-endpoint condition on A makes it impossible, in general, to exactly match u with ω (indeed, if $u \equiv \omega$ and a left endpoint $A(0)$ is given, then the right endpoint of $A(1)$ is uniquely determined), so this problem generally has a nonzero cost.

A natural measure of "u and ω being close on the average" is given by $\left\| |u-\omega| \right\|_p$,

* Received by the editors February 25, 1991; accepted for publication (in revised form) April 1, 1993. This research was supported by The Ministry of Colleges and Universities of Ontario and The Natural Sciences and Engineering Research Council of Canada, through the Fields Institute for Research in Mathematical Sciences.

† The Fields Institute for Research in Mathematical Sciences, 185 Columbia Street West, Waterloo, Ontario, Canada N2L 5Z5.

$1 \leq p \leq \infty$, and we will consider the following optimal control problem.

PROBLEM 1.1.

$$\left\| |u - \omega| \right\|_p \mapsto \inf;$$

$$\dot{A} = \hat{u}A;$$

$$A(0) = A_0 \quad \textit{and} \quad A(1) = A_1,$$

where $A(t)$ is a path in $SO(3)$, ω is a given path in $\mathbb{R}^3$, and A_0, $A_1 \in SO(3)$ are given.

We will construct and geometrically describe all the extremals for this problem and obtain what amounts to a closed-form solution; in particular, we will obtain an explicit solution in terms of quantities that can be computed independently of the variables A and u. Our main result is Theorem 3.3, which appears at the end of §3. The basic tool we will use is a change of variables that makes this nonautonomous problem equivalent to the classical, autonomous geodesic problem on $SO(3)$. We give a solved example in §4. Also, in the Appendix, we sketch an intuitive, elementary proof of the fact that length minimizing geodesics on $SO(3)$ are continuous rotations with constant angular velocity (i.e., exponentials of constant vectors) that rotate through angles in $[0, \pi]$. In the concluding remarks, we describe a generalization of this problem and mention some other, seemingly fundamental, fixed-endpoint optimal control problems on the rotation group that have not to our knowlege received much attention.

2. Equivalence to an autonomous problem. Here we establish some elementary facts about paths in $SO(3)$ and show that Problem 1.1 is equivalent to an autonomous problem under a suitable rotation of axes.

Throughout, we will assume all quantities are C^∞ smooth for simplicity. All statements may be specialized in an obvious way to the C^r category.

Remark 2.1. We will use the following notation: If $B(t)$ is a path in $SO(3)$, $\omega_B(t)$ denotes the "angular velocity" of B in space coordinates, i.e., the unique vector $\widehat{\omega_B} = \dot{B}B^t$.

The following lemma gives some basic properties of angular velocities.

LEMMA 2.2. *Let B, C, and F be paths in $SO(3)$ and $v(t)$ a path in $\mathbb{R}^3$. Then*

$$F = BC \Longrightarrow \omega_F = \omega_B + B\omega_C, \tag{2.1}$$

$$\omega_{B^t} = -B^t\omega_B. \tag{2.2}$$

Proof. To prove the "product rule" (2.1), note that $B(u \times v) = (Bu) \times (Bv)$ for any $u, v \in \mathbb{R}^3$, so that

$$\dot{F} = \dot{B}C + B\dot{C} = (\hat{\omega}_B B)\, C + B\,(\hat{\omega}_C C) = (\omega_B \widehat{+} B\omega_C)BC = (\omega_B \widehat{+} B\omega_C)F.$$

This implies (2.2):

$$0 = \frac{d}{dt}(B^tB) = (\omega_{B^t} \widehat{+} B^t\omega_B). \qquad \square$$

Now pick any $C_0 \in SO(3)$ and consider the differential equation

$$\dot{C} = \hat{\omega} C, \qquad C(0) = C_0, \tag{2.3}$$

where ω is as in Problem 1.1. The solution $C(t)$ is, clearly, a path in $SO(3)$.

PROPOSITION 2.3. *In Problem* 1.1, *let* $B \stackrel{\text{def}}{=} C^t A$, $v \stackrel{\text{def}}{=} C^t(u-\omega)$, $B_0 \stackrel{\text{def}}{=} C^t(0)A_0$, *and* $B_1 \stackrel{\text{def}}{=} C^t(1)A_1$. *Then Problem* 1.1 *is equivalent to the following problem.*

PROBLEM 2.4.

$$\left\| |v| \right\|_p \mapsto \inf;$$

$$\dot{B} = \hat{v} B;$$

$$B(0) = B_0 \quad \textit{and} \quad B(1) = B_1,$$

over paths B *in* $SO(3)$.

Proof. First note that

$$B(0) = B_0 \iff (C^t A)(0) = C^t(0)A(0) = C^t(0)A_0 \iff A(0) = A_0.$$

Similarly, $B(1) = B_1 \iff A(1) = A_1$.

It is clear that $v = \omega_B$; also, the definition of B, (2.1), and (2.2) show that

$$v = \omega_{C^t A} = \omega_{C^t} + C^t \omega_A = -C^t \omega + C^t u = C^t(u - \omega),$$

so that

$$\left\| |u - \omega| \right\|_p = \left\| |C^t(u - \omega)| \right\|_p = \left\| |v| \right\|_p. \qquad \square$$

Note that $C^t v$ and $C^t A$ are the coordinates of the vector v and the rigid body motion A observed from a coordinate system rotating with angular velocity ω. Also, Problem 2.4 is *autonomous*, so that this rotation of axes significantly simplifies the problem. Moreover, while C generally involves the numerical solution of a nonautonomous differential equation, we can compute C *in advance*, i.e., before formulating the problem. Finally, note that when $p = 2$, Problem 2.4 is just a control formulation of the classical problem of finding the free motions of a completely symmetric rigid body, since the kinetic energy of a symmetric body is a constant multiple of the square of the magnitude of its angular velocity.

3. Solutions of Problem 1.1. Assume until further notice that $p < \infty$. Since $\left\| |v| \right\|_p = \left(\int_0^1 |v|^p dt \right)^{1/p}$, we want to minimize $\int_0^1 |v|^p dt$ in Problem 2.4. Note that $|v|^p = \langle v, v \rangle^{p/2}$ defines a Riemannian metric on $TSO(3)$, so the fact that $SO(3)$ is a Lie group implies there is a path of minimal length in this metric between any two of its points, i.e., we are assured of the existence of a solution.

By standard methods, for instance, the method of symplectic reduction [2] or an elementary application of the Pontryagin Maximum Principle, the extremals for Problem 2.4 when $p > 1$ are paths of the form $B(t) = \exp(t\hat{v})$ with v constant, i.e., continuous rotations with constant angular velocity. The extremals for $p = 1$ are the same up to a time reparametrization, i.e., with the rotation angle $t|v|$ replaced by a

nondecreasing function $g(t)$ satisfying $g(0) = 0$. This follows from the case $p > 1$ and the familiar fact that the integral $2\int_0^1 |v|dt = \int_0^1 |\hat{v}|dt = \int_0^1 |\dot{A}|dt$ is invariant under time reparametrizations of $A(t)$.

LEMMA 3.1. *Let $R(\phi\mathbf{n}) \in SO(3)$ denote the rotation through the positive angle ϕ about the unit vector $\mathbf{n}$. When $p > 1$, the extremals of Problem 2.4 are of the following form:*

$$B(t) = R(t\phi\mathbf{n})B_0, \quad \text{where } \phi\mathbf{n} = v = \text{constant}. \tag{3.1}$$

When $p = 1$, the extremals are of the following form:

$$B(t) = R(g(t)\mathbf{n})B_0, \quad \text{where } g' \geq 0,\ g(0) = 0 \text{ and } v = g'(t)\mathbf{n}\ (\mathbf{n} \text{ const.}). \tag{3.2}$$

It is straightforward to classify the extremals for Problem 2.4 for given endpoints B_0, B_1. Setting $t = 1$ in $B(t)$, we require

$$R(\phi\mathbf{n}) = B_1 B_0^t \stackrel{\text{def}}{=} D, \tag{3.3}$$

where we define ϕ to be $g(1)$ in the case $p = 1$.

In the general case, i.e., when $B_1 \neq B_0$, the set of pairs $(\phi, \mathbf{n})$ with this property is countable, i.e., we have countably many extremals for Problem 2.4 (regarded as continuous rotations) with the endpoints B_0, B_1. More specifically, note that the rotation D can be expressed as $R(\phi_*\mathbf{n}_*)$ for an appropriate unit vector $\mathbf{n}_*$ and angle $\phi_* \in [0, \pi]$. The angle ϕ_* is uniquely determined. $\mathbf{n}_*$ is uniquely determined when $\phi_* \in (0, \pi)$; can be selected arbitrarily when $\phi_* = 0$ (any rotation through the angle 0 is the identity); and is determined up to multiplication by ± 1 when $\phi_* = \pi$ (the rotations in either direction about a given vector through the angle π are identical). While explicit formulas for ϕ_* and $\mathbf{n}_*$ in terms of the matrix entries of D are "standard," the formulas found in many references do not work when $\phi_* = \pi$; see [5] for a formula that works in this case.

The set of *all* angle/axis pairs satisfying (3.3) can be obtained by adding nonnegative multiples of 2π to ϕ_* and rotating about $\mathbf{n}_*$, or by adding nonnegative multiples of 2π to $2\pi - \phi_*$ and rotating about $-\mathbf{n}_*$. In other words, *the extremals for Problem 2.4 are all continuous rotations of B_0 about the line containing $\mathbf{n}_*$, and are distinguished by the number of times and direction in which they rotate.* See Fig. 1.

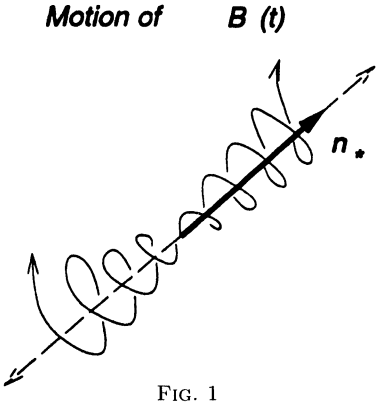


FIG. 1

It is easy to determine the minimizers for Problem 2.4. If $B = R(t\phi\mathbf{n})B_0$ is an extremal and $p > 1$, we have

$$\left\| |v| \right\|_p = \left(\int_0^1 |v|^p dt \right)^{1/p} = \left(\int_0^1 |\phi\mathbf{n}|^p dt \right)^{1/p} = \phi,$$

and it follows that the extremals with lowest cost are the ones with $(\phi, \mathbf{n}) = (\phi_*, \mathbf{n}_*)$. Similarly, when $p = 1$, we take $\mathbf{n} = \mathbf{n}_*$ and $g(t)$ to satisfy $g(1) = \phi_*$.

Finally, since $\left\| |v| \right\|_\infty = \lim_{p\to\infty} \left\| |v| \right\|_p$, it is clear that the minimizers for the case $1 < p < \infty$ are also minimizers for the case $p = \infty$. An elementary argument shows that the same uniqueness results are true here. If v is the angular velocity of another admissible path and $\left\| |v| \right\|_\infty = \phi_*$, then $\left\| |v| \right\|_2 \leq \phi_*$ (otherwise, $|v| > \phi_*$ on an interval, so that $\left\| |v| \right\|_\infty > \phi_*$). Since $|v|(t) \leq \phi_*$ and $\left\| |v| \right\|_2 = \phi_*$, we infer that $|v| \equiv \phi_*$, i.e., the motion corresponding to v is a solution of Problem 2.4 for $p = 2$.

Summarizing the results of this section, we obtain the following solutions of Problem 2.4.

PROPOSITION 3.2. *If $p > 1$, the solutions of Problem 2.4 are the extremals of the form $B_*(t) = R(t\phi_*\mathbf{n}_*)B_0$, where $R(\phi_*\mathbf{n}_*) = B_1 B_0^t$ and $\phi_* \in [0, \pi]$; there is a unique solution when $\phi_* \in [0, \pi)$ and there are two solutions when $\phi_* = \pi$.*

When $p = 1$, the solutions of Problem 2.4 are of the form $R(g(t)\mathbf{n}_*)B_0$ with $g' \geq 0$, $g(0) = 0$, $g(1) = \phi_*$, and $\mathbf{n} = \mathbf{n}_*$.

Combining this with the results of §2, we have the following solutions of Problem 1.1.

THEOREM 3.3. *The solutions of Problem 1.1 are those motions A of the form $A = CB$, where B is a solution of Problem 2.4 as given in Proposition 3.2, C is any path in $SO(3)$ satisfying $\dot{C}C^t = \hat{\omega}$, $B_0 = C^t(0)A_0$, and $B_1 = C^t(1)A_1$.*

4. An example. Consider Problem 1.1 in the situation where the vector ω is of constant magnitude and rotating about a fixed axis at a fixed rate, i.e.,

$$\omega = \rho \begin{pmatrix} \cos\beta \\ \sin\beta\cos t\gamma \\ \sin\beta\sin t\gamma \end{pmatrix}.$$

To obtain solutions of Problem 1.1, Theorem 3.3 tells us to construct a rotating frame C with angular velocity ω and, working in this frame, find the solutions of Problem 2.4.

To construct C, note that the frame $F \overset{\text{def}}{=} R(t\gamma\mathbf{i})$ has angular velocity $\gamma\mathbf{i}$ and that

$$F^t\omega = \rho \begin{pmatrix} \cos\beta \\ \sin\beta \\ 0 \end{pmatrix},$$

so that by (2.1), any frame C with angular velocity ω satisfies

$$\omega_{F^tC} = -F^t\omega_F + F^t\omega = \begin{pmatrix} \rho\cos\beta - \gamma \\ \rho\sin\beta \\ 0 \end{pmatrix} \overset{\text{def}}{=} \Omega.$$

Ω being constant, we can take $F^tC = R(t\Omega)$, so that $C = FR(t\Omega) = R(t\gamma\mathbf{i})R(t\Omega)$. It is easily seen that $B_0 = C^t(0)A_0 = A_0$ and $B_1 = C^t(1)A_1 = R(-\Omega)R(-\gamma\mathbf{i})A_1$. If

$B_0 \neq B_1$, the solutions $B(t)$ of Problem 2.4 rotate about a fixed axis between these positions as in Proposition 3.2, i.e.,

$$B(t) = \begin{cases} R(t\phi_*\mathbf{n}_*)A_0, & p > 1; \\ R(g(t)\mathbf{n}_*)A_0, & p = 1, \end{cases}$$

where $R(\phi_*\mathbf{n}_*) = B_1B_0^t = R(-\gamma\mathbf{i})R(-\Omega)A_1A_0^t$ and $\phi_* \in [0, \pi]$, and $g(t)$ is as in Proposition 3.2. By Theorem 3.3, the solutions $A(t)$ of Problem 1.1 are given by products of exponentials

$$A(t) = \begin{cases} R(t\Omega)R(t\gamma\mathbf{i})R(t\phi_*\mathbf{n}_*)A_0, & p > 1; \\ R(t\Omega)R(t\gamma\mathbf{i})R(g(t)\mathbf{n}_*)A_0, & p = 1, \end{cases}$$

where $g(t), \phi_*$, and $\mathbf{n}_*$ are as described in the last section. Intuitively, the optimal motions rotate about $\mathbf{n}_*$, as $\mathbf{n}_*$ rotates about the vector Ω, as Ω rotates about $\mathbf{i}$; see Fig. 2.

The solutions of Problem 1.1 are qualitatively of this form unless the endpoint conditions A_0, A_1 are completely adapted to the path C, i.e., the net rotation $A_1A_0^t$ is identical to $C(1)C^t(0)$, in which case C^tA is constant, the cost is identically zero, and $A(t)$ is a product of only two exponentials.

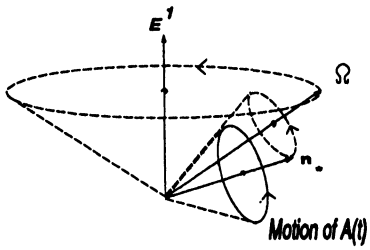


FIG. 2

It is interesting to note that, when $p = 1$ and $B(t)$ is not constant, the solutions of Problem 2.4 (distinguished by the choice of $g(t)$) are all reparametrizations of the same geometric path (the usual situation with geodesics), while this is definitely *not* the case in the context of Problem 1.1: *Different choices of $g(t)$ lead to very different-looking solutions relative to a fixed space frame*, due to the nonautonomous nature of the problem. For instance, the solutions with $g(t) \equiv t\phi_*$ and $g(t) = 0$ on $[0, \frac{1}{2}]$, $g(t) = 2(t - \frac{1}{2}\phi_*)$ on $[\frac{1}{2}, 1]$ follow quite different geometric paths in space coordinates.

5. Concluding remarks. The basic idea we used in solving Problem 1.1 is that of changing variables to autonomize the problem, essentially working around the time-dependent quantities involved. Here we were able to accomplish this by simply changing our point of view with a time-dependent rotation of axes. This philosophy can be of use in other nonautonomous problems.

The actual computation of minimizers in §3 illustrates some of the difficulties involved in obtaining complete solutions of optimal control problems on manifolds, namely, we generally need to use information about the geometry of state space. For more complicated rigid body problems, the use of quaternion geometry can frequently be helpful in this context.

A more general problem of this type is that of optimal, fixed-endpoint tracking of angular *momentum* with a fixed-endpoint motion in $SO(3)$, i.e., of minimizing $\left\| |Iu - h| \right\|_p$, where $I = ADA^t$, with D a constant, positive, and diagonal matrix, is the inertia tensor of a general rigid body. Problem 1.1 is the specialization of this problem to the case of a spherically symmetric body. This problem is treated in [5]. For this new problem, the moving frame construction we used above is not applicable; however, it can be shown that the extremals satisfy the equation

$$\frac{d}{dt}\Big(I(Iu - h)\Big) = h \times (Iu).$$

The system consisting of this equation and $\dot{A} = \hat{u}A$ is integrable by quadratures when h is constant and the body is axisymmetric (i.e., when the inertia tensor I has two equal eigenvalues) and the solution can be expressed in terms of elliptic integrals. The solutions in this case are not unlike the motions of a Lagrange top.

Another problem related to Problem 1.1 is that of minimizing the integral of a function which is small when $A(t)$ is close to a given path $C(t)$ in $SO(3)$, for instance, minimizing $\int_0^1 \Big\langle A - C, A - C \Big\rangle dt$ or, equivalently, $\int_0^1 -\Big\langle A, C \Big\rangle dt$. In some sense, when we solve Problem 1.1 are also trying to make the motions A and C close, if we choose C to satisfy $\omega_C = \omega$ and $C(0) = A_0$, since it is clear that $A \longrightarrow C$ as $u \longrightarrow \omega$. Yet the solutions are different (for instance the solutions of this new problem are generally nonsmooth), and the relationship between these problems might be interesting to investigate.

Other problems with single rigid bodies that to our knowledge have not been treated include the problem of time-optimal transfer of a general rigid body between two orientations with bounded torques, and the torque-optimal transfer of a general rigid body between two orientations on a fixed-time interval.

Other interesting problems can be obtained by adapting some of these ideas to nonholonomic mechanical systems which have recently appeared in the literature, for instance freely rotating systems of coupled rigid bodies.

6. Appendix: Direct derivation of length minimizers on $SO(3)$. We can alternatively give a proof that the minimizers of Problem 2.4 are given by Proposition 3.2 using only some basic geometric observations about $SO(3)$ and the fact that any smooth path in $SO(3)$ looks locally like the exponential of a constant vector. We briefly sketch this proof here for $p = 1$, the case $p > 1$ being similar.

As noted in §3, any rotation in $SO(3)$ can be represented as $R(\phi\mathbf{n})$ with $\phi \in [0, \pi]$. We will assume in this section that any such representation of a rotation is a standard one, with $\phi \in [0, \pi]$.

The product of two rotations $R(\phi_1\mathbf{n}_1)R(\phi_2\mathbf{n}_2)$ is a third rotation $R(\phi_3\mathbf{n}_3)$. Given $\phi_i, \mathbf{n}_i$, $i = 1, 2$, we can use formulas due to Rodrigues (see [1]) to write an expression for ϕ_3 and $\mathbf{n}_3$. Here we are only interested in the formula for ϕ_3, which is

$$\cos\frac{\phi_3}{2} = \cos\frac{\phi_1}{2}\cos\frac{\phi_2}{2} - \sin\frac{\phi_1}{2}\sin\frac{\phi_2}{2}\Big\langle \mathbf{n}_1, \mathbf{n}_2 \Big\rangle.$$

It is evident from this equation that

$$\phi_1 + \phi_2 \geq \phi_3, \tag{6.1}$$

with equality if and only if $\mathbf{n}_1 = \mathbf{n}_2$.

Let $B_*(t) = R(t\phi\mathbf{n})B_0$, where $R(\phi\mathbf{n}) = B^1B_0^t$. The statement that $B_*(t)$ is a solution of Problem 2.4 ($p = 1$) is that

$$\phi \leq \int_0^1 |v|dt \tag{6.2}$$

for any path $B(t)$ in $SO(3)$ satisfying $\dot{B} = \hat{v}B$, $B(0) = B_0$, and $B(1) = B_1$.

The idea of our proof is to approximate the path $B(t)$ with "polygonal lines," where we regard a "line" as a continuous rotation at constant angular velocity; in particular, define a sequence of paths $\{D_i\}_{i=1}^{\infty}$ as follows: Let $D_0 = B_*$, let D_1 be the path consisting of a rotation at constant angular velocity from $B(0)$ to $B(\frac{1}{2})$ on $[0, \frac{1}{2}]$ followed by a constant angular velocity path from $B(\frac{1}{2})$ to $B(1)$ on $[\frac{1}{2}, 1]$, and so on, on successive partitions of $[0, 1]$; see Fig. 3. Using matrix coordinates, it can be shown that the sequence (D_i) converges uniformly to B and that the sequence of derivatives $(\dot{D}_i)$ converges uniformly to $\dot{B}$. Consequently, since $\hat{v} = \dot{B}B^t$, the sequence (ω_{D_i}) of angular velocities converges uniformly to v, so that $\left\| |\omega_{D_i}| \right\|_1$ converges to $\left\| |v| \right\|_1$.

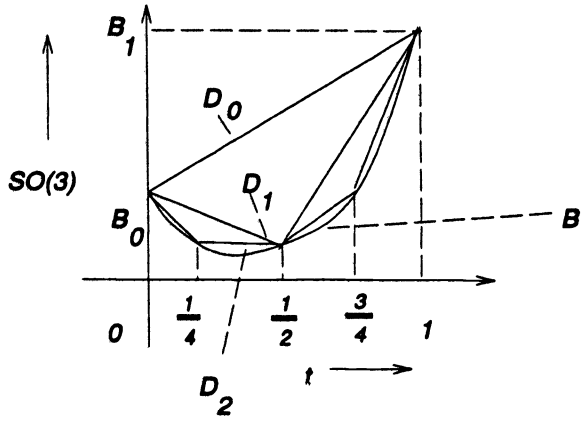


FIG. 3

Now, it is easy to show that the cost on each interval on which D_i rotates about a fixed axis is the angle rotated through, so the cost along D_i is the sum of the angles rotated through on the intervals of smoothness of D_i. Hence, by the inequality (6.1), the cost along D_1 exceeds that along D_0; a similar argument shows that the cost of D_i exceeds the cost of D_{i-1} for all i, i.e., the cost must increase in i, so that (6.2) holds.

The idea of "polygonal approximation" with geodesics that we used here can also be of use in direct sufficiency proofs for other problems.

REFERENCES

[1] S. ALTMANN, *Rotations, Quaternions, and Double Groups*, Oxford University Press, 1986.

[2] R. ABRAHAM AND J. MARSDEN, *Foundations of Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1990.

[3] A. BLOCH, P.S. KRISHNAPRASAD, J. MARSDEN, AND G. SANCHEZ DE ALVAREZ, *Stabilization of rigid body dynamics by internal and external torques*, Automatica, 28 (1992), pp. 745–756.

[4] T. DWYER, *The control of angular momentum of asymmetric rigid bodies*, IEEE Trans. Automat. Control, 27 (1982), pp. 686–688.

[5] M. J. ENOS, *Angular Momentum Optimization of Rigid Body Trajectories*, Ph.D. Thesis, Department of Mathematics, Syracuse University, Syracuse, NY, 1990.

[6] E. SONTAG AND H. SUSSMANN, *Further comments on the stabilizability of angular velocity of a rigid body*, Systems Control Lett., 12 (1987), pp. 213–217.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 4, p. 1194, July 1994

018

ERRATUM: ON THE OPTIMAL TRACKING PROBLEM*

OFER ZEITOUNI† AND MOSHE ZAKAI†

The proof of Lemma 3.2 in [1] contains an error. In contrast to what is written, (3.17) does not seem to follow from [2], and is probably incorrect.

The Lemma itself does hold. In the proof, we must replace the upper bound in (3.17) by the claim that there exists a constant C such that

$$\frac{P(||\tilde{\mu}||_T > c)}{\exp -(hx^2/2\epsilon)} \leq C/\epsilon^2, \tag{1}$$

which is more than enough to conclude.

To see (1), observe that we may represent $\tilde{\mu}_t$ as $\tilde{\mu}_t = e^{-\alpha t}(\nu_{e^{2\alpha t}} - \nu_1)/\sqrt{\alpha}$, where $\alpha = h/\epsilon$, ν_t is a standard Brownian motion and the equality is in law. Therefore, letting $\Delta = 1/\alpha^2$,

$$\begin{aligned}
P(||\tilde{\mu}||_T > c) &\leq 2P\left(\sup_{0\leq t\leq T} \frac{e^{-\alpha t}(\nu_{e^{2\alpha t}} - \nu_1)}{\sqrt{\alpha}} > c\right) \\
&\leq 2\sum_{i=1}^{\lceil 1/\Delta\rceil} P\left(\sup_{(i-1)\Delta T\leq t\leq i\Delta T} \frac{e^{-\alpha t}(\nu_{e^{2\alpha t}} - \nu_1)}{\sqrt{\alpha}} > c\right) \\
&\leq 2\sum_{i=1}^{\lceil 1/\Delta\rceil} P\left(\sup_{0\leq t\leq i\Delta T} \frac{(\nu_{e^{2\alpha t}} - \nu_1)}{\sqrt{\alpha}} > ce^{\alpha(i-1)\Delta T}\right) \\
&\leq 4\sum_{i=1}^{\lceil 1/\Delta\rceil} P(\nu_{e^{2\alpha i\Delta T}} \geq c\sqrt{\alpha}e^{\alpha(i-1)\Delta T}) \\
&\leq \frac{C_1}{\Delta}\exp(-\alpha c^2 e^{-2\alpha\Delta T}/2).
\end{aligned}$$

Inequality (1) follows by noting that by our choice, $\alpha\Delta \to_{\epsilon\to 0} 0$, while $\alpha^2\Delta = 1$.

Acknowledgment. We thank Ehud Barak for pointing out the mistake in the proof of (3.17).

REFERENCES

[1] O. ZEITOUNI AND M. ZAKAI, *On the optimal tracking problem*, SIAM J. Control Optim., 30 (1992), pp. 426–439.

[2] M. TALAGRAND, *Small tails for the supremum of a Gaussian process*, Ann. Inst. H. Poincare, 24 (1988), pp. 307–315.

* Received by the editors January 23, 1994; accepted for publication February 1, 1994.

† Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1195–1225, September 1994

STABILITY OF RECURSIVE STOCHASTIC TRACKING ALGORITHMS*

LEI GUO†

Abstract. First, the paper gives a stability study for the random linear equation $x_{n+1} = (I - A_n)x_n$. It is shown that for a quite general class of random matrices $\{A_n\}$ of interest, the stability of such a vector equation can be guaranteed by that of a corresponding scalar linear equation, for which various results are given without requiring stationary or mixing conditions. Then, these results are applied to the main topic of the paper, i.e., to the estimation of time varying parameters in linear stochastic systems, giving a unified stability condition for various tracking algorithms including the standard Kalman filter, least mean squares, and least squares with forgetting factor.

Key words. stochastic systems, adaptive systems, parameter estimation, tracking algorithms, time varying, stability, excitation

AMS subject classifications. 93C40, 93E12, 93E10

1. Introduction.

An important issue in system identification, signal processing, adaptive control and many other fields is whether the algorithms designed possess some tracking capabilities when the system parameters (or signals) to be estimated are changing with time. The basic time-varying model is that of a linear regression:

$$y_k = \varphi_k^\tau \theta_k + v_k, \qquad k \geq 0 \tag{1.1}$$

where y_k and v_k are the scalar observation and noise, respectively, and φ_k and θ_k are, respectively, the d-dimensional stochastic regressor and the unknown time-varying parameter. It is usually convenient to denote the parameter variation at instant k by Δ_k:

$$\Delta_k \triangleq \theta_k - \theta_{k-1}, \qquad k \geq 1. \tag{1.2}$$

It is well known that many problems from different application areas can be cast in the form (1.1) (see e.g., [1], [2]), and a variety of recursive algorithms have been derived for tracking the unknown parameters θ_k. These algorithms are basically of the following form:

$$\hat{\theta}_{k+1} = \hat{\theta}_k + L_k(y_k - \varphi_k^\tau \hat{\theta}_k) \tag{1.3}$$

where L_k is the adaptation gain that can be chosen in a number of ways (see e.g., [1]–[3]). In the present time-varying case, a common feature of the gain L_k is that it does not tend to zero as the time k goes to infinity. This is very natural from an intuitive point of view. When the system parameters are time-varying, the algorithm must be persistently alert to follow the parameter variations. Here we illustrate three choices of L_k that correspond to three standard algorithms.

Kalman filtering (KF) algorithm.

$$L_k = \frac{P_k \varphi_k}{R + \varphi_k^\tau P_k \varphi_k} \tag{1.4}$$

$$P_{k+1} = P_k - \frac{P_k \varphi_k \varphi_k^\tau P_k}{R + \varphi_k^\tau P_k \varphi_k} + Q, \tag{1.5}$$

* Received by the editors February 3, 1992; accepted for publication February 25, 1993. This work was supported by the National Natural Science Foundation of China.

† Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China.

where $P_0 \geq 0, R > 0, Q > 0$ and $\hat{\theta}_0$ are deterministic and can be arbitrarily chosen. Here R and Q may be regarded as the a priori estimates for the variances of v_k and Δ_k, respectively. Taking R and Q as constants is just for simplicity of discussion, and generalizations to time-varying cases are straightforward.

It is well known that (see e.g., [4, Chap. 13] and [5, Chap. 3]) if φ_k is $\mathcal{F}_{k-1}$ measurable, where $\mathcal{F}_k \triangleq \sigma\{y_i, i \leq k\}$, and if $\{\Delta_k, v_k\}$ is a Gaussian white noise process, then θ_k generated by (1.3)–(1.5) is the minimum variance estimate for θ_k, and P_k is the estimation error covariance, i.e.,

$$\hat{\theta}_k = E[\theta_k|\mathcal{F}_{k-1}], \qquad P_k = E[\tilde{\theta}_k\tilde{\theta}_k^\tau|\mathcal{F}_{k-1}] \tag{1.6}$$

provided that $Q = E\Delta_k\Delta_k^\tau$, $R = Ev_k^2$, $\hat{\theta}_0 = E\theta_0$ and $P_0 = E[\tilde{\theta}_0\tilde{\theta}_0^\tau]$, where $\tilde{\theta}_k$ is the estimation error

$$\tilde{\theta}_k = \theta_k - \hat{\theta}_k, \tag{1.7}$$

which is of prime interest to us.

Least mean squares (LMS) algorithm.

$$L_k = \mu\frac{\varphi_k}{1 + \|\varphi_k\|^2}, \tag{1.8}$$

where $\mu \in (0, 1]$ is called the step size or adaptation rate. Such an algorithm is also referred to as a gradient algorithm because the increment of the algorithm (1.3) and (1.8) is opposite to the (stochastic) gradient of the mean square error

$$e_k(\theta) = E(y_k - \varphi_k^\tau\theta)^2.$$

Thus, it is a type of steepest descent algorithm that aims at minimizing $e_k(\theta)$ recursively.

Recursive least squares (RLS) algorithm.

$$L_k = \frac{P_k\varphi_k}{\alpha + \varphi_k^\tau P_k\varphi_k} \tag{1.9}$$

$$P_{k+1} = \frac{1}{\alpha}\left[P_k - \frac{P_k\varphi_k\varphi_k^\tau P_k}{\alpha + \varphi_k^\tau P_k\varphi_k}\right], \tag{1.10}$$

where $P_0 > 0$, and $\alpha \in (0, 1)$ is a forgetting factor. This algorithm is derived by minimizing the following criterion over $\theta \in \mathbb{R}^d$:

$$V_k(\theta) = \frac{1}{k}\sum_{i=0}^{k}\alpha^{k-i}(y_i - \theta^\tau\varphi_i)^2 \tag{1.11}$$

(see e.g., [1], pp. 57–58). Note that in (1.11) old measurements are exponentially discounted, and so the estimate is expected to be representative for the current properties of the system.

All of the above-mentioned algorithms are well known and widely used in applications. The KF algorithm is attractive due to the fact that it generates the conditional expectation of the unknown parameter given the past measurements in the ideal case (see (1.6)). The LMS has been used in many applications, mainly because of its simplicity for implementation. The advantage of the RLS algorithm over LMS is that it generates more accurate estimates in the

transient phase (see e.g., [6]). In many cases, the RLS algorithm is optimal in the sense that it minimizes the criterion (1.11), while for the KF algorithm, it is not known if it is still optimal in some sense when the Gaussian assumption fails and the covariances of v_k and Δ_k are not available.

There is a vast literature on the analysis of algorithms of type (1.3). In the area of adaptive signal processing, the LMS algorithm has received a great deal of attention (see e.g., [7]–[12]). Most of the existing analysis require that the signals $\{y_k, \theta_k, \varphi_k\}$ possess some sort of stationarity, independence, or mixing properties. The KF algorithm has also attracted much research attention (e.g., [11], [13]–[15]). The first rigorous stability analysis for KF that allows $\{\varphi_k\}$ to be a large class of stochastic regressors seems to be that in [14]. Finally, for the RLS algorithm, we mention the preliminary works in [6], [16], [17], among many others.

In the related area of stochastic adaptive control, the Kalman filter was used by Meyn and Caines [31] to design the adaptive control law for a first-order stochastic system. By applying the Markov chain ergodic theory, they obtained the first concrete adaptive control result for systems with nontrivial (random) parameter variations. For high-order systems with randomly varying parameters, stability of an LMS-based adaptive minimum variance controller was demonstrated in [30]. Similar results were recently established in [32] for a KF-based model reference adaptive controller. However, the parameter tracking properties of the estimation algorithms are not studied in these papers.

In this paper, we first present a series of stability results on the vector random linear equation $x_{n+1} = (I - A_n)x_n$, where $\{A_n\}$ is a sequence of random matrices of the same dimension, which may not satisfy the usual stationary or mixing conditions. The key observation is that for a variety of $\{A_n\}$ of interest, the stability study of the vector linear equation may be reduced to that of a relatively simple scalar equation. Then we present a stability/excitation condition for recursive stochastic tracking algorithms and establish upper bounds for the tracking error.

The main contributions of the paper are as follows:

(i) The new stability condition is the weakest known and a unified one for the three standard algorithms mentioned above. This is important since establishing stability is known to be a crucial step for any further studies (see e.g., [18]).

(ii) For a large class of random models of interest in applications including time-varying autoregressive models, we can verify the present condition, whereas conditions introduced previously (see e.g., [14], [28]) cannot be verified;

(iii) For the commonly used ϕ-mixing process, we can prove that our stability condition is also a necessary one in some sense.

2. Stability of random equation $x_{n+1} = (I - A_n)x_n$.

2.1. Preliminaries. To begin, by substituting (1.1) into (1.3) and using the notations (1.2) and (1.7), we get the following error equation:

$$\tilde{\theta}_{k+1} = (I - L_k\varphi_k^\tau)\tilde{\theta}_k - L_k v_k + \Delta_{k+1}, \qquad k \geq 0. \tag{2.1}$$

Clearly, this equation falls into the following general form of linear equations:

$$x_{k+1} = (I - A_k)x_k + \xi_{k+1}, \qquad k \geq 0 \tag{2.2}$$

where $\{A_k\}$ is a sequence of $d \times d$ random matrices, and $\{\xi_{k+1}\}$ represents the disturbance. Usually, we are primarily interested in the following problem: does $\{x_k\}$ remain bounded in some sense when $\{\xi_k\}$ belongs to a certain class of random processes? To rigorously study this problem, we need to introduce some notations and definitions.

For any matrix X, its norm is defined as its maximum singular value, i.e. $\|X\| = \{\lambda_{\max}(XX^\tau)\}^{1/2}$.

DEFINITION 2.1. A random matrix (or vector) sequence $\{A_k, k \geq 0\}$ defined on the basic probability space $(\Omega, \mathcal{F}, P)$ is called L_p-stable $(p > 0)$ if $\sup_{k\geq 0} E\|A_k\|^p < \infty$.

In the sequel, we will refer to $\|A_k\|_{L_p}$ defined by

$$\|A_k\|_{L_p} \triangleq \{E\|A_k\|^p\}^{1/p} \tag{2.3}$$

as the L_p-norm of A_k.

To motivate further discussions, let us consider the following propositions.

PROPOSITION 2.1. *Consider the random equation* (2.2) *with* $x_0 = 0$. *Suppose that* $\{A_k, k \geq 0\}$ *is an independent sequence and* $\det[E(I - A_k)(I - A_k)^\tau] \neq 0$. *Then for any* $\{\xi_k\} \in \mathcal{B}$, $\{x_k\}$ *is* L_2*-stable if and only if there exist two constants* $M > 0$ *and* $\lambda \in [0, 1)$ *such that*

$$\left\| \prod_{j=i+1}^{k} (I - A_j) \right\|_{L_2} \leq M\lambda^{k-i}, \qquad \forall k \geq i, \qquad \forall i \geq 0 \tag{2.4}$$

where $\mathcal{B}$ *is a set of random processes defined by*

$$\mathcal{B} = \{\xi = (\xi_k) : \xi \text{ is } L_2\text{-stable and independent of } \{A_k\}\} \tag{2.5}$$

and where by definition

$$\prod_{j=i+1}^{k} (I - A_j) = \begin{cases} (I - A_k) \cdots (I - A_{i+1}), & k > i; \\ I, & k \leq i. \end{cases} \tag{2.6}$$

The proof is in Appendix A. Obviously, the only nontrivial conclusion in this proposition is that (2.4) is a necessary condition for L_2-stability of $\{x_k\}$. Related results in the deterministic framework may be found in [19]. We remark that when the independence assumptions are removed, similar necessity results are also true. This is the content of Proposition 2.2.

PROPOSITION 2.2. *Consider the random equation* (2.2) *with* $x_0 = 0$. *Assume that* $(I - A_k)^{-1}$ *exists for any* $k \geq 0$. *Denote*

$$\mathcal{B}^0 = \{\xi : \sup_k \|\xi_k\|_{L_2} \leq 1\}; \tag{2.7}$$

then the following property also implies (2.4):

$$\sup_{\xi\in\mathcal{B}^0} \sup_k \|x_k\|_{L_2} < \infty. \tag{2.8}$$

The proof is also given in Appendix A. These two propositions indicate that (2.4) is in some sense the necessary (and also sufficient) condition for the stability of $\{x_k\}$ generated by (2.2). This prompts us to introduce the following definition.

DEFINITION 2.2. A sequence of $d \times d$ random matrices $A = \{A_k\}$ is called stably exciting of order p, $(p \geq 1)$ with parameter $\lambda \in [0, 1)$, if it belongs to the following set

$$\mathcal{S}_p(\lambda) = \left\{ A : \left\| \prod_{j=i+1}^{k} (I - A_j) \right\|_{L_p} \leq M\lambda^{k-i}, \forall k \geq i, \forall i \geq 0, \text{ for some } M > 0 \right\}. \tag{2.9}$$

The investigation of products of random matrices has a long history (see e.g., [20]–[26] and the references therein), and almost all of the existing results rely on some stationary or mixing assumptions on the random coefficients. In particular, in [21] and [22] a time-invariant quadratic Lyapunov function was used to analyze the stability of a random linear differential equation under stationary and ergodic assumptions on the coefficients, while in [24] and [26] it was shown that under some mixing conditions, the stability of a random linear differential equation may be guaranteed by that of a corresponding "averaged" deterministic equation.

However, in general, stationary or mixing conditions cannot be directly imposed on the random coefficients in the study of tracking algorithms. Our treatment here is based on the observation that for a quite large class of matrix sequence $\{A_k\}$ of interest in applications, the study of its stably exciting property may be reduced to that of a certain class of scalar sequences. For convenience of discussion, we introduce the following subclass of $\mathcal{S}_1(\lambda)$ for scalar sequence $a = (a_k, k \geq 0)$:

$$\mathcal{S}^0(\lambda) = \left\{ a : a_k \in [0,1], E \prod_{j=i+1}^{k} (1 - a_j) \leq M\lambda^{k-i}, \forall\, k \geq i, \forall\, i \geq 0, \text{ for some } M > 0 \right\} \tag{2.10}$$

where $\lambda \in [0,1)$ is a parameter reflecting the stability margin. Note that for λ given above, $\log \lambda$ is related to the familiar concept of Lyapunov exponent (cf. [25]), and its absolute value is proportional to the exciting extent of $\{a_k\}$.

Clearly, for any constant $c \in (0,1]$, $\{c\} \in \mathcal{S}^0(1-c)$, and if $0 \leq \alpha_k \leq \beta_k \leq 1$ and $\{\alpha_k\} \in \mathcal{S}^0(\lambda)$, then $\{\beta_k\} \in \mathcal{S}^0(\lambda)$.

LEMMA 2.1. *Let $\alpha = \{\alpha_k, \mathcal{F}_k\}$ and $a = \{a_k, \mathcal{F}_k\}$ be adapted processes, such that*

$$a_k \in [0,1], \qquad E[a_{k+1}|\mathcal{F}_k] \geq \alpha_k, \qquad k \geq 0.$$

Then $\alpha \in \mathcal{S}^0(\lambda)$ implies that $a \in \mathcal{S}^0(\sqrt{\lambda})$.

Proof. We first assume that $0 \leq \alpha_k < 1$. For any $n > m, k \in [m,n]$, set

$$A_k = \left\{ \prod_{i=m}^{k} (1 - \alpha_i) \right\}^{-1}, \qquad A_{m-1} = 1$$
$$x_{k+1} = (1 - a_{k+1})x_k, \qquad x_m = 1.$$

Then

$$x_{n+1} = \prod_{i=m}^{n} (1 - a_{i+1}).$$

Note that

$$\begin{aligned} EA_k x_{k+1} &= EA_k[1 - E(a_{k+1}|\mathcal{F}_k)]x_k \\ &\leq EA_k(1 - \alpha_k)x_k = EA_{k-1}x_k. \end{aligned}$$

Hence

$$EA_n x_{n+1} \leq EA_{n-1}x_n \leq \cdots \leq EA_{m-1}x_m = 1.$$

Consequently,

$$E\prod_{i=m}^{n}(1-a_{i+1}) = Ex_{n+1} \le E\sqrt{x_{n+1}}$$
$$= E\sqrt{x_{n+1}A_n}\sqrt{A_n^{-1}} \le \sqrt{E(x_{n+1}A_n)EA_n^{-1}} \le \sqrt{EA_n^{-1}}$$
$$\le \left\{E\prod_{i=m}^{n}(1-\alpha_i)\right\}^{1/2} \le \sqrt{M}(\sqrt{\lambda})^{n-m+1}.$$

Hence $a \in \mathcal{S}^0(\sqrt{\lambda})$.

Next, we consider the general case $\alpha_k \in [0,1]$. By the monotonic convergence theorem, it is known that

$$\lim_{\varepsilon\to 1^-} E\prod_{k=m}^{n}(1-\varepsilon\alpha_k) \le M\lambda^{n-m+1}.$$

Hence there exists $0 < \varepsilon^* < 1$ such that for any $\varepsilon \in (\varepsilon^*, 1)$,

$$E\prod_{k=m}^{n}(1-\varepsilon\alpha_k) \le 2M\lambda^{n-m+1}.$$

Hence by $\varepsilon\alpha_k \in (0,1)$ and the fact proved above we have

$$E\prod_{k=m}^{n}(1-\varepsilon a_{k+1}) \le \sqrt{2M}(\sqrt{\lambda})^{n-m+1}.$$

Thus, by noticing that $\varepsilon a_{k+1} \le a_{k+1}$, we have $a \in \mathcal{S}^0(\sqrt{\lambda})$. This completes the proof. □

LEMMA 2.2. *Let* $\{\alpha_k, \mathcal{F}_k\}$ *be an adapted process,* $\alpha_k \in [0,1]$. *If for some integer* $h > 0$, $\{E[\alpha_{k+h}|\mathcal{F}_k]\} \in \mathcal{S}^0(\lambda)$, *then* $\{\alpha_k\} \in \mathcal{S}^0(\lambda^{2^{-h}})$.

Proof. Set $a_k = E[\alpha_{k+h-1}|\mathcal{F}_k]$. Then since

$$E[a_{k+1}|\mathcal{F}_k] = E\{E[\alpha_{k+h}|\mathcal{F}_{k+1}]|\mathcal{F}_k\} = E\{\alpha_{k+h}|\mathcal{F}_k\},$$

we know by Lemma 2.1 that $a_k \in \mathcal{S}^0(\sqrt{\lambda})$ or

$$\{E[\alpha_{k+h-1}|\mathcal{F}_k]\} \in \mathcal{S}^0(\sqrt{\lambda}).$$

Continuing this procedure h times, we finally get $\{\alpha_k\} \in \mathcal{S}^0(\lambda^{2^{-h}})$. □

LEMMA 2.3. *Let* $\{\alpha_k\} \in \mathcal{S}^0(\lambda)$, *and* $\alpha_k \le \alpha^* < 1$, *where* α^* *is a constant. Then for any* $0 < \varepsilon < 1$, $\{\varepsilon\alpha_k\} \in \mathcal{S}^0(\lambda^{(1-\alpha^*)\varepsilon})$.

Proof. We will need the following inequality ([14, p. 145])

$$1 - x \le (1-tx)^{\frac{(1-\alpha)}{t}}, \qquad t > 1, \qquad 0 \le tx \le \alpha < 1, \tag{2.11}$$

which can be proven by using standard differentiation methods.

Let M and $\lambda \in (0,1)$ be such that

$$E\prod_{k=m+1}^{n}(1-\alpha_k) \le M\lambda^{n-m}.$$

Then using the inequality (2.11) we have by taking $x = \varepsilon\alpha_k, t = 1/\varepsilon$,

$$E\prod_{k=m+1}^{n}(1-\varepsilon\alpha_k) \le E\left[\prod_{k=m+1}^{n}(1-\alpha_k)^{(1-\alpha^*)\varepsilon}\right]$$
$$\le \left\{E\prod_{k=m+1}^{n}(1-\alpha_k)\right\}^{(1-\alpha^*)\varepsilon} \le M^{(1-\alpha^*)\varepsilon}[\lambda^{(1-\alpha^*)\varepsilon}]^{n-m},$$

which implies the desired result. □

We now give some examples to illustrate the class $\mathcal{S}^0(\lambda)$.

Example 2.1. Nonzero strictly stationary processes do not necessarily belong to $\mathcal{S}^0(\lambda)$. Consider the process $\alpha_k \equiv \alpha$, with α being uniformly distributed on $[0, 1]$. Obviously, $\{\alpha_k\}$ is a stationary process. For any $n > 0$, we have

$$E\prod_{k=1}^{n}(1-\alpha_k) = E(1-\alpha)^n = \int_0^1 (1-x)^n\,dx = \frac{1}{n+1}.$$

This implies that $\{\alpha_k\} \notin \mathcal{S}^0(\lambda)$ for any $\lambda \in [0, 1)$, since the convergence rate of $E\prod_{k=1}^{n}(1-\alpha_k)$ is not exponentially fast.

Example 2.2. Let $\{\alpha_k, \mathcal{F}_k\}$ be any adapted process, $\alpha_k \in [0, 1]$. If there exists some constant $\alpha > 0$ and an integer $h > 0$, such that $E[\alpha_{k+h}|\mathcal{F}_k] \ge \alpha$, then $\{\alpha_k\} \in \mathcal{S}^0((1-\alpha)^{2^{-h}})$.

This fact can be easily proved by using Lemma 2.2. Example 2.2 contains many standard signals, for example, ϕ-mixing processes. To be precise, let ξ_k be a ϕ-mixing process, i.e., there exists a sequence $\phi(n) \xrightarrow[n\to\infty]{} 0$, such that

$$\sup_{A\in\mathcal{F}_{t+s}^{\infty}, B\in\mathcal{F}_0^t} |P(A|B) - P(A)| \le \phi(s), \quad \forall\, t, s,$$

where $\mathcal{F}_t^s \triangleq \sigma\{\xi(u), t \le u \le s\}$. Then for any $\mathcal{F}_t^{\infty}$-measurable f_t, with $|f_t| \le 1$, the following inequality holds (cf. [10], p. 82)

$$|E[f_{t+h}|\mathcal{F}_0^t] - Ef_{t+h}| \le 2\phi(h), \quad \forall\, t, h. \tag{2.12}$$

Hence if we take $f_t = f(\xi(t))$ and assume that $Ef_t \ge \alpha > 0$, for all t, where $f(\cdot) \in [0, 1]$ is a measurable function, then there exists an integer $h > 0$, such that $E[f_{t+h}|\mathcal{F}_0^t] \ge \alpha/2 > 0$, for all t. This verifies the conditions of Example 2.2 for ϕ-mixing processes.

2.2. A_k **nonnegative definite.** We are now in a position to study the more general class $\mathcal{S}_p(\lambda)$ defined by (2.9). We first study the stably exciting properties of nonnegative matrices $A_k, k \ge 1$, and see how the verification of $\{A_k\} \in \mathcal{S}_p(\lambda)$ can be transferred to that of a certain scalar sequence in $\mathcal{S}^0(\lambda)$.

THEOREM 2.1. *Let $\{A_i, \mathcal{F}_i\}$ be an adapted sequence of random matrices, $0 \le A_i \le I$. If there exists an integer $h > 0$, such that $\{\lambda_k\} \in \mathcal{S}^0(\lambda)$, where λ_k is defined by*

$$\lambda_k \triangleq \lambda_{\min}\left\{E\left[\frac{1}{1+h}\sum_{i=kh+1}^{(k+1)h} A_i|\mathcal{F}_{kh}\right]\right\},$$

then $\{A_k\} \in \mathcal{S}_2(\lambda^{\alpha})$, with $\alpha = 1/[8h(1+h)^2]$.

Proof. Recursively define

$$\Phi(n+1, m) = (I - A_n)\Phi(n, m), \quad \Phi(m, m) = I, \quad n \ge m \ge 0. \tag{2.13}$$

Then it can be shown that (see Appendix B) for any $m \geq 1$,

$$\begin{aligned}&\lambda_{\max}\{E[\Phi^\tau((m+1)h+1, mh+1)\Phi((m+1)h+1, mh+1)|\mathcal{F}_{mh}]\}\\ &\quad\leq 1 - \frac{\lambda_m}{(1+h)}.\end{aligned} \tag{2.14}$$

Now, for any $n > m + h$, let us define

$$k_0 = \min\{k : m \leq kh + 1 \leq n\}, \qquad k_1 = \max\{k : m \leq kh + 1 \leq n\}.$$

Then it is clear that

$$E\|\Phi(n, m)\|^2 \leq E\|\Phi(k_1 h + 1, k_0 h + 1)\|^2 \tag{2.15}$$

and

$$(k_1 + 1)h + 1 > n, \qquad (k_0 - 1)h + 1 < m. \tag{2.16}$$

Hence for $\{A_i\} \in \mathcal{S}_2(\lambda^\alpha)$, it suffices to find a constant c which is free of k_1 and k_0 such that, for all $k_1 \geq k_0$,

$$E\|\Phi(k_1 h + 1, k_0 h + 1)\|^2 \leq c\lambda^{2\alpha h(k_1 - k_0 + 1)}. \tag{2.17}$$

To prove this, we consider the following equation:

$$x_k = \Phi(kh + 1, (k - 1)h + 1)x_{k-1}, \qquad k \geq k_0 + 1 \tag{2.18}$$

where x_{k_0} is deterministic and $\|x_{k_0}\| = 1$. It is easily seen that $x_k \in \mathcal{F}_{kh}$, and $x_{k_1} = \Phi(k_1 h+1, k_0 h+1)x_{k_0}$. Therefore, for (2.17), we need only to prove that for any deterministic x_{k_0} with $\|x_{k_0}\| = 1$,

$$E\|x_{k_1}\|^2 \leq c\lambda^{2\alpha h(k_1 - k_0)} \tag{2.19}$$

where c is independent of k_0, k_1 and x_{k_0}.

Let us set for any $k \geq k_0 + 1$,

$$\alpha_k = \begin{cases} 1 - \dfrac{\|\Phi(kh + 1, (k - 1)h + 1)x_{k-1}\|}{\|x_{k-1}\|}, & \text{if } \|x_{k-1}\| \neq 0; \\ 1, & \text{otherwise.} \end{cases} \tag{2.20}$$

Since $0 \leq A_i \leq I$, $i \geq 0$, implies $\|\Phi(n, m)\| \leq 1$, for all $n \geq m$, $m \geq 0$, it is clear that $\alpha_k \in [0, 1]$, $\alpha_k \in \mathcal{F}_{kh}$, and by (2.18) and (2.20),

$$\|x_k\| \leq (1 - \alpha_k)\|x_{k-1}\|$$

and

$$\|x_{k_1}\| \leq \prod_{k=k_0+1}^{k_1} (1 - \alpha_k). \tag{2.21}$$

We now show that

$$E[\alpha_{k+1}|\mathcal{F}_{kh}] \geq \frac{\lambda_k}{2(1+h)}. \tag{2.22}$$

Set $\Omega_k = \{\omega : \|x_k\| = 0\}$. Then $\Omega_k \in \mathcal{F}_{kh}$, and by (2.20)

$$I_{\Omega_k} E[\alpha_{k+1}|\mathcal{F}_{kh}] = E[I_{\Omega_k}\alpha_{k+1}|\mathcal{F}_{kh}] = I_{\Omega_k}.$$

Hence by noting $\lambda_k < 1$ we see that (2.22) is true on the set Ω_k.

To prove that (2.21) is also true on the set Ω_k^c, we first note that by (2.14), we have

$$\begin{aligned}
&E[\|\Phi((k+1)h+1, kh+1)x_k\||\mathcal{F}_{kh}] \\
&\le \{E[\|\Phi((k+1)h+1, kh+1)x_k\|^2|\mathcal{F}_{kh}]\}^{1/2} \\
&\le \{x_k^\tau E[\Phi^\tau((k+1)h+1, kh+1)\Phi((k+1)h+1, kh+1)|\mathcal{F}_{kh}]x_k\}^{1/2} \\
&\le \left\{x_k^\tau\left(1 - \frac{\lambda_k}{1+h}\right)x_k\right\}^{1/2} \le \left(1 - \frac{\lambda_k}{2(1+h)}\right)\|x_k\|.
\end{aligned}$$

Consequently, by (2.20) we have

$$\begin{aligned}
I_{\Omega_k^c} E[\alpha_{k+1}|\mathcal{F}_{kh}] &\ge I_{\Omega_k^c}\left(1 - \left(1 - \frac{\lambda_k}{2(1+h)}\right)\right) \\
&= \frac{\lambda_k}{2(1+h)} I_{\Omega_k^c}.
\end{aligned} \tag{2.23}$$

Hence (2.22) is also true on Ω_k^c.

Since $\{\lambda_k\} \in \mathcal{S}^0(\lambda)$ and $\lambda_k \le h/(1+h)$, then by Lemma 2.3 we know that $\{\lambda_k/[2(1+h)]\} \in \mathcal{S}^0(\lambda^{4h\alpha})$. From this, (2.22) and Lemma 2.1 (together with its proof), we know that

$$\prod_{k=k_0+1}^{k_1}(1-\alpha_k) \le c\lambda^{2h\alpha(k_1-k_0)},$$

for some constant c independent of k_1, k_0, and x_{k_0}. Consequently, by (2.21) we see that (2.19) is true. Hence the proof of Theorem 2.1 is complete. □

COROLLARY 2.1. *Under the same conditions and notations as in Theorem* 2.1, *the following property holds*:

$$\{A_k\} \in \begin{cases} \mathcal{S}_p(\lambda^\alpha), & 1 \le p \le 2; \\ \mathcal{S}_p(\lambda^{2\alpha/p}), & p > 2. \end{cases} \tag{2.24}$$

Proof. For $1 \le p \le 2$, we use the monotonicity of the norm $\|\cdot\|_{L_p}$, while for $p > 2$ we apply the simple inequality $\|I - A_j\| \le 1$, and then derive

$$\left\|\prod_{j=i+1}^{k}(I - A_j)\right\|_{L_p} \le \begin{cases} \|\prod_{j=i+1}^{k}(I - A_j)\|_{L_2}, & 1 \le p \le 2; \\ \|\prod_{j=i+1}^{k}(I - A_j)\|_{L_2}^{2/p}, & p > 2. \end{cases}$$

Consequently (2.24) follows from this and Theorem 2.1. □

THEOREM 2.2. *Let* $\{A_i, \mathcal{F}_i\}$ *be an adapted sequence of random matrices,* $0 \le A_i \le I$. *If* $\{A_i\} \in \mathcal{S}_1(\lambda)$ *for some* $\lambda \in [0, 1)$, *then there exists an integer* $h > 0$ *such that*

$$\inf_m \lambda_{\min}\left\{\sum_{i=mh+1}^{(m+1)h} EA_i\right\} > 0.$$

Proof. By the assumption we know that there exists a suitably large integer $h > 0$ such that

$$E\left\|\prod_{i=mh+1}^{(m+1)h}(I - A_i)\right\| \leq M\lambda^h < \frac{1}{2}, \qquad \forall m. \tag{2.25}$$

Let ρ_m be the smallest eigenvalue of the matrix $E[\sum_{i=mh+1}^{(m+1)h} A_i]$, and x_m be its corresponding unit eigenvector. Then we have

$$\rho_m = E\left[\sum_{i=mh+1}^{(m+1)h} x_m^\tau A_i x_m\right].$$

Hence for any integers $i_j \in [mh+1, (m+1)h], j = 1, \ldots, k, k \leq h$,

$$\begin{aligned}
&Ex_m^\tau A_{i_1}\cdots A_{i_k}x_m \leq E\|x_m^\tau A_{i_1}^{1/2}\|\|A_{i_1}^{1/2}A_{i_2}\cdots A_{i_k}^{1/2}\|\|A_{i_k}^{1/2}x_m\| \\
&\leq E\|x_m^\tau A_{i_1}^{1/2}\|\cdot\|A_{i_k}^{1/2}x_m\| \leq \{E\|x_m^\tau A_{i_1}^{1/2}\|^2\cdot E\|A_{i_k}^{1/2}x_m\|^2\}^{1/2} \\
&= \{E(x_m^\tau A_{i_1}x_m)E(x_m^\tau A_{i_k}x_m)\}^{1/2} \leq \max_{mh+1\leq i\leq (m+1)h} E(x_m^\tau A_i x_m) \leq \rho_m.
\end{aligned}$$

Consequently, by (2.25) we have

$$\begin{aligned}
\frac{1}{2} &> E\left\|\prod_{i=mh+1}^{(m+1)h}(I - A_i)\right\| \geq Ex_m^\tau \prod_{i=mh+1}^{(m+1)h}(I - A_i)x_m \\
&= 1 - \sum_{k=1}^{h}\sum_{mh+1\leq i_1<\ldots<i_k\leq(m+1)h} E(x_m^\tau A_{i_1}\cdots A_{i_k}x_m) \\
&\geq 1 - \sum_{k=1}^{h}\sum_{mh+1\leq i_1<\ldots<i_k\leq(m+1)h} \rho_m = 1 - \sum_{k=1}^{h}\binom{h}{k}\rho_m,
\end{aligned}$$

which implies that

$$\rho_m \geq \frac{1}{2\sum_{k=1}^{h}\binom{h}{k}}.$$

Hence Theorem 2.2 is true. □

We remark that the converse assertion of Theorem 2.2 is not true in general. This fact can be seen from Example 2.1. However, it will be true if we impose additional assumptions on $\{A_k\}$, for example, the ϕ-mixing properties. The following theorem provides necessary and sufficient conditions for such a matrix process to be in $\mathcal{S}_1(\lambda)$.

THEOREM 2.3. *If $\{A_k, k \geq 0\}$ is a ϕ-mixing matrix sequence with dimension $d \times d$, and $0 \leq A_k \leq I$, then the following three properties are equivalent*:

(i) $\{A_k\} \in \mathcal{S}_1(\lambda)$ *for some* $\lambda \in [0, 1)$;

(ii) *There is an integer* $h_0 > 0$ *such that*

$$\delta \triangleq \inf_m \lambda_{\min}\left\{\sum_{i=mh_0+1}^{(m+1)h_0} EA_i\right\} > 0;$$

(iii) *There exist some* $h > 0, \lambda \in (0,1)$, *such that* $\{\lambda_k\} \in \mathcal{S}^0(\lambda)$ *where* λ_k *is defined as in Theorem* 2.1 *with* $\mathcal{F}_k \triangleq \sigma\{A_i, i \le k\}$.

Proof. By Theorems 2.1 and 2.2 we need only to prove that (ii) implies (iii).

Let the mixing rate of $\{A_k, k \ge 0\}$ be $\phi(k)$. Then applying the inequality (2.12), we are easily convinced of the following property:

$$\|E[A_{t+k}|\mathcal{F}_t] - EA_{t+k}\| \le 2d\phi(k), \quad \forall t, k. \tag{2.26}$$

Since $\phi(n) \xrightarrow[n\to\infty]{} 0$, we can find a constant (integer) M such that

$$\phi(k) \le \frac{\delta}{4(2h_0+1)d}, \quad \forall k \ge M, \tag{2.27}$$

where δ is defined in (ii).

Set $h = M + 2h_0 + 1$. Then by (ii) and the assumption $A_i \ge 0$, it is easy to convince oneself that

$$\lambda_{\min}\left\{\sum_{k=mh+1+M}^{(m+1)h} EA_k\right\} \ge \delta, \quad \forall m \ge 0. \tag{2.28}$$

Finally, combining (2.26)–(2.28) we conclude that for any $m \ge 0$,

$$\begin{aligned}
(1+h)\lambda_m &= \lambda_{\min}\left\{E\left[\sum_{k=mh+1}^{(m+1)h} A_k|\mathcal{F}_{mh}\right]\right\} \\
&\ge \lambda_{\min}\left\{E\left[\sum_{k=mh+1+M}^{(m+1)h} A_k|\mathcal{F}_{mh}\right]\right\} \\
&\ge \lambda_{\min}\left\{E\left[\sum_{k=mh+1+M}^{(m+1)h} A_k\right]\right\} - \left\|E\left[\sum_{k=mh+1+M}^{(m+1)h} A_k|\mathcal{F}_{mh}\right] - E\left[\sum_{k=mh+1+M}^{(m+1)h} A_k\right]\right\| \\
&\ge \delta - (h-M)\frac{2d\delta}{4(2h_0+1)d} = \delta - \frac{\delta}{2} = \frac{\delta}{2} > 0.
\end{aligned}$$

Hence for the h defined above, we have proved that $\{\lambda_k\} \in \mathcal{S}^0\{1 - \delta/[2(1+h)]\}$, i.e., (iii) holds. This completes the proof. □

2.3. A_k **nonsymmetric.** We now turn to the case where A_k is possibly nonsymmetrical and see how to transfer the study of $\{A_k\} \in \mathcal{S}_p(\lambda)$ to that of a scalar random sequence in $\mathcal{S}^0(\lambda)$.

Before pursuing this further, it is worth mentioning that in the continuous-time case, if $\{A(t)\}$ is a stationary ergodic matrix process and satisfies

$$E\lambda_{\max}\{A(0)^\tau + PA(0)P^{-1}\} < 0$$

for some positive definite matrix P, then the results of [21] and [22] state that the random differential equation $\dot{x}(t) = A(t)x(t)$ is almost surely asymptotically stable. This result may be generalized to the discrete-time case. However, this kind of results have the following limitations: (i) ergodicity is required; (ii) exponential stability can not be guaranteed, and (iii) applications to stochastic tracking algorithms are difficult.

Here, we will present a result that does not have the above-mentioned limitations. For this, we introduce the following recursive random Lyapunov equation:

$$P_{k+1} = (I - A_k)P_k(I - A_k)^\tau + Q_k, \qquad P_0 > 0, \quad k \geq 0, \tag{2.29}$$

where $\{Q_k\}$ is a sequence of nonnegative random matrices.

THEOREM 2.4. *Let $\{A_k\}$ be a sequence of $d \times d$ random matrices, and $\{Q_k\}$ be a sequence of positive definite random matrices. Then for $\{P_k\}$ recursively defined by* (2.29) *we have, for all $n > m$,*

$$\left\| \prod_{k=m}^{n-1} (I - A_k) \right\|^2 \leq \prod_{k=m}^{n-1} \left(1 - \frac{1}{1 + \|Q_k^{-1} P_{k+1}\|} \right) \|P_n\| \cdot \|P_m^{-1}\|. \tag{2.30}$$

Hence if $\{P_k\}$ satisfies the following two conditions,

$$\text{(i)} \quad \left\{ \frac{1}{1 + \|Q_k^{-1} P_{k+1}\|} \right\} \in \mathcal{S}^0(\lambda), \quad \textit{for some } \lambda \in [0, 1);$$

$$\text{(ii)} \quad \sup_{n \geq m \geq 0} \|(\|P_n\| \cdot \|P_m^{-1}\|)\|_{L_p} < \infty, \quad \textit{for some } p \geq 1,$$

then $\{A_k\} \in \mathcal{S}_p(\lambda^{1/2p})$.

Proof. Let us consider the following equation for $n > m$,

$$x_{k+1} = (I - A_k)x_k, \qquad k \in [m, n-1]$$

where x_m is taken to be deterministic and $\|x_m\| = 1$. Then

$$x_n = \prod_{i=m}^{n-1} (I - A_i)x_m. \tag{2.31}$$

Next we consider the following Lyapunov function $V_k = x_k^\tau P_k^{-1} x_k$. Then by denoting $B_k = I - A_k$, we have

$$V_{k+1} = x_{k+1}^\tau P_{k+1}^{-1} x_{k+1} = x_k^\tau B_k^\tau P_{k+1}^{-1} B_k x_k. \tag{2.32}$$

But, by (2.29) and the matrix inversion formula (see e.g., [27, p. 824]) we have

$$\begin{aligned} B_k^\tau P_{k+1}^{-1} B_k &= B_k^\tau [B_k P_k B_k^\tau + Q_k]^{-1} B_k \\ &= P_k^{-1} - [P_k + P_k B_k^\tau Q_k^{-1} B_k P_k]^{-1} \\ &= P_k^{-1/2} \{ I - [I + P_k^{1/2} B_k^\tau Q_k^{-1} B_k P_k^{1/2}]^{-1} \} P_k^{-1/2} \\ &\leq \{ 1 - [1 + \|Q_k^{-1} B_k P_k B_k^\tau\|]^{-1} \} P_k^{-1} \leq \left(1 - \frac{1}{1 + \|Q_k^{-1} P_{k+1}\|} \right) P_k^{-1}, \end{aligned}$$

which in conjunction with (2.32) yields

$$V_{k+1} \leq \left(1 - \frac{1}{1 + \|Q_k^{-1} P_{k+1}\|} \right) V_k$$

and so

$$V_n \le \prod_{k=m}^{n-1} \left(1 - \frac{1}{1 + \|Q_k^{-1} P_{k+1}\|}\right) V_m.$$

Hence by this, (2.31) and the dependence of V_k on x_m we have

$$\begin{aligned}
\left\| \prod_{k=m}^{n-1} (I - A_i) \right\|^2 &= \max_{\|x_m\|=1} \|x_n\|^2 = \max_{\|x_m\|=1} \|x_n^\tau P_n^{-1/2} P_n^{1/2}\|^2 \\
&\le \max_{\|x_m\|=1} \|x_n^\tau P_n^{-1/2}\|^2 \|P_n^{1/2}\|^2 = \max_{\|x_m\|=1} (V_n \|P_n\|) \\
&\le \left\{ \prod_{k=m}^{n-1} \left(1 - \frac{1}{1 + \|Q_k^{-1} P_{k+1}\|}\right) \right\} \left\{ \|P_n\| \max_{\|x_m\|=1} V_m \right\} \\
&\le \left\{ \prod_{k=m}^{n-1} \left(1 - \frac{1}{1 + \|Q_k^{-1} P_{k+1}\|}\right) \right\} \{\|P_n\| \cdot \|P_m^{-1}\|\}.
\end{aligned}$$

Hence (2.30) holds. The second assertion $\{A_k\} \in \mathcal{S}_p(\lambda^{1/2p})$ follows directly from (2.30) and the Hölder inequality.

This theorem does not require that $A_i's$ are nonnegative definite matrices and means that the verification of $\{A_k\} \in \mathcal{S}_p(\lambda^{1/2p})$ can be reduced to two relatively simple tasks: (i) to verify that a certain scalar sequence is in $\mathcal{S}^0(\lambda)$, and (ii) to prove that a certain process is "L_p-stable." We remark that suitably choosing the sequence $\{Q_k\}$ is crucial in simplifying the tasks (i) and (ii). In §4, we will see that for the analysis of KF or RLS algorithms, the sequence $\{P_k\}$ may simply be taken as that defined by (1.5) or (1.10). □

3. Stability/excitation condition. For the basic time-varying model (1.1), we will need the following excitation condition for estimating $\{\theta_k\}$.

CONDITION 3.1 (Excitation condition). *The regressor $\{\varphi_k, \mathcal{F}_k\}$ is an adapted sequence of random vectors (i.e., φ_k is $\mathcal{F}_k$-measurable, for all k, where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ-algebras), and there exists an integer $h > 0$ such that $\{\lambda_k\} \in \mathcal{S}^0(\lambda)$ for some $\lambda \in (0, 1)$, where λ_k is defined by*

$$\lambda_k \triangleq \lambda_{\min} \left\{ E \left[\frac{1}{1+h} \sum_{i=kh+1}^{(k+1)h} \frac{\varphi_i \varphi_i^\tau}{1 + \|\varphi_i\|^2} \,\middle|\, \mathcal{F}_{kh} \right] \right\}. \tag{3.1}$$

In the next section, we will show that this condition guarantees the L_p-stability of all three standard algorithms described in §1. The main purpose of this section is to illustrate this condition by several propositions and examples of interest in application.

PROPOSITION 3.1. *Let $\{\varphi_k\}$ be a ϕ-mixing process; then the necessary and sufficient condition for Condition* 3.1 *to be satisfied is that there exists an integer $h > 0$ such that*

$$\inf_{k \ge 0} \lambda_{\min} \left\{ \sum_{i=kh+1}^{(k+1)h} E \left[\frac{\varphi_i \varphi_i^\tau}{1 + \|\varphi_i\|^2} \right] \right\} > 0. \tag{3.2}$$

This fact directly follows from the equivalence of the assertions (ii) and (iii) in Theorem 2.3, since $\{\varphi_i \varphi_i^\tau / (1 + \|\varphi_i\|^2)\}$ is also a ϕ-mixing process. The ϕ-mixing process is commonly used in the literature (e.g., [8], [9], [17], [18]). It includes a large class of important processes,

for instance, deterministic processes, M-dependent processes and processes generated from bounded white noise filtered through a stable finite-dimensional linear filter. However, as is well known, ϕ-mixing is not perfect as a model in many applications, so next we show that Condition 3.1 is still satisfied by another important class of regressors that does not verify the ϕ-mixing condition.

In the sequel, for convenience of discussion we set $\mathcal{G}_k = \mathcal{F}_{kh}$ where h is defined in Condition 3.1. Note that λ_k is $\mathcal{G}_k$-measurable for any $k \geq 1$.

PROPOSITION 3.2. *If for some $h > 0$, $\{\lambda_k\}$ defined by* (3.1) *has the following time-varying lower bound*:

$$\lambda_k \geq \frac{1}{a_k}, \qquad \forall\, k \geq 1,$$

where $\{a_k, \mathcal{G}_k\}$ is an adapted sequence, $a_k \geq 1, Ea_0 < \infty$, and

$$E[a_k|\mathcal{G}_{k-1}] \leq \alpha a_{k-1} + \beta, \qquad 0 \leq \alpha < 1, 0 < \beta < \infty, \qquad \forall\, k \geq 1. \tag{3.3}$$

Then $\{\lambda_k\} \in \mathcal{S}^0(\lambda)$ for some $\lambda \in (0, 1)$, i.e., Condition 3.1 *holds.*

Proof. By Lemma 4 in [14], we know that there exists a constant $\lambda \in (0, 1)$ such that $\{1/a_k\} \in \mathcal{S}^0(\lambda)$. Hence Condition 3.1 follows immediately. □

Remark 3.1. Intuitively speaking, in order to guarantee $\{\lambda_k\} \in \mathcal{S}^0(\lambda)$, the lower bound $\{1/a_k\}$ should not "diminish" or equivalently, $\{a_k\}$ should not "grow unboundedly." Condition (3.3) effectively is a growth constraint on the random process $\{a_k\}$. If in (3.3) we take $\alpha = 0$ and $a_k = \beta$, then we get the excitation condition used in [14]. Moreover, if we assume that $\{a_k\}$ satisfies $a_k \in \mathcal{G}_k, a_k \geq 1$ and

$$a_k \leq \alpha a_{k-1} + \eta_k, \qquad \alpha \in [0, 1), \qquad E[|\eta_k|^{1+\delta}|\mathcal{G}_{k-1}] \leq M, \quad \forall\, k \geq 1 \tag{3.4}$$

for some constants $\delta > 0$, and $M < \infty$, then we get the excitation condition proposed in [28], which obviously satisfies (3.3). Therefore, the condition of Proposition 3.2 (and hence Condition 3.1) is weaker than those proposed in [14] and [28]. Consequently, all examples presented in [14] and [28] satisfy the condition of Proposition 3.2. In particular, we have Example 3.1.

Example 3.1. Let the regressor $\{\varphi_k\}$ be generated by the following state space model:

$$\begin{aligned} x_k &= Ax_{k-1} + B\xi_k, & E\|x_0\|^4 < \infty \\ \varphi_k &= Cx_k + \zeta_k, & k \geq 0, \end{aligned}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times q}$ and $C \in \mathbb{R}^{d\times n}$ are deterministic matrices, A is stable, (A, B, C) is output controllable and $\{\xi_k, \zeta_k\}$ is an independent process with zero mean, and

$$E\xi_k\xi_k^\tau \geq \varepsilon I > 0, \qquad E[\|\xi_k\|^4 + \|\zeta_k\|^4] \leq M, \qquad \forall\, k \geq 0,$$

where ε and M are constants. Then the condition of Proposition 3.2 is satisfied.

The proof of this example is essentially the same as that for Example 2 in [28], but here the moment condition imposed on the driving signal $\{\xi_k, \zeta_k\}$ is weaker. It is also worth noting that to verify the condition in [14] we have to assume that $\{\xi_k, \zeta_k\}$ is uniformly bounded in the sample path (see [14, p. 142]).

We now turn to the main task of this section, i.e., to study the case where $\{\varphi_k\}$ is generated by a time-varying $AR(p)$ model. This model not only is a natural extension of the standard time-invariant $AR(p)$ models extensively studied in a variety of areas, but also is closely related to the closed-loop systems resulting from adaptive control (cf. [30]). We remark

that in this case, the existing excitation conditions (e.g., in [14] and [28]) do not seem to be satisfied. The basic reason is that the "contraction" factor α in (3.4) is a random process rather than a constant.

Let the time-varying $AR(p)$ model be described by

$$\begin{aligned} y_k &= a_1(k)y_{k-1} + \cdots + a_p(k)y_{k-p} + v_k \\ &\triangleq \theta_k^\tau \varphi_k + v_k, \qquad k \geq 0 \end{aligned} \tag{3.5}$$

where θ_k and φ_k are p-dimensional vectors defined in a standard way, and where $\{v_k\}$ is an independent sequence that is independent of φ_0 and satisfies

$$Ev_k = 0, \qquad Ev_k^2 \geq \sigma_v^2 > 0, \qquad \sup_k E|v_k|^9 < \infty. \tag{3.6}$$

Obviously, the regressor satisfies the following state space equation:

$$\varphi_{k+1} = A_k \varphi_k + bv_k \tag{3.7}$$

where

$$A_k = \begin{bmatrix} a_1(k) & \cdots & \cdots & a_p(k) \\ 1 & \cdots & \cdots & 0 \\ \vdots & \ddots & & \\ 0 & \cdots & 1 & 0 \end{bmatrix}, \qquad b = [1, 0 \cdots 0]^\tau. \tag{3.8}$$

Example 3.2. Consider the $AR(p)$ model (3.5)–(3.6). Let $\{A_k\}$ defined by (3.8) be an independent sequence that is independent of $\{v_k\}$. If

$$\sup_k \|A_k\|_{L_q} < \infty, \qquad \left\| \prod_{i=kp}^{(k+1)p-1} A_i \right\|_{L_4} \leq \delta, \quad \forall k \geq 0, \tag{3.9}$$

where $q = \max\{4, 2(p-1)\}$ and $\delta \in (0, 1)$, then the condition of Proposition 3.2 is satisfied.

The proof is given in Appendix C.

When the coefficient sequence $\{A_k\}$ is (strongly) dependent, the analysis becomes more complicated. We now consider a standard situation.

PROPOSITION 3.3. Consider the $AR(p)$ model (3.5)–(3.6). Let $\{A_k, \mathcal{F}_k'\}$ be an adapted sequence that can be decomposed as

$$A_k = A + \bar{A}_k \tag{3.10}$$

where A is a stable matrix and $\{\bar{A}_k, \mathcal{F}_k'\}$ is dominated by a nonnegative linear process:

$$\|\bar{A}_k\| \leq \beta_k, \qquad \beta_k = \beta\beta_{k-1} + e_k, \qquad 0 \leq \beta < 1, \tag{3.11}$$

where $e_k \geq 0$, $e_k \in \mathcal{F}_k'$ and e_{k+1} is independent of $\mathcal{F}_k'$. Assume that $\mathcal{F}_\infty' \triangleq \sigma\{\cup_i \mathcal{F}_i'\}$ is independent of $\{v_k\}$ and that for some constants $\varepsilon > 0$ and $b > 0$

$$\log\{E[\exp(be_k)]\} \leq \varepsilon, \qquad \forall k \geq 0. \tag{3.12}$$

Then Condition 3.1 is satisfied provided that ε and b are suitably small and large respectively.

The proof of this proposition is given in Appendix C.

Example 3.3. Let the parameter θ_k in (3.5) be the superposition of a "nominal" parameter θ and a "fluctuation" $\bar{\theta}_k$, i.e., $\theta_k = \theta + \bar{\theta}_k$. Moreover, let the time-invariant $AR(p)$ model obtained by replacing θ_k by θ in (3.5) be stable. If either $\|\bar{\theta}_k\|$ is small or $\bar{\theta}_k$ is generated by a stable *ARMA* model:

$$\bar{\theta}_k + F_1\bar{\theta}_{k-1} + \cdots + F_q\bar{\theta}_{k-q} = w_k + G_1 w_{k-1} + \cdots + G_r w_{k-r}$$

where $\{w_k\}$ is a Gaussian white noise sequence which is independent of $\{v_k\}$ with small variance. Then conditions (3.10)–(3.12) of Proposition 3.3 hold.

The proof of this example is straightforward and the details are omitted.

Remark 3.2. Conditions in Example 3.2, Proposition 3.3, and Example 3.3 are stronger than necessary as can be easily seen from the proof; they are used for simplicity of discussion. Certainly, various generalizations are possible, for example, a more general state space model (3.7) may be considered without requiring that A_k and b have the canonical form (3.8); in Example 3.2, the independence assumption of $\{A_k\}$ can be replaced by some weakly dependent conditions; and in Example 3.3, the Gaussian assumption on $\{w_k\}$ can be weakened by assuming that the distribution of $\{w_k\}$ has exponentially decaying tail (a condition similar to (3.12)).

The following result plays an essential role in the proof of Proposition 3.3 and will also be used in the next section.

LEMMA 3.1. *Let* $\{x_k, \mathcal{F}_k\}$ *be an adapted process,* $x_k \geq 1$, *and*

$$x_{k+1} \leq \alpha_{k+1}x_k + \xi_{k+1}, \qquad k \geq 0, \qquad Ex_0^2 < \infty \tag{3.13}$$

where $\{\alpha_k, \mathcal{F}_k\}$ *and* $\{\xi_k, \mathcal{F}_k\}$ *are adapted nonnegative processes with properties*:

$$\alpha_k \geq \varepsilon_0 > 0, \quad \forall k, \quad \left\| \prod_{k=m}^{n} E[\alpha_{k+1}^4 | \mathcal{F}_k] \right\|_{L_1} \leq M\gamma^{n-m+1}, \quad \forall n \geq m, \quad \forall m \tag{3.14}$$

and

$$E[\xi_{k+1}^2 | \mathcal{F}_k] \leq N < \infty, \qquad \forall k \tag{3.15}$$

where $\varepsilon_0, M, N,$ *and* $\gamma \in (0,1)$ *are constants. Then*

(i) $\left\| \prod_{k=m}^{n} \alpha_k \right\|_{L_2} \leq M^{1/4}\gamma^{(1/4)(n-m+1)}, \qquad \forall n \geq m, \quad \forall m;$

(ii) $\sup_k E\|x_k\| < \infty;$

(iii) $\{1/x_k\} \in S^0(\lambda)$ *for some* $\lambda \in (0,1)$.

Proof. Denote $\beta_k = E[\alpha_{k+1}^4 | \mathcal{F}_k]$, and set $z_{k+1} = \left(\prod_{i=m}^{k} \beta_i\right)^{-1} \prod_{i=m}^{k} \alpha_{i+1}^4$. Then we have $z_{k+1} = z_k\beta_k^{-1}\alpha_{k+1}^4$, and so

$$Ez_{k+1} = E\{E[z_{k+1}|\mathcal{F}_k]\} = Ez_k = \cdots = Ez_{m+1} = 1, \qquad \forall k \geq m.$$

Consequently, for all $n \geq m$,

$$E\prod_{i=m}^{n} \alpha_{i+1}^2 = E\sqrt{z_{n+1}}\sqrt{\prod_{m}^{n} \beta_i}$$

$$\leq \sqrt{Ez_{n+1}}\sqrt{E\prod_{m}^{n} \beta_i} \leq \sqrt{M}\sqrt{\gamma^{n-m+1}}$$

so (i) holds, while (ii) follows immediately from (i), (3.15), and (3.13). We now proceed to prove the last assertion (iii).

We first consider the case where N defined by (3.15) is less than one. In this case, by (3.15) we have $E[\xi_{k+1}|\mathcal{F}_k] \le 1$.

For any $n > m$, set for $k \in [m, n]$

$$y_k = \left(1 - \frac{1}{x_k}\right) y_{k-1}, \qquad y_{m-1} = 1. \tag{3.16}$$

Then $y_k \in \mathcal{F}_k$ and by (3.13) we have

$$x_k y_k = (x_k - 1)y_{k-1} \le (\alpha_k x_{k-1} + \xi_k - 1)y_{k-1}$$

so with $\gamma_k \triangleq E[\alpha_{k+1}|\mathcal{F}_k]$ by noticing that $E[\xi_k|\mathcal{F}_{k-1}] \le 1$ we get

$$E[x_k y_k|\mathcal{F}_{k-1}] \le \gamma_{k-1}(x_{k-1}y_{k-1}), \qquad k \ge m. \tag{3.17}$$

Denote $z_k = \left(\prod_{i=m-1}^{k-1} \gamma_i\right)^{-1} x_k y_k, k \ge m-1$. Then by (3.17) we have for $k \ge m$,

$$E[z_k|\mathcal{F}_{k-1}] \le \left(\prod_{i=m-1}^{k-2} \gamma_i\right)^{-1} x_{k-1}y_{k-1} = z_{k-1}.$$

Consequently,

$$Ez_k \le Ez_{k-1} \le \cdots \le Ez_{m-1} = Ex_{m-1}. \tag{3.18}$$

Hence by (ii) we have for some constant $M_0 < \infty$, $\sup_{m\ge 0} \sup_{k\ge m} Ez_k \le M_0$. Thus by the Schwarz inequality and (3.14) we have

$$\begin{aligned} E\prod_{k=m}^{n}\left(1 - \frac{1}{x_k}\right) = Ey_n \le E\sqrt{x_n y_n} &= E\sqrt{z_n \prod_{i=m-1}^{n-1} \beta_i} \\ &\le \sqrt{Ez_n} \cdot \sqrt{E\prod_{i=m-1}^{n-1} \beta_i} \le \sqrt{M_0} M^{1/8} \gamma^{1/8(n-m+1)} \end{aligned}$$

where for the last inequality (3.14) has been used. Hence (iii) holds.

Next, we consider the general case where N in (3.15) is an arbitrary constant. By (3.15) we may take a constant c large enough such that

$$E[\xi_{k+1} I(\xi_{k+1} \ge c)|\mathcal{F}_k] \le 1, \quad \text{and} \quad \delta \triangleq (1 + \varepsilon_0)\frac{c}{1+c} > 1.$$

Then we have by (3.13),

$$x_{k+1} \le \alpha_{k+1} x_k + c + \xi_{k+1} I(\xi_{k+1} > c), \qquad k \ge 0. \tag{3.19}$$

Without loss of generality, we may assume that the equality in (3.19) holds for all k. Hence by setting $\bar{x}_k = x_k/(1+c)$ we get

$$\bar{x}_{k+1} = \alpha_{k+1}\bar{x}_k + \eta_{k+1} \tag{3.20}$$

where $\eta_{k+1} = [c + \xi_{k+1} I(\xi_{k+1} > c)]/(1+c)$. It is clear that $E[\eta_{k+1}|\mathcal{F}_k] \le 1$. Then by the fact we have just proved we know that $\{1/\bar{x}_k\} \in S^0(\gamma^{1/8})$, where γ is given in (3.14).

Note that by (3.14) and (3.20)

$$\bar{x}_{k+1} \ge \alpha_{k+1}\left(\frac{c}{1+c}\right) + \frac{c}{1+c} \ge \frac{c}{1+c}(1+\varepsilon_0) > 1, \qquad k \ge 1.$$

Hence applying Lemma 2.3 with $\varepsilon = 1/(1+c)$ we know that $\{1/x_k\} \in S^0(\lambda)$, for some $\lambda \in (0,1)$. This completes the proof of Lemma 3.1. □

We remark that the condition $x_k \ge 1$ in Lemma 3.1 is by no means a restrictive condition in applications since if $x_k \ge 0$ satisfies (3.13), then the shifted process $x'_k \triangleq x_k + 1$ satisfies both $x'_k \ge 1$ and $x'_{k+1} \le \alpha_{k+1} x'_k + \xi'_{k+1}$ where $\xi'_{k+1} = \xi_{k+1} + 1$.

COROLLARY 3.1. *Let $\{x_k\}$ satisfy conditions in Lemma* 3.1. *If $\{y_n, \mathcal{F}_n\}$ is a nonnegative adapted process and satisfies*:

$$y_{k+1} \le \beta y_k + \eta_{k+1}, \qquad 0 \le \beta < 1, \qquad \forall\, k \tag{3.21}$$

where $E[\eta_k^{2q}|\mathcal{F}_{k-1}] \le M_1 < \infty$, M_1 is a positive constant and $q > \log \varepsilon_0 / \log \beta$ is a positive integer and ε_0 is defined in (3.14), *then $\{1/(x_k + y_k)\} \in S^0(\lambda)$ for some $\lambda \in (0,1)$.*

Proof. Take ε so small such that $(1+\varepsilon)\beta^q \le \varepsilon_0$, and define $T_k = (1/q) y_k^q + (1/s)$, where $s = (1 - 1/q)^{-1}$. Note that for any $\varepsilon > 0$ and $q > 0$ there is a constant $M > 0$ depending on ε and q such that

$$(x+y)^q \le (1+\varepsilon)x^q + My^q, \qquad \forall\, x \ge 0, \quad \forall\, y \ge 0. \tag{3.22}$$

Then we have

$$\begin{aligned} T_k &\le \frac{1}{q}[\beta y_{k-1} + \eta_k]^q + \frac{1}{s} \\ &\le \frac{1}{q}[(1+\varepsilon)(\beta y_{k-1})^q + M\eta_k^q] + \frac{1}{s} \\ &\le \varepsilon_0 \left[\frac{1}{q} y_{k-1}^q + \frac{1}{s}\right] + \frac{M}{q}\eta_k^q + \frac{1}{s} \le \varepsilon_0 T_{k-1} + \frac{M}{q}\eta_k^q + \frac{1}{s}. \end{aligned}$$

Hence

$$\begin{aligned} x_k + T_k &\le \alpha_k x_{k-1} + \xi_k + \varepsilon_0 T_{k-1} + \frac{M}{q}\eta_k^q + \frac{1}{s} \\ &\le \alpha_k (x_{k-1} + T_{k-1}) + \xi_k + \frac{M}{q}\eta_k^q + \frac{1}{s}. \end{aligned}$$

Applying Lemma 3.1 we know that $\{1/(x_k + T_k)\} \in S^0(\lambda)$, for some $\lambda \in (0,1)$. Finally note that $y_k \le T_k$; we conclude that $\{1/(x_k + y_k)\} \in S^0(\lambda)$. □

4. Tracking error bounds. In this section we establish tracking error bounds for the standard algorithms introduced in §1. We first present a lemma.

LEMMA 4.1. *Let $\{c_{nk}, n \ge k \ge 0\}, \{d_{nk}, n \ge k \ge 0\}$, and $\{\xi_k, k \ge 0\}$ be three nonnegative random processes satisfying*:

(i) $c_{nk} \in [0,1], Ec_{nk} \le M\lambda^{n-k}$, *for all $n \ge k \ge 0$, for some $M > 0$ and $\lambda \in [0,1)$;*

(ii) *There exist some constants $\varepsilon > 0$ and $\alpha > 0$ such that*

$$\sup_{n \ge k \ge 0} E[\exp(\varepsilon d_{nk}^{1/\alpha})] < \infty;$$

(iii) $\sigma_p \triangleq \sup_k \|\xi_k \log^\beta(e+\xi_k)\|_{L_p} < \infty$, *for some* $p \geq 1, \beta > 0$.

Then

$$\sum_{k=0}^{n} \|c_{nk} d_{nk} \xi_k\|_{L_p} \leq c\sigma_p f(\sigma_p^{-1}), \qquad \forall n \geq 0, \tag{4.1}$$

where c *is a constant independent of* σ_p, *and*

$$f(\sigma_p^{-1}) = \begin{cases} \log^{1+(\beta/2)}(e+\sigma_p^{-1}), & \text{if } \beta > 2\max(1,\alpha); \\ \log^{\beta}(e+\sigma_p^{-1}), & \text{if } \{c_{nk}\} \text{ is deterministic and } \beta = \alpha; \\ \log(e+\sigma_p^{-1}), & \text{if } \{d_{nk}\} \text{ is deterministic and } \beta > 1. \end{cases} \tag{4.2}$$

The proof is given in Appendix D.

We now proceed to analyze the Kalman filter algorithm. To apply Theorem 2.4 we need to prove some boundedness properties of $\{P_k\}$ first.

LEMMA 4.2. *For* $\{P_k\}$ *generated by* (1.5), *if Condition* 3.1 *holds, then there exists a constant* $\varepsilon^* > 0$ *such that for any* $\varepsilon \in [0, \varepsilon^*)$,

$$\sup_{k \geq 0} E \exp(\varepsilon \|P_k\|) < \infty.$$

Proof. Denote

$$T_m = \sum_{k=(m-1)h+1}^{mh} tr(P_{k+1}), \qquad T_0 = 0. \tag{4.3}$$

Then $T_m \in \mathcal{G}_m \triangleq \mathcal{F}_{mh}$, and similar to Lemma 3 in [28] we have

$$T_{m+1} \leq (1 - a_{m+1}) T_m + b \tag{4.4}$$

where

$$a_{m+1} = \frac{tr\left[(P_{mh+1} + hQ)^2 \sum_{k=mh+1}^{(m+1)h} \dfrac{\varphi_k \varphi_k^\tau}{1+\|\varphi_k\|^2}\right]}{h(R+1)[1+\lambda_{\max}(P_{mh+1}+hQ)]tr(P_{mh+1}+hQ)}, \qquad b = \frac{3}{2}h(h+1)trQ.$$

Similar to (39) and (40) in [28] we have $a_{m+1} \in [0, 1/(1+R)]$ and

$$E[a_{m+1} | \mathcal{G}_m] \geq \frac{(1+h)\|Q\|\lambda_m}{d(R+1)(1+h\|Q\|)} \tag{4.5}$$

where λ_m is defined by (3.1). By using Condition 3.1 and applying Lemmas 2.1 and 2.3, it is easy to see that $\{a_{k+1}\} \in S^0(\lambda)$ for some $\lambda \in [0, 1)$. Hence, the rest of the proof is completely the same as that for Lemma 4 in [28], because the key property (43) in [28] is still true. □

LEMMA 4.3. *Let* $\{P_k\}$ *be generated by* (1.5). *Then under Condition* 3.1, *for any* $\mu \in (0, 1]$ *there is a constant* $\lambda \in (0, 1)$ *such that* $\{\mu/(1 + \|Q^{-1}\| \cdot \|P_k\|)\} \in S^0(\lambda)$.

Proof. Denote $x_k = \mu^{-1}(h + \|Q^{-1}\| T_k)$, where T_k is defined by (4.3). Then it follows from (4.4) that

$$x_{k+1} \leq (1 - a_{k+1}) x_k + \mu^{-1}(h + b\|Q^{-1}\|). \tag{4.6}$$

It is easy to see from (4.5), Condition 3.1, and Lemma 2.3 that Lemma 3.1 is applicable to (4.6); hence, we have $\{1/x_k\} \in S^0(\gamma)$, for some $\gamma \in (0,1)$. Note that $x_k = \sum_{i=(k-1)h+1}^{kh} \mu^{-1} \cdot [1 + \|Q^{-1}\| tr(P_{i+1})]$; hence, it is easy to conclude that $\{\mu/[1 + \|Q^{-1}\| tr(P_k)]\} \in S^0(\lambda)$ for some $\lambda \in (0,1)$ (see the proof of Lemma 5 in [14]), which ensures the desired result. □

THEOREM 4.1. *Consider the time-varying model* (1.1) *and the Kalman filter algorithm* (1.3)–(1.5). *Suppose that Condition* 3.1 *is satisfied and that for some* $p \geq 1$ *and* $\beta > 2$,

$$\sigma_p \triangleq \sup_k \|\xi_k \log^{\beta}(e + \xi_k)\|_{L_p} < \infty \tag{4.7}$$

and

$$\|\tilde{\theta}_0\|_{L_{2p}} < \infty \tag{4.8}$$

where $\xi_k = |v_k| + \|\Delta_{k+1}\|, \tilde{\theta}_0 = \theta_0 - \hat{\theta}_0$, *and* v_k *and* Δ_{k+1} *are given by* (1.1) *and* (1.2), *respectively. Then the tracking error* $\{\theta_k - \hat{\theta}_k, k \geq 0\}$ *is* L_p*-stable and*

$$\limsup_{k\to\infty} \|\theta_k - \hat{\theta}_k\|_{L_p} \leq c[\sigma_p \log^{1+\beta/2}(e + \sigma_p^{-1})], \tag{4.9}$$

where c *is a finite constant depending on* $\{\varphi_k\}, R, Q$ *and* p *only; its precise value may be found from the proof.*

Proof. By (1.4) we may rewrite (1.5) as

$$P_{k+1} = (I - L_k\varphi_k^{\tau})P_k(I - L_k\varphi_k^{\tau})^{\tau} + Q_k$$

where $Q_k = RL_kL_k^{\tau} + Q$. It is easy to see that $Q_k \geq Q$ and $P_{k+1} \geq Q$. Hence by applying Theorem 2.4 we have for all $n > m$,

$$\left\| \prod_{k=m}^{n-1} (I - L_k\varphi_k^{\tau}) \right\| \leq \prod_{k=m}^{n-1} \left(1 - \frac{1}{1 + \|Q^{-1}\| \cdot \|P_{k+1}\|}\right)^{1/2} \cdot \|P_n\|^{1/2}\|Q^{-1}\|^{1/2}. \tag{4.10}$$

Note also that $\|L_k\| \leq \|P_k\|^{1/2}/(2\sqrt{R})$, so by (2.1) we get

$$\begin{aligned} \|\tilde{\theta}_{k+1}\|_{L_p} \leq{}& \left\| \prod_{i=0}^{k} (I - L_i\varphi_i^{\tau})\tilde{\theta}_0 \right\|_{L_p} + \|Q^{-1}\|^{1/2} \sum_{i=0}^{k} \\ & \cdot \left\| \prod_{j=i+1}^{k} \left(1 - \frac{1}{2(1 + \|Q^{-1}\| \cdot \|P_{j+1}\|)}\right) \|P_{k+1}\|^{1/2} \left(1 + \frac{\|P_i\|^{1/2}}{2\sqrt{R}}\right) \xi_i \right\|_{L_p}. \end{aligned} \tag{4.11}$$

Note that by the Schwarz inequality and Lemma 4.2,

$$\sup_{k\geq i} E \exp(\varepsilon \|P_{k+1}\|^{1/2}\|P_i\|^{1/2}) \leq \sup_{k\geq i} [E \exp(\varepsilon \|P_{k+1}\|)]^{1/2} [E \exp(\varepsilon \|P_i\|)]^{1/2} < \infty.$$

So by noting Lemma 4.3 and applying Lemma 4.1 to the second term on the right-hand side of (4.11), we get the desired result. □

Next, we consider the LMS algorithm.

THEOREM 4.2. *Consider the time-varying model* (1.1) *and the LMS algorithm* (1.3) *and* (1.8). *Suppose that Condition* 3.1 *holds and that for some* $p \geq 1$ *and* $\beta > 1$, (4.7) *and* (4.8) *hold. Then* $\{\theta_k - \hat{\theta}_k, k \geq 0\}$ *is* L_p*-stable, and*

$$\limsup_{k\to\infty} \|\theta_k - \hat{\theta}_k\|_{L_p} \leq c[\sigma_p \log(e + \sigma_p^{-1})], \tag{4.12}$$

where σ_p is defined by (4.7) *and c is a constant.*

Proof. Let $c_{ki} = \left\| \prod_{j=i+1}^{k} \left(I - \mu \frac{\varphi_j \varphi_j^\tau}{1+\|\varphi_j\|^2} \right) \right\|$. Then by Condition 3.1, Lemma 2.3, and Theorem 2.1 we know that $\{c_{ki}\}$ satisfies conditions in Lemma 4.1. Note that $\|L_k\| \le \mu$, so by (2.1) we have

$$\|\tilde{\theta}_{k+1}\|_{L_p} \le \|c_{k,-1}\tilde{\theta}_0\|_{L_p} + \sum_{i=0}^{k} \|c_{ki}\xi_i\|_{L_p}$$

and the desired result (4.12) follows by applying Lemma 4.1. □

Remark 4.1. Combining Propositions 2.1 and 2.2 with Theorem 2.3, we see that Condition 3.1 is also a necessary one for the stability of the LMS algorithm *in some sense.*

Finally, we study the recursive least squares algorithm.

LEMMA 4.4. *Let $\{P_k\}$ be generated by* (1.10) *with forgetting factor $\alpha \in (0,1)$. If Condition* 3.1 *holds, then for any $p \ge 1$*

$$\sup_{k\ge 0} E\|P_k\|^p < \infty,$$

provided that α satisfies $\lambda^{[16hd(2h-1)p]^{-1}} < \alpha < 1$, where λ and h are given by Condition 3.1, *and d is the dimension of $\{\varphi_k\}$.*

Proof. The proof ideas are similar to those for Lemmas 1 and 2 in [14] for the Kalman filter algorithm. For any $m \ge 0$ by (1.10) we have

$$P_k \le \frac{1}{\alpha} P_{k-1} \le \cdots \le \left(\frac{1}{\alpha}\right)^{h-1} P_{mh+1}, \qquad k \in [mh+1, (m+1)h].$$

Then by the matrix inverse formula from (1.10) again we have for $k \in [mh+1, (m+1)h]$,

$$\begin{aligned} P_{k+1} &= [\alpha P_k^{-1} + \varphi_k \varphi_k^\tau]^{-1} \le [\alpha \alpha^{h-1} P_{mh+1}^{-1} + \varphi_k \varphi_k^\tau]^{-1} \\ &= \left(\frac{1}{\alpha}\right)^h \left[P_{mh+1} - \frac{P_{mh+1}\varphi_k \varphi_k^\tau P_{mh+1}}{\alpha^h + \varphi_k^\tau P_{mh+1} \varphi_k} \right] \\ &\le \left(\frac{1}{\alpha}\right)^h \left[P_{mh+1} - \frac{P_{mh+1}\varphi_k \varphi_k^\tau P_{mh+1}}{[\alpha^h + \lambda_{\max}(P_{mh+1})][1+\|\varphi_k\|^2]} \right]. \end{aligned} \tag{4.13}$$

Denote

$$T_m = \sum_{k=(m-1)h+1}^{mh} tr(P_{k+1}), \qquad a_{m+1} = \frac{tr\left[P_{mh+1}^2 \sum_{k=mh+1}^{(m+1)h} \frac{\varphi_k \varphi_k^\tau}{1+\|\varphi_k\|^2} \right]}{[\alpha^h + \lambda_{\max}(P_{mh+1})] h tr(P_{mh+1})}. \tag{4.14}$$

Then summing up both sides of (4.13) we get

$$T_{m+1} \le \alpha^{-h}[1 - a_{m+1}] h tr(P_{mh+1}). \tag{4.15}$$

But by the inequality $P_{k+1} \le \alpha^{-1} P_k$ it follows that

$$\begin{aligned} h tr(P_{mh+1}) &= \sum_{k=(m-1)h+1}^{mh} tr(P_{mh+1}) \\ &\le \sum_{k=(m-1)h+1}^{mh} \alpha^{k-mh} tr(P_{k+1}) = \alpha^{1-h} T_m. \end{aligned}$$

Hence by (4.15)

$$T_{m+1} \le \alpha^{1-2h}[1 - a_{m+1}]T_m. \tag{4.16}$$

For any $p \ge 1$, denote

$$b_{m+1} = \alpha^{(1-2h)p}\left[1 - \frac{a_{m+1}}{2}\right] I(tr(P_{mh+1}) \ge 1).$$

Then by (4.15) and (4.16),

$$\begin{aligned} T_{m+1}^p &\le T_{m+1}^p[I(tr(P_{mh+1}) \ge 1) + I(tr(P_{mh+1} \le 1))] \\ &\le b_{m+1}T_m^p + (h\alpha^{-h})^p. \end{aligned} \tag{4.17}$$

By the definition of a_{m+1} in (4.14) and the fact that $tr(P_k^2) \ge d^{-1}(trP_k)^2$,

$$\begin{aligned} E[a_{m+1}|\mathcal{F}_{mh}] &\ge \frac{(h+1)\lambda_m tr(P_{mh+1}^2)}{h(1 + tr(P_{mh+1}))tr(P_{mh+1})} \\ &\ge \frac{(h+1)\lambda_m}{2hd}, \quad \text{on } \{tr(P_{mh+1}) \ge 1\}. \end{aligned}$$

Hence by the definition of b_{m+1},

$$E[b_{m+1}|\mathcal{F}_{mh}] \le \alpha^{(1-2h)p}\left(1 - \frac{(h+1)\lambda_m}{4hd}\right) I(tr(P_{mh+1}) \ge 1). \tag{4.18}$$

Denote

$$\alpha_{m+1} = \begin{cases} b_{m+1}, & \text{if } tr(P_{mh+1}) \ge 1; \\ \alpha^{(1-2h)p}\left(1 - \dfrac{(1+h)\lambda_m}{4hd}\right), & \text{otherwise.} \end{cases} \tag{4.19}$$

Then we have by (4.17)

$$T_{m+1}^p \le \alpha_{m+1}T_m^p + (h\alpha^{-h})^p. \tag{4.20}$$

By Condition 3.1, $\lambda_m \in S^0(\lambda)$ for some $\lambda \in (0,1)$. Since $\lambda_m \le h/(1+h)$, by Lemma 2.3 we know that $\{[(1+h)/(4hd)]\lambda_m\} \in S^0(\lambda^{(4hd)^{-1}})$. Hence by (4.18) and (4.19) and the assumption that $\lambda^{[16hd(2h-1)p]^{-1}} < \alpha$, it is easy to see that Lemma 3.1 is applicable to (4.20) and thus we get $\sup_m ET_m^p < \infty$. So Lemma 4.4 holds. □

THEOREM 4.3. *Consider the time-varying model* (1.1) *together with the forgetting factor algorithm* (1.3), (1.9), *and* (1.10). *Suppose that the following conditions are satisfied*:

(i) *Conditions* 3.1 *holds, i.e.*, $\lambda_m \in S^0(\lambda)$ *for some* $\lambda \in (0,1)$ *and some integer* $h > 0$, *where* λ_m *is defined by* (3.1);

(ii) *For some* $p \ge 1$

$$\sup_k(\|v_k\|_{L_{3p}} + \|\Delta_k\|_{L_{3p}}) \le \sigma_{3p};$$

(iii) $\sup_k \|\varphi_k\|_{L_{6p}} < \infty$;

(iv) *The forgetting factor* α *satisfies* $\lambda^{[48hd(2h-1)p]^{-1}} < \alpha < 1$, *where* d *is the dimension of* $\{\varphi_k\}$.

Then there exists a constant c such that

$$\limsup_{k\to\infty} \|\theta_k - \hat{\theta}_k\|_{L_p} \le c\sigma_{3p}.$$

Proof. We may complete the proof by using Theorem 2.4 just as it has been used for Theorem 4.1. However, in the present case the following analysis appears to be more straightforward.

By the matrix inverse formula, it follows from (1.10) that

$$P_{k+1}^{-1} = \alpha P_k^{-1} + \varphi_k \varphi_k^\tau. \tag{4.21}$$

Multiplying P_k^{-1} from both sides of (1.10) and using (1.9) we get $[I - L_k\varphi_k^\tau] = \alpha P_{k+1}P_k^{-1}$, and so

$$\prod_{j=i+1}^{k} (I - L_j\varphi_j^\tau) = \alpha^{k-i} P_{k+1} P_{i+1}^{-1}. \tag{4.22}$$

On the other hand, multiplying φ_k from both sides of (1.10) we have $P_{k+1}^{-1}L_k = \varphi_k$. Hence by (2.1) and (4.22),

$$\begin{aligned}\|\theta_{k+1} - \hat{\theta}_{k+1}\|_{L_p} &\le \alpha^k \|P_{k+1}P_0^{-1}\tilde{\theta}_0\|_{L_p} \\ &\quad + \sum_{i=0}^{k} \alpha^{k-i} (\|P_{k+1}\varphi_i v_i\|_{L_p} + \|P_{k+1}P_{i+1}^{-1}\Delta_{i+1}\|_{L_p}).\end{aligned}$$

By the Hölder inequality, Assumptions (i)–(iv), and Lemma 4.4 we know that the proof will be complete if we can show that $\sup_i \|P_{i+1}^{-1}\|_{L_{3p}} < \infty$. But, this can be easily seen from (4.21) and Assumption (iii), since

$$\|P_{k+1}^{-1}\|_{L_{3p}} \le \alpha \|P_k^{-1}\|_{L_{3p}} + \|\varphi_k\|_{L_{6p}}^2, \qquad \forall\, k \ge 0. \qquad \square$$

Remark 4.2. Under additional statistical assumptions on the processes $\{\varphi_k, v_k, \Delta_k\}$, a refined upper bound for the tracking error of the forgetting factor RLS can be derived (see [33]).

Conclusions. In this paper, stability and tracking error bounds are established for several standard estimation algorithms under a very general excitation condition. The various stability results presented in the paper are believed to be necessary preliminaries for further study of tracking properties, e.g., approximate expressions of the variance of the tracking errors (see e.g., [18]). Also, applications of the results to adaptive control systems as studied in e.g., [30] are possible. These issues will be discussed in detail elsewhere.

Appendix A.

Proof of Proposition 2.1. The solution of (2.2) may be expressed by

$$x_{n+1} = \sum_{i=0}^{n} \left[\prod_{j=i+1}^{n} (I - A_j) \right] \xi_{i+1}. \tag{A.1}$$

From this and the independence of $\{A_k\}$ and $\{\xi_k\}$ we know that the sufficiency of (2.4) is obvious.

To prove the necessity, we take $\{\xi_k\}$ to be an independently and identically distributed (i.i.d.) sequence with zero mean and unit variance. Then by denoting $B_k = I - A_k$, we have for some $c > 0$ and for any $n \geq k > 0$,

$$\begin{aligned} c &\geq E\|x_{n+1}\|^2 = tr \sum_{i=0}^{n} E\left[\prod_{j=i+1}^{n} B_j\right]\left[\prod_{j=i+1}^{n} B_j\right]^\tau \\ &\geq \sum_{i=k}^{n} tr\left\{E\left[\prod_{j=i+1}^{n} B_j\right]\left[\prod_{j=i+1}^{n} B_j\right]^\tau\right\}. \end{aligned} \tag{A.2}$$

Denote

$$a(n,i) = tr\left\{E\left[\prod_{j=i+1}^{n} B_j\right]\left[\prod_{j=i+1}^{n} B_j\right]^\tau\right\}.$$

It is easy to verify that $a(n,i) > 0$, for all $n \geq i$. Then by the independency of $\{A_j\}$ we have for any $n \geq i \geq k$,

$$\begin{aligned} a(n,k) &= trE[B_n \cdots B_{k+1} B_{k+1}^\tau \cdots B_n^\tau] \\ &= trE\{B_n \cdots B_{i+1} E[B_i \cdots B_{k+1} B_{k+1}^\tau \cdots B_i^\tau] B_{i+1}^\tau \cdots B_n^\tau\} \\ &\leq trE[B_n \cdots B_{i+1} B_{i+1}^\tau \cdots B_n^\tau] trE[B_i \cdots B_{k+1} B_{k+1}^\tau \cdots B_i^\tau] \\ &= a(n,i)a(i,k). \end{aligned}$$

Hence by (A.2) we have

$$c \geq a(n,k) \sum_{i=k}^{n} a^{-1}(i,k)$$

or

$$\sum_{i=k}^{n} a^{-1}(i,k) \leq c a^{-1}(n,k), \quad \forall n \geq k \geq 0. \tag{A.3}$$

From this we have

$$\begin{aligned} \sum_{i=k}^{n} a^{-1}(i,k) &= a^{-1}(n,k) + \sum_{i=k}^{n-1} a^{-1}(i,k) \\ &\geq \left(1 + \frac{1}{c}\right) \sum_{i=k}^{n-1} a^{-1}(i,k) \geq \cdots \\ &\geq \left(1 + \frac{1}{c}\right)^{n-k} a^{-1}(k,k) = \left(1 + \frac{1}{c}\right)^{n-k} d. \end{aligned}$$

Therefore, by (A.3)

$$c a^{-1}(n,k) \geq \left(1 + \frac{1}{c}\right)^{n-k} d$$

or

$$a(n,k) \le \frac{c}{d}\left(\frac{c}{1+c}\right)^{n-k}, \qquad \forall n \ge k, \quad \forall k \ge 0.$$

So (2.4) holds with $\lambda = [c/(1+c)]^{1/2} < 1$. □

Proof of Proposition 2.2. Denote $\psi(i,k) = \prod_{j=k+1}^{i}(I - A_j)$, and set for any fixed $k \ge 0$,

$$\xi_{i+1} = \psi(i,k)[E\psi(i,k)^\tau \psi(i,k)]^{-1/2}\eta_{i+1}, \tag{A.4}$$

where $\{\eta_{i+1}\}$ is a d-dimensional i.i.d. sequence independent of $\{A_i\}$ with $E\eta_{i+1} = 0$, $E\eta_i\eta_i^\tau = (1/d)I$. It is easy to see that

$$\begin{aligned} E\|\xi_{i+1}\|^2 &= tr[E\xi_{i+1}\xi_{i+1}^\tau] \\ &= \frac{1}{d}\, trE\{\psi(i,k)[E\psi(i,k)^\tau\psi(i,k)]^{-1}\psi(i,k)^\tau\} = 1. \end{aligned}$$

Hence for any $k \ge 0, \xi \in \mathcal{B}^0$. Substituting (A.4) into (A.1) and calculating the covariance, we get

$$Ex_{n+1}x_{n+1}^\tau = \frac{1}{d}E\sum_{i=0}^{n}\psi(n,k)[E\psi(i,k)^\tau\psi(i,k)]^{-1}\psi(n,k)^\tau, \qquad \forall k \ge 0,$$

and so

$$\begin{aligned} &[E\psi(n,k)^\tau\psi(n,k)]^{1/2}\sum_{i=0}^{n}[E\psi(i,k)^\tau\psi(i,k)]^{-1}[E\psi(n,k)^\tau\psi(n,k)]^{1/2} \\ &\le tr\, E\left\{\sum_{i=0}^{n}\psi(n,k)[E\psi(i,k)^\tau\psi(i,k)]^{-1}\psi(n,k)^\tau\right\} I \\ &= dE\|x_{n+1}\|^2 I \le cI, \quad \forall n \ge k, \quad \forall k, \end{aligned}$$

where for the last inequality we have used the assumption (2.8) and where c is a finite constant. This inequality implies that

$$\sum_{i=0}^{n}[E\psi(i,k)^\tau\psi(i,k)]^{-1} \le c[E\psi(n,k)^\tau\psi(n,k)]^{-1}.$$

Hence by denoting $a^{-1}(i,k) \triangleq \lambda_{\min}\{[E\psi(i,k)^\tau\psi(i,k)]^{-1}\}$, we obtain

$$\sum_{i=k}^{n}a^{-1}(i,k) \le \sum_{i=0}^{n}a^{-1}(i,k) \le ca^{-1}(n,k), \qquad \forall n \ge k \ge 0.$$

This inequality is exactly the same as (A.3). Hence by the same arguments as those in the proof of Proposition 2.1, we get

$$a(n,k) \le \frac{c}{d}\left(\frac{c}{1+c}\right)^{n-k}, \quad \forall n \ge k, \qquad \forall k \ge 0.$$

Finally, observing that $a(n,k) = \lambda_{\max}\{E[\psi(n,k)^\tau\psi(n,k)]\}$, we get the desired result (2.4). □

Appendix B.

Proof of (2.14). For simplicity of notations, set $k = mh + 1$. Following the ideas in the proofs of Theorem 4.5 and Lemma 10.7 in [5], we denote z_{k-1} as the unit eigenvector corresponding to the largest eigenvalue ρ_{k-1} of the matrix $E[\Phi^\tau(k+h,k)\Phi(k+h,k)|\mathcal{F}_{k-1}]$, and recursively define z_j by

$$z_j = (I - A_j)z_{j-1}, \qquad j \geq k. \tag{B.1}$$

It follows from (2.13) that $z_{k+h-1} = \Phi(k+h,k)z_{k-1}$. Hence we have

$$\begin{aligned} E(\|z_{k+h-1}\|^2|\mathcal{F}_{k-1}) &= z_{k-1}^\tau E[\Phi^\tau(k+h,k)\Phi(k+h,k)|\mathcal{F}_{k-1}]z_{k-1} \\ &= \rho_{k-1}\|z_{k-1}\|^2 = \rho_{k-1}. \end{aligned} \tag{B.2}$$

By (B.1) we have

$$z_j = z_{k-1} - \sum_{i=k}^{j} A_i z_{i-1}, \quad \forall\, j \in [k, k+h-1].$$

Hence by the Schwarz inequality

$$\begin{aligned} E[\|z_{j-1} - z_{k-1}\|^2|\mathcal{F}_{k-1}] &= E\left[\left\|\sum_{i=k}^{j-1} A_i z_{i-1}\right\|^2 |\mathcal{F}_{k-1}\right] \\ &\leq E\left[\left(\sum_{i=k}^{j-1}\|A_i^{1/2}z_{i-1}\|^2\right)\sum_{i=k}^{j-1}\|A_i^{1/2}\|^2|\mathcal{F}_{k-1}\right] \\ &\leq hE\left[\sum_{i=k}^{j-1} z_{i-1}^\tau A_i z_{i-1}|\mathcal{F}_{k-1}\right], \qquad j \in [k, k+h]. \end{aligned} \tag{B.3}$$

By the definition of λ_m and the Minkowski inequality we have

$$\begin{aligned} &\sqrt{(1+h)}\lambda_m^{1/2} \\ &\leq \left\{ z_{k-1}^\tau E\left[\sum_{i=mh+1}^{(m+1)h} A_i|\mathcal{F}_{mh}\right] z_{k-1}\right\}^{1/2} = \left\{E\left[\sum_{i=k}^{k+h-1}\|A_i^{1/2}z_{k-1}\|^2|\mathcal{F}_{k-1}\right]\right\}^{1/2} \\ &\leq \left\{E\left[\sum_{i=k}^{k+h-1}\|A_i^{1/2}z_{i-1}\|^2|\mathcal{F}_{k-1}\right]\right\}^{1/2} + \left\{E\left[\sum_{i=k}^{k+h-1}\|z_{i-1}-z_{k-1}\|^2|\mathcal{F}_{k-1}\right]\right\}^{1/2}. \end{aligned}$$

From this and (B.3) it follows that

$$\sqrt{(1+h)}\lambda_m^{1/2} \leq (1+h)\left\{E\left[\sum_{i=k}^{k+h-1}\|A_i^{1/2}z_{i-1}\|^2|\mathcal{F}_{k-1}\right]\right\}^{1/2}$$

or

$$E\left[\sum_{i=k}^{k+h-1}\|A_i^{1/2}z_{i-1}\|^2|\mathcal{F}_{k-1}\right] \geq \frac{\lambda_m}{1+h}. \tag{B.4}$$

By (B.1) and the fact that $0 \le A_i \le I$ it is easily derived that

$$z_j^\tau z_j \le z_{j-1}^\tau z_{j-1} - z_{j-1}^\tau A_j z_{j-1},$$

from which we have

$$\begin{aligned}\|z_{k+h-1}\|^2 &\le \|z_{k-1}\| - \sum_{i=k}^{k+h-1} z_{i-1}^\tau A_i z_{i-1} \\ &= 1 - \sum_{i=k}^{k+h-1} z_{i-1}^\tau A_i z_{i-1}.\end{aligned}$$

Combining this with (B.2) and (B.4) we get

$$\begin{aligned}\rho_{k-1} &= E[\|z_{k+h-1}\|^2 | \mathcal{F}_{k-1}] \\ &\le 1 - E\left[\sum_{i=k}^{k+h-1} z_{i-1}^\tau A_i z_{i-1} | \mathcal{F}_{k-1}\right] \le 1 - \frac{\lambda_m}{1+h},\end{aligned}$$

which is tantamount to (2.14). □

Appendix C. We first prove Proposition 3.3. The proof is divided into two steps.

Step 1. We first prove that

$$\lambda_k \ge \frac{1}{P(\beta_{kp-1})[1 + \|\varphi_{kp}\|^4]}, \quad \forall k \ge 1 \tag{C.1}$$

where λ_k is defined by (3.1) with $h = p$ and $\mathcal{F}_k = \sigma\{\mathcal{F}_i', v_i, i \le k-1\}$ and where $P(x)$ is a polynomial of x with nonnegative coefficients.

By (3.7) we have

$$\varphi_{k+1} = \left(\prod_{j=s}^{k} A_j\right)\varphi_s + \sum_{i=s}^{k}\left(\prod_{j=i+1}^{k} A_j\right) b v_i, \quad \forall k \ge s, \quad \forall s \ge 0. \tag{C.2}$$

By (3.1) with $h = p$ and the Schwarz inequality it is easy to show that (cf. [14] or [28], p. 168)

$$\begin{aligned}\lambda_k &\ge \frac{1}{p+1}\lambda_{\min}\left\{E\left[\frac{\varphi_{(k+1)p}\varphi_{(k+1)p}^\tau}{1 + \|\varphi_{(k+1)p}\|^2} \,\middle|\, \mathcal{F}_{kp}\right]\right\} \\ &\ge \frac{1}{1+p}\frac{\{\lambda_{\min}(E[\varphi_{(k+1)p}\varphi_{(k+1)p}^\tau | \mathcal{F}_{kp}])\}^2}{E[(\|\varphi_{(k+1)p}\|^4 + \|\varphi_{(k+1)p}\|^2) | \mathcal{F}_{kp}]}.\end{aligned} \tag{C.3}$$

We first analyze the numerator. Denote the controllability Gramian by H_{kp+1}:

$$H_{kp+1} \triangleq \sum_{i=kp}^{(k+1)p-1}\left(\prod_{j=i+1}^{(k+1)p-1} A_j\right) bb^\tau \left(\prod_{j=i+1}^{(k+1)p-1} A_j\right)^\tau. \tag{C.4}$$

Then by (C.2), (3.6), and the independence assumptions we have

$$E[\varphi_{(k+1)p}\varphi_{(k+1)p}^\tau | \mathcal{F}_{kp}] \ge \sigma_v^2 E[H_{kp+1} | \mathcal{F}_{kp}]. \tag{C.5}$$

By (3.8) and (C.4) it is easy to verify that $det[H_{kp+1}] = 1$, and hence by (C.5)

$$\begin{aligned}\lambda_{\min}\{E[\varphi_{(k+1)p}\varphi_{(k+1)p}^{\tau}|\mathcal{F}_{kp}]\} &\geq \sigma_v^2 E[\lambda_{\min}(H_{kp+1})|\mathcal{F}_{kp}] \\ &\geq \sigma_v^2 E\left[\frac{det(H_{kp+1})}{\{\lambda_{\max}(H_{kp+1})\}^{p-1}}\,\middle|\,\mathcal{F}_{kp}\right] \\ &\geq \frac{\sigma_v^2}{E\{\|H_{kp+1}\|^{p-1}|\mathcal{F}_{kp}\}}.\end{aligned} \tag{C.6}$$

Concerning the denominator in (C.3), we first note that

$$E[\|\varphi_{(k+1)p}\|^2|\mathcal{F}_{kp}] \leq \frac{1}{2} + \frac{1}{2}E[\|\varphi_{(k+1)p}\|^4|\mathcal{F}_{kp}] \tag{C.7}$$

and that by (C.2)

$$\begin{aligned}E[\|\varphi_{(k+1)p}\|^4|\mathcal{F}_{kp}] \leq\ & 8E\left[\left\|\prod_{i=kp}^{(k+1)p-1} A_i\right\|^4 \,\middle|\, \mathcal{F}_{kp}\right]\|\varphi_{kp}\|^4 \\ &+ 8p^3\|b\|^4 \sup_k Ev_k^4 \sum_{i=kp}^{(k+1)p-1} E\left[\left\|\prod_{j=i+1}^{(k+1)p-1} A_j\right\|^4 \,\middle|\, \mathcal{F}_{kp}\right].\end{aligned}$$

Then, substituting this, (C.6), and (C.7) into (C.3) and using (3.10)–(3.11) together with the Markovian properties of β_k, it is not difficult to conclude (C.1). □

Step 2. We prove that

$$\left\{\frac{1}{P(\beta_{kp-1})[1 + \|\varphi_{kp}\|^4]}\right\} \in S^0(\lambda), \quad \text{for some } \lambda \in (0,1). \tag{C.8}$$

Since A is a stable matrix, there is a norm $\|\cdot\|_\delta$ on $\mathbb{R}^p$ such that its induced norm on $\mathbb{R}^{p\times p}$ (also denoted by $\|\cdot\|_\delta$) satisfies $\|A\|_\delta \triangleq \delta < 1$. Clearly, there is a constant $c > 1$ such that, for all $x \in \mathbb{R}^p$, $\|x\| \leq c\|x\|_\delta$. In order to apply Corollary 3.1 we denote

$$\begin{aligned}x_k &= \|\varphi_{kp}\|_\delta^8 + \beta_{kp-1}^L + 1 \\ y_k &= \frac{c^8}{2}P^2(\beta_{kp-1}), \qquad \mathcal{G}_k = \sigma\{\mathcal{F}_i', v_i, i \leq kp-1\}\end{aligned}$$

where L is a suitably large number defined later on. Then both$\{x_k, \mathcal{G}_k\}$ and $\{y_k, \mathcal{G}_k\}$ are adapted processes. Clearly, $P(\beta_{kp-1})[1 + \|\varphi_{kp}\|^4] \leq x_k + y_k$. Hence, (C.8) will be proved if conditions in Corollary 3.1 can be verified.

By (3.11) and the convexity of the function $P^2(x), x \geq 0$, we have

$$\begin{aligned}y_{k+1} &\leq \frac{c^8}{2}P^2\left(\beta^p\beta_{kp-1} + \sum_{i=kp}^{(k+1)p-1} e_i\right) \\ &\leq \frac{c^8}{2}P^2\left(\beta\beta_{kp-1} + (1-\beta)\frac{1}{1-\beta}\sum_{i=kp}^{(k+1)p-1} e_i\right) \\ &\leq \beta y_k + \frac{c^8}{2}(1-\beta)P^2\left(\frac{1}{1-\beta}\sum_{i=kp}^{(k+1)p-1} e_i\right).\end{aligned}$$

Hence $\{y_k\}$ satisfies the required properties.

Now, it only remains to prove that $\{x_k\}$ satisfies conditions in Lemma 3.1. By (3.10)–(3.11) we have

$$\|A_k\|_\delta \le \delta + \|\bar{A}_k\|_\delta \le \delta + c_\delta \beta_k,$$

where $c_\delta > 0$ is a constant. This motivates us to set $\alpha_{k+1} \triangleq \prod_{i=kp}^{(k+1)p-1}(\delta + c_\delta \beta_i)$. Clearly, $\alpha_k \in \mathcal{G}_k$, and by a completely similar argument as that used in [29] we know that under condition (3.12) (with small ε and large b) there are constants $M > 0, \gamma \in (0,1)$ such that

$$\left\| \prod_{k=m}^{n} E[\alpha_{k+1}^4 | \mathcal{G}_k] \right\| \le M\gamma^{n-m+1}, \quad \forall n \ge m, \quad \forall m \ge 0.$$

Let α be a positive number such that $(1+\alpha)^4\gamma < 1$, where γ is defined above. It is easy to see from (C.2) and the definition of α_{k+1} that there is a constant $M_\alpha > 0$ such that for any $\varepsilon_0 > 0$,

$$\text{(C.9)} \quad \|\varphi_{(k+1)p}\|_\delta^8 \le (1+\alpha)\alpha_{k+1}\|\varphi_{kp}\|_\delta^8 + M_\alpha \left\| \sum_{i=kp}^{(k+1)p-1} \left(\prod_{j=i+1}^{(k+1)p-1} A_j \right) b v_i \right\|_\delta^8$$

$$\le (1+\alpha)\alpha_{k+1}\|\varphi_{kp}\|_\delta^8 + \frac{\varepsilon_0}{2}\beta_{kp-1}^L$$

$$\text{(C.10)} \quad + c_1 \left(\sum_{i=kp}^{(k+1)p-1} e_i + 1 \right)^L + c_2 \left(\sum_{i=kp}^{(k+1)p-1} |v_i| + 1 \right)^9$$

for some constants L, c_1, and c_2, where for the last inequality we have used the fact that $\|A_j\|_\delta \le \delta + c_\delta \beta_j$ together with the Markovian property of $\{\beta_j\}$.

Without loss of generality we may assume that L in (C.10) is so large that

$$\text{(C.11)} \quad 4\beta^{pL} \le (1+\alpha)\delta^{8p} \triangleq \varepsilon_0.$$

By (C.11) and (3.11) it is easy to see that there is a constant $c_3 > 0$ such that

$$\text{(C.12)} \quad \beta_{kp-1}^L \le \frac{\varepsilon_0}{2}\beta_{(k-1)p-1}^L + c_3 \left(\sum_{i=(k-1)p}^{kp-1} e_i \right)^L.$$

Combining (C.11) and (C.12), using the definition of x_k and the fact that $(1+\alpha)\alpha_{k+1} \ge \varepsilon_0$, we get for some constant $c_4 > 0$,

$$x_{k+1} \le (1+\alpha)\alpha_{k+1}x_k + c_4 \left[1 + \left(\sum_{i=kp}^{(k+1)p-1} e_i + 1 \right)^L + \left(\sum_{i=kp}^{(k+1)p-1} |v_i| + 1 \right)^9 \right].$$

Hence both $\{x_k\}$ and $\{y_k\}$ satisfy conditions of Corollary 3.1, and so $\{1/(x_k + y_k)\} \in S^0(\lambda)$ for some $\lambda \in (0,1)$. This proves (C.8). Finally, combining (C.1) and (C.8) we know that Proposition 3.3 is true. □

Proof of Example 3.2. Set $\mathcal{F}_k = \sigma\{A_i, v_i, i \le k-1\}$. Since $\{A_i\}$ is an independent sequence, similar to the proof of (C.1) we have for some constant $c_5 \ge 1$,

$$\lambda_k \ge \frac{1}{a_k}, \qquad a_k = c_5(1 + \|\varphi_{kp}\|^4), \qquad k \ge 0. \tag{C.13}$$

So we need only to prove that $\{a_k\}$ verifies (3.3). Let $\alpha > 0$ be such that $(1+\alpha)\delta^4 < 1$ where δ is given by (3.9). By (C.2), (3.9) and the independence assumptions we know that

$$E[\|\varphi_{(k+1)p}\|^4 | \mathcal{F}_{kp}] \le (1+\alpha)\delta^4 \|\varphi_{kp}\|^4 + c_6$$

for some constant $c_6 > 0$. Consequently, we have

$$E[a_{k+1} | \mathcal{F}_{kp}] \le (1+\alpha)\delta^4 \alpha_k + c_5(1+c_6).$$

Hence (3.3) is true. □

Appendix D.

Proof of Lemma 4.1. We first consider the case where $\beta > 2\max(1, \alpha)$. Let $\delta_p > 0$ be such that

$$\sup_{n \ge k \ge 0} \|(d_{nk}\xi_k)\log^{\beta/2}(e + d_{nk}\xi_k)\|_{L_p} \le \delta_p.$$

Then exactly the same argument as that used for Lemma 8 in [28] yields

$$\sum_{k=0}^{n} \|c_{nk} d_{nk} \xi_k\|_{L_p} \le c[\delta_p \log(e + \delta_p^{-1})]. \tag{D.1}$$

So we need only to find a relationship between δ and σ. By inequality (52) in [28] we know that

$$xy \le \sigma \exp(\varepsilon x^{1/\alpha}) + c_1 y[\log^{\alpha}(e + \sigma^{-1}) + \log^{\alpha}(e + y)] \tag{D.2}$$

holds for all $\sigma > 0, \varepsilon > 0$, and $\alpha > 0$, where c_1 is a constant depending only on ε and α. Applying (D.2) with $x = d_{nk}^p \log^{p\beta/2}(e + d_{nk}), y = \xi_k^p \log^{p\beta/2}(e + \xi_k), \sigma = \sigma_p^p, \alpha = p\beta/2$ we have

$$\begin{aligned} & Ed_{nk}^p \xi_k^p \log^{p\beta/2}(e + d_{nk}\xi_k) \le 2^{p\beta/2} Exy \\ & \le 2^{p\beta/2} E\{\sigma_p^p \exp(\varepsilon x^{2/(p\beta)}) + c_1 y[\log^{p\beta/2}(e + \sigma_p^{-p}) + \log^{p\beta/2}(e+y)]\} \\ & \le c\sigma_p^p \log^{p\beta/2}(e + \sigma_p^{-1}), \quad \text{for some constant } c. \end{aligned} \tag{D.3}$$

Hence we may take $\delta_p = c\sigma_p \log^{\beta/2}(e + \sigma_p^{-1})$. Substituting this into (D.1) we know that the first case in (4.2) is true, while the second case can be proved in a similar way. Finally, the last case can be derived from (D.1) by noting $\delta_p \le c_1 \sigma_p$ for some $c_1 > 0$. □

REFERENCES

[1] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.

[2] B. Widrow and S. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[3] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[4] R. S. Lipster and A. N. Shiryayev, *Statistics of Random Processes II*, Springer-Verlag, New York, 1977.

[5] H. F. CHEN AND L. GUO, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.
[6] O. MACCHI AND E. EWEDA, *Compared speed and accuracy of RLS and LMS algorithms with constant forgetting factors*, Traitement Signal, 22 (1988), pp. 255–267.
[7] B. WIDROW, J. M. MCCOOL, M. G. LARIMORE, AND C. R. JOHNSON, JR., *Stationary and nonstationary learning characteristics of the LMS adaptive filter*, Proc. IEEE, 64 (1976), pp. 1151–1162.
[8] B. BITMEAD, *Convergence properties of LMS adaptive estimators with unbounded dependent inputs*, IEEE Trans. Automat. Control, 29 (1984), pp. 477–479.
[9] O. MACCHI, *Optimization of adaptive identification for time-varying filters*, IEEE Trans. Automat. Control, 31 (1986), pp. 283–287.
[10] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes With Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
[11] A. BENVENISTE, *Design of adaptive algorithms for the tracking of time-varying systems*, Internat. J. Adapt. Control Signal Process., 1 (1987), pp. 1–29.
[12] V. SOLO, *The limit behavior of LMS*, IEEE Trans. Acoustics, Speech, Signal Process., 37 (1989), pp. 1909–1922.
[13] L. GUO, L. G. XIA, AND J. B. MOORE, *Tracking randomly varying parameters: analysis of a standard algorithm*, in Proc. 27th IEEE CDC, Austin, TX, 1988, pp. 1514–1519.
[14] L. GUO, *Estimating time-varying parameters by Kalman filter based algorithm: stability and convergence*, IEEE Trans. Automat. Control, 35 (1990), pp. 141–147.
[15] G. KITAGAWA AND W. GERSCH, *A smoothness priors time varying AR coefficient modelling of nonstationary covariance time series*, IEEE Trans. Automat. Control, 30 (1985), pp. 48–56.
[16] M. NIEDŹWIECKI AND L. GUO, *Nonasymptotic results for finite-memory WLS filters*, IEEE Trans. Automat. Control, 36 (1991), pp. 198–206.
[17] S. BITTANTI AND M. CAMPI, *Adaptive RLS algorithms under stochastic excitation—L^2 convergence analysis*, IEEE Trans. Automat. Control, 36 (1991), pp. 963–967.
[18] L. LJUNG AND P. PRIOURET, *A result on the mean square error obtained using general tracking algorithms*, Internat. J. Adapt. Control Signal Process., 5 (1991), pp. 231–250.
[19] B.D.O. ANDERSON, *Internal and external stability of linear time-varying systems*, SIAM J. Control Optim., 20 (1982), pp. 408–413.
[20] H. FURSTENBERG AND H. KESTEN, *Products of random matrices*, Ann. Math. Statist., 31 (1960), pp. 457–469.
[21] E. F. INFANTE, *On the stability of some linear nonautonomous random systems*, J. Appl. Mech. 35 (1968), pp. 7–12.
[22] G. BLANKENSHIP, *Stability of linear differential equations with random coefficients*, IEEE Trans. Automat. Control, 22 (1987), pp. 834–838.
[23] R. Z. HASMINSKII, *Stochastic Stability of Differential Equations*, Sijthoff and Nordhoff., Alphen aan den Rijn, The Netherlands, 1980.
[24] S. GEMAN, *Some averaging and stability results for random differential equations*, SIAM J. Appl. Math., 36 (1979), pp. 87–107.
[25] L. ARNOLD AND V. WIHSTUTZ, EDS., *Lyapunov exponents*, Lecture Notes in Math. 1186, Springer-Verlag, New York, Berlin, 1987.
[26] L. GERENCSER, *Almost sure exponential stability of random linear differential equations*, Stochastics Stochastics Rep., 36 (1991), pp. 91–107.
[27] P. E. CAINES, *Linear Stochastic Systems*, Wiley, New York, 1988.
[28] J. F. ZHANG, L. GUO, AND H. F. CHEN, *L_p-stability of estimation errors of Kalman filter for tracking time-varying parameters*, Internat. J. Adapt. Control Signal Process., 5 (1991), pp. 155–174.
[29] S. P. MEYN AND L. GUO, *Geometric ergodicity of a bilinear time series model*, J. Time Ser. Anal., 14 (1993), pp. 93–108.
[30] L. GUO, *On adaptive stabilization of time-varying stochastic systems*, SIAM J. Control Optim., 28 (1990), pp. 1432–1451.
[31] S. P. MEYN AND P. E. CAINES, *A new approach to stochastic adaptive control*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 220–226.
[32] S. P. MEYN AND L. BROWN, *Model Reference Adaptive Control of Time-Varying and Stochastic Systems*, Tech. Rep., Coordinated Science Laboratory, University of Illinois, Urbana, IL, 1992.
[33] L. GUO, L. LJUNG, AND P. PRIOURET, *Tracking performance analysis of the forgetting factor RLS algorithm*, in Proc. 31st IEEE CDC, Tucson, AZ, 1992.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1226–1246, September 1994

002

UNIFORM EXPONENTIAL STABILITY AND APPROXIMATION IN CONTROL OF A THERMOELASTIC SYSTEM*

ZHUANGYI LIU† AND SONGMU ZHENG‡

Abstract. This paper has two objectives. First, necessary and sufficient conditions are given to characterize the uniform exponential stability of a sequence of c_0-semigroups $T_n(t)$ on Hilbert space H_n. Secondly, approximation in control of a one-dimensional thermoelastic system, subject to Dirichlet–Dirichlet as well as Dirichlet–Neumann boundary conditions, is considered. The uniform exponential stability and strong convergence of corresponding semigroups associated with approximate scheme are proved. Numerical experimental results are also presented.

Key words. linear thermoelastic system, uniform exponential stability, semigroup, approximation in control

AMS subject classifications. 93C20, 93D20, 73C25

1. Introduction. For a homogeneous rod with uniform cross section, in general, the equation of one-dimensional linear thermoelasticity can be written as (see [D])

$$u_{tt} - c^2 u_{xx} + c^2 \gamma \theta_x = 0, \qquad (0,\pi) \times (0,+\infty), \tag{1.1}$$

$$\theta_t + \gamma u_{xt} - \theta_{xx} = 0, \qquad (0,\pi) \times (0,+\infty), \tag{1.2}$$

where u is proportional to the displacement and θ is the relative temperature about the stress-free reference temperature. The constants $\gamma > 0$ and $c > 0$ represent, respectively, the amount of thermal-mechanical coupling and the small-amplitude wave speed about a constant temperature state. (See [D] for a precise definition of γ and c.) In most materials of interest, γ is several orders of magnitude smaller than 1.

By introducing new variable (velocity)

$$v = u_t, \tag{1.3}$$

(1.1), (1.2) is reduced to the following abstract first-order evolution equation:

$$\frac{dz}{dt} = \mathcal{A}z \tag{1.4}$$

with

$$z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} \equiv \begin{pmatrix} u \\ v \\ \theta \end{pmatrix}, \tag{1.5}$$

and

$$\mathcal{A} = \begin{pmatrix} 0 & I & 0 \\ c^2 D^2 & 0 & -c^2 \gamma D \\ 0 & -\gamma D & D^2 \end{pmatrix}. \tag{1.6}$$

Here we have used the notation $D = \partial/\partial x$, $D^2 = \partial^2/\partial x^2$.

* Received by the editors September 9, 1991; accepted for publication (in revised form) March 1, 1993.

† Department of Mathematics and Statistics, University of Minnesota, Duluth, Minnesota 55812. This author was supported in part by Grant-in-Aid 0350-3744-04 and a summer research fellowship of the Graduate School of the University of Minnesota.

‡ Institute of Mathematics, Fudan University, Shanghai 200433, China. This author's research was carried out while visiting the University of Minnesota at Duluth.

If both ends of the rod are clamped and kept at the reference temperature, then we have the Dirichlet–Dirichlet boundary conditions

$$u|_{x=0,\pi} = \theta|_{x=0,\pi} = 0, \quad \text{for } t > 0. \tag{1.7}$$

Let the state space be

$$H = H_0^1(\Omega) \times L^2(\Omega) \times L^2(\Omega) \tag{1.8}$$

equipped with the norm

$$\|z\|_H = \left(\|Dz_1\|_{L^2}^2 + \frac{1}{c^2} \|z_2\|_{L^2}^2 + \|z_3\|_{L^2}^2 \right)^{1/2}. \tag{1.9}$$

It is shown in [BLM] that the operator $\mathcal{A}$ with $\mathcal{D}(\mathcal{A}) = H^2 \cap H_0^1 \times H_0^1 \times H^2 \cap H_0^1$ generates a c_0-semigroup $T(t)$ on the Hilbert space H. Our study is motivated by the linear quadratic Gaussian (LQG) optimal control problem of the thermoelastic system and its approximation; we refer to [GRT] for the detailed description. Recall that a c_0-semigroup $T(t)$ on a Hilbert space H is said to be exponentially stable if there exist positive constants M and α such that

$$\|T(t)\|_{\mathcal{L}(H,H)} \leq Me^{-\alpha t} \quad \forall t > 0. \tag{1.10}$$

Accordingly, a sequence of C_0-semigroups $T_n(t)$ on the Hilbert spaces H_n are said to be uniformly exponentially stable if

$$\|T_n(t)\|_{\mathcal{L}(H_n,H_n)} \leq Me^{-\alpha t} \quad \forall t > 0, \quad \forall n. \tag{1.11}$$

As shown in [GRT], the exponential stability of semigroup $T(t)$ associated with the open-loop system (1.4) and the uniformly exponential stability of approximating semigroups $T_n(t)$, if such results are available, will play a very important role in the study of the corresponding LQG optimal control problem and the convergence of the approximating optimal controls. The exponential stability of the semigroup associated with the thermoelastic system (1.4) subject to Dirichlet–Neumann boundary condition was proved by Hansen [Ha] in 1990. Another approach of the proof was given by Gibson, Rosen, and Tao (see [GRT]). But the problem for the Dirichlet–Dirichlet boundary conditions had remained open until Kim [K] and Liu and Zheng [LZ] independently suceeded in proving exponential energy decay rate for this case.

The purpose of this paper is to study the uniform exponentially stable approximation and its application to the thermoelastic system (1.4) with Dirichlet–Dirichlet boundary conditions. As mentioned before, it is crucial to show that (1.11) holds. Let us explain the difficulties in proving (1.11). For the wave equation with internal or boundary friction damping, the dissipation is relatively strong so that the energy method can be applied to obtain the exponential stability (1.10) (see [C], [L1], [L2]) as well as the uniformly exponential stability (1.11) for the approximation (see [BIW]). However, the dissipation in the thermoelastic system, due to heat conduction, is much weaker. To our knowledge, the energy method has been used to obtain the exponential stability in more regular Sobolev's space $\mathcal{D}(\mathcal{A})$ as Slemrod [S] and Rivera [R] did, but not in the primary Hilbert space H, at least for the case of Dirichlet–Dirichlet boundary conditions. As far as (1.11) is concerned, the existence of a positive constant M independent of n is also a difficult part of the proof. As the counterexample in [Hu] shows, even for semigroups $T_n(t) = e^{\mathcal{A}_n t}$ with $\mathcal{A}_n$ being an $n \times n$ matrix, uniform negative boundedness away from zero of the spectrum $\sigma(\mathcal{A}_n)$ of $\mathcal{A}_n$ does not guarantee the existence of such an M. As can be seen in [GRT], the existence of a uniform constant M is somehow related

to the independence of eigenvectors of $\mathcal{A}_n$. In the Dirichlet–Neumann boundary conditions case, we could indeed use such information coming from decoupling to prove (1.11). But in the case of Dirichlet–Dirichlet boundary conditions a new approach must be adopted.

The paper is organized as follows. In §2 we give necessary and sufficient conditions to characterize the uniform exponential stability of a sequence of semigroups $T_n(t)$. In §3, as an application of the results in §2, the uniform exponential stability (1.11) and convergence are proved for the so-called modal approximation of the thermoelastic system subject to Dirichlet–Dirichlet or Dirichlet–Neumann boundary conditions. In §4 numerical experimental results are presented.

2. Uniform exponential stability of C_0-semigroups. In this section we give necessary and sufficient conditions to characterize the uniform exponential stability of $T_n(t)$, a sequence of c_0-semigroups on Hilbert spaces H_n.

For a single semigroup $T(t)$, the characteristic condition of exponential stability was given by Huang [Hu]. In what follows we extend his results to a sequence of semigroups $T_n(t)$.

THEOREM 2.1. *Let $T_n(t)$ $(n = 1, \ldots)$ be a sequence of c_0-semigroups of operators on the Hilbert spaces H_n and let $\mathcal{A}_n$ be the corresponding infinitesimal generators. Then $T_n(t)$ are uniformly exponentially stable if and only if the following three conditions hold*:

$$\sup_{n\in N}\{\operatorname{Re}\lambda;\ \lambda\in\sigma(\mathcal{A}_n)\}=\sigma_0<0; \tag{2.1}$$

there exist $\sigma \in (\sigma_0, 0)$ such that

$$\sup_{\operatorname{Re}\lambda\geq\sigma,n\in N}\{\|(\lambda I-\mathcal{A}_n)^{-1}\|\}=M_0<\infty; \tag{2.2}$$

and there exist $M_1 > 0$ such that

$$\|T_n(t)\|_{\mathcal{L}(H_n,H_n)}\leq M_1<\infty\quad\forall t>0,\qquad n\in N. \tag{2.3}$$

We postpone the proof until the end of this section.

THEOREM 2.2. *Let $T_n(t)$ $(n = 1, \ldots)$ be a sequence of semigroups of operators on the Hilbert spaces H_n and let $\mathcal{A}_n$ be the corresponding infinitesimal generators. Then $T_n(t)$ are uniformly exponentially stable if and only if* (2.1), (2.3), *and*

$$\sup_{\operatorname{Re}\lambda\geq 0,n\in N}\{\|(\lambda I-\mathcal{A}_n)^{-1}\|\}<\infty \tag{2.4}$$

hold.

Proof. We only need to prove that (2.1), (2.3), and (2.4) imply (2.1)–(2.3). Let

$$M=\sup_{\operatorname{Re}\lambda\geq 0,n\in N}\{\|\lambda I-\mathcal{A}_n)^{-1}\|\}<\infty, \tag{2.5}$$

and $\lambda = (\tau + \boldsymbol{i}\omega)$, $\tau \in [-1/2M, 0]$. Then

$$\begin{aligned}\|(I+\tau(\boldsymbol{i}\omega I-\mathcal{A}_n)^{-1})x\| &\geq \|x\|-\frac{1}{2M}\|(\boldsymbol{i}\omega I-\mathcal{A}_n)^{-1}x\|\\ &\geq \|x\|-\frac{1}{2}\|x\|=\frac{1}{2}\|x\|.\end{aligned} \tag{2.6}$$

This implies that

$$\|(I+\tau(\boldsymbol{i}\omega I-\mathcal{A}_n)^{-1})^{-1}\|\leq 2. \tag{2.7}$$

By

$$\lambda I - \mathcal{A}_n = (I + \tau(\boldsymbol{i}\omega I - \mathcal{A}_n)^{-1})(\boldsymbol{i}\omega I - \mathcal{A}_n), \tag{2.8}$$

we conclude that $\lambda I - \mathcal{A}_n$ is invertible and

$$\begin{aligned} \|(\lambda I - \mathcal{A}_n)^{-1}\| &= \|(\boldsymbol{i}\omega I - \mathcal{A}_n)^{-1}(I + \tau(\boldsymbol{i}\omega I - \mathcal{A}_n)^{-1})^{-1}\| \\ &\le 2\|(\boldsymbol{i}\omega I - \mathcal{A}_n)^{-1}\| \le 2M. \end{aligned} \tag{2.9}$$

Let

$$\sigma = \max\left(-\frac{1}{2M}, \frac{\sigma_0}{2}\right). \tag{2.10}$$

Then

$$\sigma_0 < \sigma < 0, \quad \sigma \in \left[-\frac{1}{2M}, 0\right). \tag{2.11}$$

It turns out from (2.9) that (2.2) is satisfied. □

In particular, (2.3) holds if $T_n(t)$ is a sequence of semigroups of contraction. Thus we have the following result.

COROLLARY 2.3. *Let $T_n(t)$ be a sequence of semigroups of contraction on the Hilbert spaces H_n and $\mathcal{A}_n$ be the corresponding infinitesimal generators. Then $T_n(t)$ are uniformly exponentially stable if and only if* (2.1) *and* (2.4) *hold.*

In what follows we give the proof of Theorem 2.1.

Proof of Theorem 2.1. If $T_n(t)$ are uniformly exponentially stable, i.e., there exist M, $\alpha > 0$ such that

$$\|T_n(t)\|_{\mathcal{L}(H_n, H_n)} \le Me^{-\alpha t} \quad \forall t > 0, \qquad n \in N, \tag{2.12}$$

then

$$\omega_0(\mathcal{A}_n) \overset{\text{def}}{=} \lim_{t \to +\infty} \frac{\ln \|T_n(t)\|}{t} \le -\alpha. \tag{2.13}$$

Thus (2.1) follows from the following property:

$$\sigma_0(\mathcal{A}_n) \overset{\text{def}}{=} \sup_{n \in N} \{\text{Re } \lambda; \lambda \in \sigma(\mathcal{A}_n)\} \le \omega_0(\mathcal{A}_n) \le -\alpha. \tag{2.14}$$

Let $\sigma = -\alpha/2$. Then $\sigma_0 < \sigma < 0$. For Re $\lambda \ge \sigma$, we have

$$\begin{aligned} \|(\lambda I - \mathcal{A}_n)^{-1}x\| &= \left\| \int_0^{+\infty} e^{-\lambda t} T_n(t)x \, dt \right\| \\ &\le M\|x\| \int_0^{+\infty} e^{-\text{Re}\lambda t} e^{-\alpha t} \, dt = \frac{M\|x\|}{\alpha + \text{Re } \lambda} \le 2\frac{M\|x\|}{\alpha}. \end{aligned} \tag{2.15}$$

This implies that (2.2) holds. Furthermore, (2.3) immediately follows from (2.12). Thus the proof of the "only if" part is complete.

On the other hand, suppose (2.1)–(2.3) hold. Let

$$\tilde{\mathcal{A}}_n = \mathcal{A}_n - \frac{\sigma}{2} I. \tag{2.16}$$

Then

$$\sup_{n\in N}\{\operatorname{Re}\lambda;\lambda\in\sigma(\tilde{\mathcal{A}}_n)\}\le-\frac{\sigma}{2}+\sigma_0<\frac{\sigma}{2}<0, \tag{2.17}$$

and

$$\sup_{\operatorname{Re}\lambda\ge(\sigma/2),n\in N}\{\|(\lambda I-\tilde{\mathcal{A}}_n)^{-1}\|\}=M_0<\infty. \tag{2.18}$$

In what follows we prove that (2.17), (2.18), and (2.3) imply that there exists a positive constant $M>0$ independent of n such that the corresponding semigroups $\tilde{T}_n(t)=T_n(t)e^{-(\sigma/2)t}$ with infinitesimal generators $\tilde{\mathcal{A}}_n$ satisfy

$$\|\tilde{T}_n(t)\|\le M, \tag{2.19}$$

which results in (2.12) with $\alpha=-\frac{\sigma}{2}>0$.

To prove (2.19), we use the same technique as Huang did in [Hu]. First, by (2.3) we have

$$\|\tilde{T}_n(t)\|\le M_1e^{-(\sigma/2)t}. \tag{2.20}$$

Therefore

$$\omega_0(\tilde{\mathcal{A}}_n)\le-\frac{\sigma}{2}. \tag{2.21}$$

Our next step is to prove the following estimate:

$$|(\tilde{T}_n(t)x,y)|\le\frac{cM_1^2}{2\pi}\|x\|\,\|y\|\quad\text{for } t\ge1,\qquad x,y\in H_n\quad\forall n \tag{2.22}$$

with constant $c>0$. For this purpose we first prove the following two lemmas.

LEMMA 2.4. *For any $x\in H_n$, $\tau>-\sigma/2$, as a function of $\omega\in R$,*

$$\|((\tau+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\|\in L^2(R),\qquad\|((\tau+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\|\to0\quad\text{as }|\omega|\to\infty.$$

Moreover,

$$\int_{-\infty}^{+\infty}\|((\tau+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\|^2\,d\omega\le\frac{\pi M_1^2\|x\|^2}{\tau+\frac{\sigma}{2}}, \tag{2.23}$$

$$\int_{-\infty}^{+\infty}\|((\tau-\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n^*)^{-1}x\|^2\,d\omega\le\frac{\pi M_1^2\|x\|^2}{\tau+\frac{\sigma}{2}}. \tag{2.24}$$

Proof. By Hille–Yosida's theorem we have

(2.25)

$$\begin{aligned}\|((\tau+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\|^2&=\left(\int_0^\infty e^{-((\tau+\boldsymbol{i}\omega)I)t}\tilde{T}_n(t)x\,dt,\int_0^\infty e^{-((\tau+\boldsymbol{i}\omega)I)s}\tilde{T}_n(s)x\,ds\right)\\&=\int_0^\infty\int_0^\infty e^{-\tau(t+s)}e^{-\boldsymbol{i}\omega(t-s)}(\tilde{T}_n(t)x,\tilde{T}_n(s)x)\,dt\,ds\\&=\int_0^\infty\int_{-s}^\infty e^{-\tau(u+2s)}e^{-\boldsymbol{i}\omega u}(\tilde{T}_n(u+s)x,\tilde{T}_n(s)x)\,du\,ds\\&=\int_0^\infty e^{-\boldsymbol{i}\omega u}\left(\int_0^\infty e^{-\tau(u+2s)}(\tilde{T}_n(u+s)x,\tilde{T}_n(s)x)\,ds\right)du\\&\quad+\int_{-\infty}^0 e^{-\boldsymbol{i}\omega u}\left(\int_{-u}^\infty e^{-\tau(u+2s)}(\tilde{T}_n(u+s)x,\tilde{T}_n(s)x)\,ds\right)du\\&\overset{\text{def}}{=}\int_{-\infty}^{+\infty}f(u)e^{-\boldsymbol{i}\omega u}\,du\end{aligned}$$

with

$$f(u) = \begin{cases} \displaystyle\int_0^\infty e^{-\tau(u+2s)}(\tilde{T}_n(u+s)x, \tilde{T}_n(s)x)\,ds, & u > 0, \\ \displaystyle\int_{-u}^\infty e^{-\tau(u+2s)}(\tilde{T}_n(u+s)x, \tilde{T}_n(s)x)\,ds, & u < 0. \end{cases} \tag{2.26}$$

Therefore, we have for $u > 0$

$$\begin{aligned} |f(u)| &\le \int_0^\infty e^{-\tau(u+2s)}\|\tilde{T}_n(u+s)x\| \cdot \|\tilde{T}_n(s)x\|\,ds \\ &\le M_1^2\|x\|^2 \int_0^\infty e^{-\tau(u+2s)}e^{-(\sigma/2)(u+s)}e^{-(\sigma/2)s}\,ds \\ &= \frac{M_1^2\|x\|^2}{2\left(\tau+\frac{\sigma}{2}\right)}\,e^{-(\tau+(\sigma/2))u}, \end{aligned} \tag{2.27}$$

and for $u < 0$,

$$\begin{aligned} |f(u)| &\le M_1^2\|x\|^2 \int_{-u}^\infty e^{-\tau(u+2s)}e^{-(\sigma/2)(u+s)}e^{-(\sigma/2)s}\,ds \\ &= \frac{M_1^2\|x\|^2}{2\left(\tau+\frac{\sigma}{2}\right)}\,e^{(\tau+(\sigma/2))u}. \end{aligned} \tag{2.28}$$

It turns out that $f \in L^1(R) \cap L^\infty(R)$,

$$\|f\|_{L^\infty} \le \frac{M_1^2\|x\|^2}{2\left(\tau+\frac{\sigma}{2}\right)}. \tag{2.29}$$

By [HS, p. 401b] we conclude that $\|((\tau+i\omega)I - \tilde{\mathcal{A}}_n)^{-1}x\|^2 \in L^1(R)$ and

$$\int_{-\infty}^{+\infty} \|((\tau+i\omega)I - \tilde{\mathcal{A}}_n)^{-1}x\|^2\,d\omega \le 2\pi\|f\|_{L^\infty}. \tag{2.30}$$

In addition, from the Riemann–Lebesgue theorem, $\|((\tau+i\omega)I - \tilde{\mathcal{A}}_n)^{-1}x\| \to 0$. Combining (2.30) with (2.29) yields (2.23). Inequality (2.24) can be proved in the same way. □

LEMMA 2.5. *For any $x \in H_n$, $\omega \in R$ we have*

$$\|(i\omega I - \tilde{\mathcal{A}}_n)^{-1}x\| \le 2^m\|((-\sigma+i\omega)I - \tilde{\mathcal{A}}_n)^{-1}x\| \tag{2.31}$$

with m being an integer such that

$$m - 1 < -2M_0\sigma \le m. \tag{2.32}$$

Proof. Let

$$\tau_m = -\sigma, \quad \Delta\tau = \frac{1}{2M_0}, \quad \tau_i = \tau_m - (m-i)\,\Delta\tau, \qquad (i = m-1, \ldots, 0). \tag{2.33}$$

Then $\tau_0 \le 0$. Since

(2.34)
$$\|[I-(\tau_i-\tau)((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}]x\| \geq \|x\|-(\tau_i-\tau)\|((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\| \geq \|x\|-\Delta\tau\cdot M_0\cdot\|x\| \geq \tfrac{1}{2}\|x\| \quad \text{for } \tau\in[\tau_{i-1},\tau_i],$$

we have

$$\|[I-(\tau_i-\tau)((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}]^{-1}\| \leq 2. \tag{2.35}$$

For any fixed i, $(i=m,\dots,1)$ and any $\tau\in[\tau_{i-1},\tau_i]$, we obtain

(2.36)
$$\begin{aligned}\|((\tau+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\| &= \|[I-(\tau_i-\tau)((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}]^{-1}((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\|\\ &\leq \|[I-(\tau_i-\tau)((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}]^{-1}\|\\ &\quad\cdot\|((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\|\\ &\leq 2\|((\tau_i+\boldsymbol{i}\omega)I-\tilde{\mathcal{A}}_n)^{-1}x\|,\end{aligned}$$

which results in (2.31). □

COROLLARY 2.6. *We have*

$$\int_{-\infty}^{+\infty}\|(\boldsymbol{i}\omega I-\tilde{\mathcal{A}}_n)^{-1}x\|^2\,d\omega \leq cM_1^2\|x\|^2 \tag{2.37}$$

and

$$\int_{-\infty}^{+\infty}\|(-\boldsymbol{i}\omega I-\tilde{\mathcal{A}}_n^*)^{-1}x\|^2\,d\omega \leq cM_1^2\|x\|^2 \tag{2.38}$$

with a positive constant c depending only on σ.

We are now in position to prove (2.22). Let $\tau_0>\tau_1>-(\sigma/2)>0$. Then for $x\in\mathcal{D}(\tilde{\mathcal{A}}_n^2)$, $y\in H_n$, by the inverse formula [Pa, Cor. 7.5, p. 29] we have

$$\begin{aligned}(\tilde{T}_n(t)x,y) &= \frac{1}{2\pi\boldsymbol{i}}\lim_{\omega\to+\infty}\int_{\tau_1-\boldsymbol{i}\omega}^{\tau_1+\boldsymbol{i}\omega}e^{\lambda t}((\lambda I-\tilde{\mathcal{A}}_n)^{-1}x,y)\,d\lambda\\ &= \frac{1}{2\pi\boldsymbol{i}}\lim_{\omega\to+\infty}\left[\frac{e^{\lambda t}}{t}((\lambda I-\tilde{\mathcal{A}}_n)^{-1}x,y)\Big|_{\tau_1-\boldsymbol{i}\omega}^{\tau_1+\boldsymbol{i}\omega}\right.\\ &\quad\left.+\int_{\tau_1-\boldsymbol{i}\omega}^{\tau_1+\boldsymbol{i}\omega}\frac{e^{\lambda t}}{t}((\lambda I-\tilde{\mathcal{A}}_n)^{-2}x,y)\,d\lambda\right].\end{aligned} \tag{2.39}$$

From Lemma 2.4 we have that

$$(\tilde{T}_n(t)x,y) = \lim_{\omega\to+\infty}\frac{1}{2\pi\boldsymbol{i}}\int_{\tau_1-\boldsymbol{i}\omega}^{\tau_1+\boldsymbol{i}\omega}\frac{e^{\lambda t}}{t}((\lambda I-\tilde{\mathcal{A}}_n)^{-2}x,y)\,d\lambda. \tag{2.40}$$

Since $\sup_{n\in N}\{\text{Re }\lambda;\lambda\in\sigma(\tilde{\mathcal{A}}_n)\}<(\sigma/2)<0$, for $t>0$, $e^{\lambda t}/t((\lambda I-\tilde{\mathcal{A}}_n)^{-2}x,y)$ is analytic in the domain $\{\lambda;\text{Re }\lambda\in(\sigma/2,\tau_0)\}$. Let Γ_ω be the curve composed of $\Gamma_0=\{\text{Re }\lambda=\tau_1,-\omega\leq\text{ Im }\lambda\leq\omega\}$, $\Gamma_{1,2}=\{0\leq\text{ Re }\lambda\leq\tau_1,\text{ Im }\lambda=\pm\omega\}$, and $\Gamma_3=\{\text{Re }\lambda=0,-\omega\leq\text{ Im }\lambda\leq\omega\}$. From

$$\int_{\Gamma_\omega}\frac{e^{\lambda t}}{t}((\lambda I-\tilde{\mathcal{A}}_n)^{-2}x,y)\,d\lambda=0 \tag{2.41}$$

and, due to Lemma 2.4,

$$\lim_{\omega\to+\infty}\int_{\Gamma_{1,2}}\frac{e^{\lambda t}}{t}((\lambda I-\tilde{\mathcal{A}}_n)^{-2}x,y)\,d\lambda=0, \tag{2.42}$$

it follows that

$$\begin{aligned}(\tilde{T}_n(t)x,y) &= \frac{1}{2\pi i}\lim_{\omega\to+\infty}\int_{-i\omega}^{i\omega}\frac{e^{\lambda t}}{t}((\lambda I-\tilde{\mathcal{A}}_n)^{-2}x,y)\,d\lambda\\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{i\omega t}}{t}((i\omega I-\tilde{\mathcal{A}}_n)^{-2}x,y)\,d\omega.\end{aligned} \tag{2.43}$$

Therefore,

$$\begin{aligned}|(\tilde{T}_n(t)x,y)| &\le \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{1}{t}\,\|(i\omega I-\tilde{\mathcal{A}}_n)^{-1}x\|\cdot\|(-i\omega I-\tilde{\mathcal{A}}_n^*)^{-1}y\|\,d\omega\\ &\le \frac{1}{2\pi t}\left(\int_{-\infty}^{+\infty}\|(i\omega I-\tilde{\mathcal{A}}_n)^{-1}x\|^2\,d\omega\right)^{1/2}\left(\int_{-\infty}^{+\infty}\|(-i\omega I-\tilde{\mathcal{A}}_n^*)^{-1}y\|^2\,d\omega\right)^{1/2}.\end{aligned} \tag{2.44}$$

Combining it with (2.37), (2.38) yields (2.22) for $x\in\mathcal{D}(\tilde{\mathcal{A}}_n^2)$, $y\in H_n$. Since $\mathcal{D}(\tilde{\mathcal{A}}_n^2)$ is dense in H_n, (2.22) also holds for any $x,y\in H_n$. By taking $y=\tilde{T}_n(t)x$, we conclude that for $t\ge 1$

$$\|\tilde{T}_n(t)\|\le\frac{cM_1^2}{2\pi}. \tag{2.45}$$

For $0\le t\le 1$, by (2.3) we have

$$\|\tilde{T}_n(t)\|=\|T_n(t)e^{-\frac{\sigma}{2}t}\|\le M_1e^{-\sigma/2}. \tag{2.46}$$

Therefore,

$$\|\tilde{T}_n(t)\|\le\max\left(\frac{cM_1^2}{2\pi},M_1e^{-\sigma/2}\right)\stackrel{\text{def}}{=}M\quad\forall t>0 \tag{2.47}$$

which results in

$$\|T_n(t)\|=\|\tilde{T}_n(t)e^{\frac{\sigma}{2}t}\|\le Me^{\frac{\sigma}{2}t},\qquad\sigma<0\quad\forall t>0. \tag{2.48}$$

Thus the proof of Theorem 2.1 is complete.

3. Approximations of the thermoelastic system. In this section we first present a general approximate scheme for thermoelastic system (1.4). Then we will use Corollary 2.3 to show the uniform exponential stability of a particular approximate scheme that is often referred to as the modal method. We also provide a convergence proof of this scheme.

Let

$$E_j=\begin{pmatrix}\phi_j\\0\\0\end{pmatrix},\quad E_{n+j}=\begin{pmatrix}0\\\psi_j\\0\end{pmatrix},\quad E_{2n+j}=\begin{pmatrix}0\\0\\\xi_j\end{pmatrix},\qquad j=1,\dots,n \tag{3.1}$$

be a basis for the finite-dimensional space $H_n=H_1^n(\Omega)\times H_2^n(\Omega)\times H_3^n(\Omega)\subset H_0^1(\Omega)\times H_0^1(\Omega)\times H_0^1(\Omega)\subset H=H_0^1(\Omega)\times L^2(\Omega)\times L^2(\Omega)$. The inner product on H_n is the one

induced by the H-product. For simplicity, we take $c^2 = 1$ in H-norm (1.9) without affecting the proof of our results. We consider the approximation to the solution of (1.4) of the form

$$z_n = \sum_{j=1}^{3n} \tilde{z}_j(t) E_j(x), \tag{3.2}$$

which is required to satisfy the following variational system:

$$(\dot{z}_n, E_j)_H = (\mathcal{A} z_n, E_j)_H, \qquad j = 1, \dots, 3n. \tag{3.3}$$

Then we have

$$\begin{aligned} M_n \dot{\tilde{z}}_n &= \begin{bmatrix} M_n^{(1)} & & \\ & M_n^{(2)} & \\ & & M_n^{(3)} \end{bmatrix} \begin{bmatrix} \dot{\tilde{z}}_n^{(1)} \\ \dot{\tilde{z}}_n^{(2)} \\ \dot{\tilde{z}}_n^{(3)} \end{bmatrix} \\ &= \begin{bmatrix} 0 & \tilde{D}_n^T & 0 \\ -\tilde{D}_n & 0 & -\gamma \tilde{F}_n \\ 0 & \gamma \tilde{F}_n^T & -G_n \end{bmatrix} \begin{bmatrix} \tilde{z}_n^{(1)} \\ \tilde{z}_n^{(2)} \\ \tilde{z}_n^{(3)} \end{bmatrix} = \tilde{A}_n \tilde{z}_n \end{aligned} \tag{3.4}$$

with

$$\begin{aligned} &\left(M_n^{(1)}\right)_{ij} = (D\phi_i, D\phi_j)_{L^2}, \quad \left(M_n^{(2)}\right)_{ij} = (\psi_i, \psi_j)_{L^2}, \quad \left(M_n^{(3)}\right)_{ij} = (\xi_i, \xi_j)_{L^2}, \\ &(\tilde{D}_n)_{ij} = (D\phi_i, D\psi_j)_{L^2}, \quad (\tilde{F}_n)_{ij} = (D\xi_i, \psi_j)_{L^2}, \quad (G_n)_{ij} = (D\xi_i, D\xi_j)_{L^2} \end{aligned} \tag{3.5}$$

and

$$\tilde{z}_n^{(i)} = (\tilde{z}_{(i-1)n+1}, \dots, \tilde{z}_{in})^T, \qquad i = 1, 2, 3. \tag{3.6}$$

By construction, the matrix $M_n^{(i)}$ is symmetric and positive definite. Therefore, there exists a lower triangle matrix $L_n^{(i)}$ such that $M_n^{(i)} = (L_n^{(i)})^T (L_n^{(i)})$. Let $L_n = \operatorname{diag}(L_n^{(1)}, L_n^{(2)}, L_n^{(3)})$ and denote $L_n \tilde{z}_n$ by $\bar{z}_n$. Then to obtain approximate solution z_n we are led to solving ordinary differential equations

$$\dot{\bar{z}}_n = A_n \bar{z}_n \tag{3.7}$$

with

$$A_n = \begin{bmatrix} 0 & (L_1^T)^{-1} \tilde{D}_n^T L_2^{-1} & 0 \\ -(L_2^T)^{-1} \tilde{D}_n L_1^{-1} & 0 & -\gamma (L_2^T)^{-1} \tilde{F}_n L_3^{-1} \\ 0 & \gamma (L_3^T)^{-1} \tilde{F}_n^T L_2^{-1} & -(L_3^T)^{-1} G_n L_3^{-1} \end{bmatrix}. \tag{3.8}$$

It is easy to see that

$$(A_n \bar{z}_n, \bar{z}_n)_{C^{3n}} = -\left(G_n L_3^{-1} \bar{z}_n^{(3)}, L_3^{-1} \bar{z}_n^{(3)}\right)_{C^n} \le 0 \tag{3.9}$$

provided that G_n is semipositive definite. In that case, A_n generates a C_0-semigroup $T_n(t)$ of contraction on H.

The modal approximation scheme is to choose the eigenvectors of the system as the basis vectors. Here, we will use the eigenvectors of the uncoupled thermoelastic system, i.e., $\gamma = 0$ in (1.4). Thus we still call it modal approximation. Let

$$\phi_j = \sqrt{\frac{2}{\pi}} \frac{1}{j} \sin jx, \quad \psi_j = \sqrt{\frac{2}{\pi}} \sin jx, \quad \xi_j = \sqrt{\frac{2}{\pi}} \sin jx, \qquad j = 1, \dots, n. \tag{3.10}$$

A straightforward calculation following (3.5) and (3.8) yields

$$A_n = \begin{bmatrix} 0 & D_n & 0 \\ -D_n & 0 & -\gamma F_n \\ 0 & \gamma F_n^T, & -D_n^2 \end{bmatrix} \tag{3.11}$$

with

$$D_n = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & n \end{bmatrix}, \qquad F_{ij} = \begin{cases} -\dfrac{4}{\pi}\dfrac{ij}{i^2 - j^2}, & |i-j| = \text{odd}, \\ 0, & \text{otherwise}. \end{cases} \tag{3.12}$$

THEOREM 3.1. *The semigroups generated by A_n defined in* (3.11) *are uniformly exponentially stable, i.e., there exist positive constants M and α, independent of n, such that*

$$\|T_n(t)\|_{\mathcal{L}(H_n, H_n)} \leq Me^{-\alpha t}. \tag{3.13}$$

Proof. By Corollary 2.3 we need only to prove that (2.1) and (2.4) hold. This will be done by contradiction. We first point out that the real parts of the eigenvalues of the matrix A_n are strictly negative for every n.

(i) If (2.4) is not true, then there must exist a sequence of $\lambda_n \in C$ with $\mathrm{Re}\lambda_n \geq 0$, $\mathrm{Re}\lambda_n \to 0$ (as $n \to +\infty$), a sequence of $h_n \in C^{3n}$ with $\|h_n\|_{C^{3n}} = 1$, and a subsequence of A_n, still denoting by A_n such that

$$\|(\lambda_n I - A_n)h_n\|_{C^{3n}} \to 0. \tag{3.14}$$

The matrix

$$A_0^n = \begin{bmatrix} 0 & D_n \\ -D_n & 0 \end{bmatrix} \tag{3.15}$$

has eigenvalues

$$\pm \boldsymbol{i}j, \qquad j = 1, \ldots, n, \tag{3.16}$$

and corresponding eigenvectors

$$\mathcal{E}_j = \frac{1}{\sqrt{2}} \begin{pmatrix} e_j \\ \boldsymbol{i}e_j \end{pmatrix}, \quad \mathcal{E}_{-j} = \frac{1}{\sqrt{2}} \begin{pmatrix} e_j \\ -\boldsymbol{i}e_j \end{pmatrix}, \qquad j = 1, \ldots, n, \tag{3.17}$$

with e_j being the jth unit vector.

It follows from

$$\mathrm{Re}((\lambda_n I - A_n)h_n, h_n)_{C^{3n}} \to 0 \tag{3.18}$$

that

$$\mathrm{Re}\lambda_n + \|D_n h_n^{(3)}\|^2 \to 0. \tag{3.19}$$

Hereafter we denote by $\|\cdot\|$ the l^2 norm in C^n. Therefore, we have

$$\|h_n^{(3)}\| \leq \|D_n h_n^{(3)}\| \to 0. \tag{3.20}$$

Taking the first $2n$ rows of (3.14) into consideration gives

$$\left\| (\lambda_n I - A_0^n) \begin{pmatrix} h_n^{(1)} \\ h_n^{(2)} \end{pmatrix} + \begin{pmatrix} 0 \\ \gamma F_n h_n^{(3)} \end{pmatrix} \right\|_{C^{2n}} \to 0. \tag{3.21}$$

We now claim that

$$\|F_n h_n^{(3)}\| \to 0. \tag{3.22}$$

In fact,

$$\begin{aligned} \left(F_n h_n^{(3)}\right)_i &= \frac{2}{\pi} \left(D \sin ix, \sum_{j=1}^{n} (h_n^{(3)})_j \sin jx \right)_{L^2} \\ &= -\frac{2}{\pi} \left(\sin ix, \sum_{j=1}^{n} j \cos jx (h_n^{(3)})_j \right)_{L^2}. \end{aligned} \tag{3.23}$$

By Parsaval's inequality,

$$\|F_n h_n^{(3)}\|^2 \le \frac{2}{\pi} \left\| \sum_{j=1}^{n} j \cos jx (h_n^{(3)})_j \right\|_{L^2}^2 \le \sum_{j=1}^{n} j^2 |(h_n^{(3)})_j|^2 = \|D_n h_n^{(3)}\|^2. \tag{3.24}$$

Thus, (3.22) follows from (3.20). Moreover, we deduce from (3.21) that

$$\left\| (\lambda_n I - A_0^n) \begin{pmatrix} h_n^{(1)} \\ h_n^{(2)} \end{pmatrix} \right\|_{C^{2n}} \to 0. \tag{3.25}$$

Since the eigenvectors $\{\mathcal{E}_{\pm j}\}$ form a basis in C^{2n}, we have

$$\begin{pmatrix} h_n^{(1)} \\ h_n^{(2)} \end{pmatrix} = \sum_{j=-n, j\neq 0}^{n} \alpha_{nj} \mathcal{E}_j. \tag{3.26}$$

It follows from $\|h_n\|_{C^{3n}} = 1$ and $\|h_n^{(3)}\| \to 0$ that

$$\left\| \begin{pmatrix} h_n^{(1)} \\ h_n^{(2)} \end{pmatrix} \right\|_{C^{2n}}^2 = \sum_{j=-n, j\neq 0}^{n} |\alpha_{nj}|^2 \to 1. \tag{3.27}$$

Substituting (3.26) into (3.25), we obtain

$$\begin{aligned} \left\| (\lambda_n I - A_0^n) \begin{pmatrix} h_n^{(1)} \\ h_n^{(2)} \end{pmatrix} \right\|_{C^{2n}}^2 &= \left\| \sum_{j=-n, j\neq 0}^{n} (\lambda_n - \boldsymbol{i} j) \alpha_{nj} \mathcal{E}_j \right\|_{C^{2n}}^2 \\ &= \sum_{j=-n, j\neq 0}^{n} |\lambda_n - \boldsymbol{i} j|^2 |\alpha_{nj}|^2 \to 0, \quad \text{as } n \to +\infty. \end{aligned} \tag{3.28}$$

If for n large enough, $|\lambda_n - \boldsymbol{i} j| \ge \delta > 0$ for all j, then $\sum_{j=-n, j\neq 0}^{n} |\alpha_{n,j}|^2 \to 0$, a contradiction with (3.27). Thus we derive from (3.27) and (3.28) that there exists $j(n) \in \{\pm 1, \pm 2, \ldots, \pm n\}$

such that as $n \to +\infty$

$$\begin{aligned} \lambda_n - \boldsymbol{i}j(n) &\to 0, \\ \sum_{j=-n, j\neq 0, j(n)}^{n} |\alpha_{n,j}|^2 &\to 0, \\ |\alpha_{n,j(n)}| &\to 1, \end{aligned} \tag{3.29}$$

and

$$\left\| \begin{pmatrix} h_n^{(1)} \\ h_n^{(2)} \end{pmatrix} - \alpha_{n,j(n)} \mathcal{E}_{j(n)} \right\|_{C^{2n}} \to 0. \tag{3.30}$$

Taking the last n rows of (3.14) into consideration, we obtain

$$\|\lambda_n h_n^{(3)} + D_n^2 h_n^{(3)} - \gamma F_n^T h_n^{(2)}\| \to 0. \tag{3.31}$$

By (3.29), (3.31) and $j(n) \neq 0$, if we denote $h_n^{(3)}/j(n)$ by y_n, then

$$\|g_n\| \overset{\text{def}}{=} \left\| \boldsymbol{i}j(n)y_n + D_n^2 y_n - \boldsymbol{i}\frac{\gamma\alpha_{n,j(n)}}{\sqrt{2}|j(n)|} F_n^T e_{|j(n)|} \right\| \to 0. \tag{3.32}$$

Taking real part of the inner product of g_n with $D_n^2 y_n$ yields

$$\operatorname{Re}(g_n, D_n^2 y_n) = \|D_n^2 y_n\|^2 - \operatorname{Re}\left(\boldsymbol{i}\frac{\gamma\alpha_{n,j(n)}}{\sqrt{2}|j(n)|} F_n^T e_{|j(n)|}, D_n^2 y_n \right). \tag{3.33}$$

We now estimate the last term on the right hand side of (3.33). Indeed,

(3.34)

$$\begin{aligned} \operatorname{Re}\left(\frac{1}{|j(n)|} F_n^T e_{|j(n)|}, D_n^2 y_n \right) &= \frac{2}{\pi j(n)} \operatorname{Re} \sum_{i=1}^{n} \int_0^\pi -(D \sin ix) \sin j(n)x dx \cdot \overline{i^2 (y_n)_i} \\ &= \frac{2}{\pi} \operatorname{Re} \sum_{i=1}^{n} i \int_0^\pi \sin ix \cos j(n)x dx \cdot \overline{i(y_n)_i} \\ &= -\frac{2}{\pi} \operatorname{Re} \sum_{i=1}^{n} \int_0^\pi \cos ix \sin j(n)x dx \cdot \overline{i(h_n^{(3)})_i} \\ &\quad + \frac{2}{\pi} \operatorname{Re} \sum_{i=1}^{n} (\cos ix \cos j(n)x) \Big|_0^\pi \cdot \frac{1}{j(n)} \overline{i(h_n^{(3)})_i}. \end{aligned}$$

Therefore,

$$\begin{aligned} I &\equiv \left| \operatorname{Re}\left(\boldsymbol{i}\frac{\gamma\alpha_{n,j(n)}}{\sqrt{2}|j(n)|} F_n^T e_{|j(n)|}, D_n^2 y_n \right) \right| \\ &\leq \frac{\gamma|\alpha_{n,j(n)}|}{\sqrt{\pi}} (\|\sin j(n)x\|_{L^2} \|D_n h_n^{(3)}\| + 2\|Dz_n\|_{L^\infty}) \\ &= \frac{\gamma|\alpha_{n,j(n)}|}{\sqrt{\pi}} \left(\frac{\pi}{2} \|D_n h_n^{(3)}\| + 2\|Dz_n\|_{L^\infty} \right), \end{aligned} \tag{3.35}$$

where

$$z_n = \sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} (y_n)_i \sin ix. \tag{3.36}$$

By the well-known Nirenberg inequality, we have

$$\begin{aligned} \|Dz_n\|_{L^\infty} &\le c_1 \|D^2 z_n\|_{L^2}^{1/2} \|Dz_n\|_{L^2}^{1/2} + c_2 \|Dz_n\|_{L^2} \\ &= c_1 \|D_n^2 z_n\|^{1/2} \|D_n z_n\|^{1/2} + c_2 \|D_n z_n\| \end{aligned} \tag{3.37}$$

with c_1, c_2 being positive constants independent of z_n.

Combining (3.35) with (3.37) and applying Young's inequality yields

$$I \le \frac{\gamma |\alpha_{n,j(n)}|}{\sqrt{\pi}} \left(\frac{\pi}{2} \|D_n h_n^{(3)}\| + 2c_2 \|D_n y_n\| \right) + \frac{1}{4} \|D_n^2 y_n\| + c_3 \|D_n y_n\|^{2/3}. \tag{3.38}$$

On the other hand,

$$|\mathrm{Re}(g_n, D_n^2 y_n)| \le \frac{1}{2} \|g_n\|^2 + \frac{1}{2} \|D_n^2 y_n\|^2. \tag{3.39}$$

Thus combining (3.38), (3.39) with (3.33), we obtain

$$\|D_n^2 y_n\| \to 0. \tag{3.40}$$

Let w_n be the unique solution to the equation

$$\boldsymbol{i} j(n) w_n + D_n^2 w_n - \boldsymbol{i} \frac{\gamma \alpha_{n,j(n)}}{\sqrt{2}|j(n)|} F_n^T e_{|j(n)|} = 0. \tag{3.41}$$

Then

$$(w_n)_i = \begin{cases} \dfrac{-\mathrm{sign}(j(n)) \boldsymbol{i} 2\sqrt{2} \gamma \alpha_{n,j(n)} i}{\pi(\boldsymbol{i} j(n) + i^2)(i^2 - j^2(n))}, & |i - j(n)| = \text{odd}, \\ 0, & \text{otherwise}. \end{cases} \tag{3.42}$$

From (3.32) and (3.41), we obtain

$$\|\boldsymbol{i} j(n)(y_n - w_n) + D_n^2 (y_n - w_n)\| \to 0, \tag{3.43}$$

which results in, in a similar way as before,

$$\|D_n^2 (y_n - w_n)\| \to 0. \tag{3.44}$$

It follows immediately from (3.40), (3.44) that

$$\|D_n^2 w_n\| \to 0. \tag{3.45}$$

On the other hand, since $|\alpha_{n,j(n)}| \to 1$, as $n \to \infty$, we obtain for n large enough

$$\begin{aligned} \|D_n^2 w_n\|^2 &= \sum_{i=1, |i-j(n)|=\text{odd}}^{n} \frac{8\gamma^2 |\alpha_{n,j(n)}|^2 i^6}{\pi^2 (j^2(n) + i^4)(i^2 - j^2(n))^2} \\ &\ge \begin{cases} \dfrac{2\gamma^2 |\alpha_{n,j(n)}|^2 (|j(n)| + 1)^6}{\pi^2 ((|j(n)| + 1)^4 + j^2(n))(2|j(n)| + 1)^2}, & |j(n)| < n \\ \dfrac{2\gamma^2 |\alpha_{n,j(n)}|^2 (n-1)^6}{\pi^2 ((n-1)^4 + n^2)(2n-1)^2}, & |j(n)| = n \end{cases} \\ &\ge \delta > 0 \end{aligned} \tag{3.46}$$

with δ being a constant independent of $n, j(n)$. Thus, we have a contradiction.

(ii) If (2.1) is not true, then there must exist a sequence of $\lambda_n \in C$ with $\lambda_n \in \sigma(A_n)$, $\operatorname{Re} \lambda_n \to 0$, a sequence of $h_n \in \mathcal{D}(A_n)$ with $\|h_n\|_{C^{3n}} = 1$ and a subsequence of A_n, still denoting by A_n, such that

$$(\lambda_n I - A_n)h_n = 0. \tag{3.47}$$

Taking real part of the inner product of (3.47) with h_n, we obtain

$$\operatorname{Re} \lambda_n + \|D_n h_n^{(3)}\|^2 = 0, \tag{3.48}$$

which also results in (3.20). Taking the first $2n$ rows of (3.47) into consideration, we again obtain (3.25). Repeating the same argument as before leads to a contradiction again. Thus the proof of Theorem 3.1 is complete. □

Remark 3.1. For the cases of Dirichlet–Neumann and Neumann–Dirichlet boundary conditions, we can show that the approximate semigroups by the modal method are also uniformly exponentially stable. For example, if u satisfies the Dirichlet boundary condition and θ satisfies the Neumann boundary condition, then $\int_0^\pi \theta(x,t)\,dx = \int_0^\pi \theta_0(x)\,dx$, which is derived by integrating (1.2) with respect to x and t, where $\theta_0(t)$ is the initial temperature distribution of the rod.

After changing to the new dependent variable

$$\bar{\theta} = \theta - \frac{1}{\pi}\int_0^\pi \theta_0(x)\,dx, \tag{3.49}$$

we can choose the state space $H = \{(y_1, y_2, y_3) \in H_0 \times L^2 \times L^2 | \int_0^\pi y_3\,dx = 0\}$, choose ϕ_j, ψ_j as before and ξ_j to be $\sqrt{2/\pi}\cos jx$. In this case, the matrix D_n is the same as (3.12). Moreover, the matrix $F_n = -D_n$. Therefore, the proof of $\|D_n^2 y_n\| \to 0$ and $\|D_n^2 w_n\| \to 0$ can be even more easily carried out. Accordingly, we have

$$(w_n)_i = \frac{-\operatorname{sign}(j(n))\gamma\alpha_{n,j(n)}}{\sqrt{2}(i^2 + ij(n))} \tag{3.50}$$

and

$$\begin{aligned} \|D_n^2 w_n\|^2 &= \sum_{i=1}^{n} \frac{\gamma^2 |\alpha_{n,j(n)}|^2 i^4}{2(i^4 + j^2(n))} \\ &\geq \frac{\gamma^2 |\alpha_{n,j(n)}|^2 j^4(n)}{2(j^4(n) + j^2(n))} \geq \delta > 0. \end{aligned} \tag{3.51}$$

The strong convergence of the approximating semigroups $T_n(t)$ to $T(t)$ and $T_n^*(t)$ to $T^*(t)$ is another important issue for the approximation of LQG problems (see [GRT, Hypoths. 4.3, 4.4]). Let P_n be the orthogonal projection from H to H_n. Then the matrix A_n in (3.11) is the matrix representation of the operator $\mathcal{A}_n = P_n \mathcal{A} P_n$. Let

$$\mathcal{D} = \mathcal{D}(\mathcal{A}) \cap (H^4 \times H^3 \times H^4). \tag{3.52}$$

It is easy to see that $\mathcal{D}$ is dense in H. Since $(I - \mathcal{A})\mathcal{D}(\mathcal{A}) = H$, we also know that $(I - \mathcal{A})\mathcal{D}$ is dense in H. With the dissipativeness of $\mathcal{A}$ and $\mathcal{A}_n$, by the Trotter–Kato theorem (see [P, Thm. 4.5]), we only need to show $\mathcal{A}_n z \to \mathcal{A} z$ in H for all $z \in \mathcal{D}$ for the strong convergence of the approximation semigroups $T_n(t)$ to $T(t)$.

THEOREM 3.2. *$T_n(t), T_n^*(t) \xrightarrow{s} T(t), T^*(t)$ in H, respectively. Moreover, the convergence is uniform in bounded t-intervals.*

Proof. Let $z \in \mathcal{D}$. Then

$$z = \sum_{j=1}^{\infty} \left[a_j \begin{pmatrix} \frac{1}{j} \sin jx \\ 0 \\ 0 \end{pmatrix} + b_j \begin{pmatrix} 0 \\ \sin jx \\ 0 \end{pmatrix} + c_j \begin{pmatrix} 0 \\ 0 \\ \sin jx \end{pmatrix} \right] \tag{3.53}$$

with $\{a_j j^3, b_j j^3, c_j j^4\}_1^{\infty}$ being l^2 sequences. Furthermore, we have

$$\mathcal{A}z = \begin{pmatrix} \sum_{i=1}^{\infty} b_i \sin ix \\ \sum_{i=1}^{\infty} \left(-a_i i - \gamma \sum_{j=1}^{\infty} c_j j f_{ij} \right) \sin ix \\ \sum_{i=1}^{\infty} \left(-\gamma \sum_{j=1}^{\infty} b_j j f_{ij} - c_i i^2 \right) \sin ix \end{pmatrix}, \tag{3.54}$$

and

$$\mathcal{A}_n z = \begin{pmatrix} \sum_{i=1}^{n} b_i \sin ix \\ \sum_{i=1}^{n} \left(-a_i i - \gamma \sum_{j=1}^{n} c_j j f_{ij} \right) \sin ix \\ \sum_{i=1}^{n} \left(-\gamma \sum_{j=1}^{n} b_j j f_{ij} - c_i i^2 \right) \sin ix \end{pmatrix}, \tag{3.55}$$

where $f_{ij} = 2/\pi \langle \cos jx, \sin ix \rangle_{L^2}$. Now $\mathcal{A}z - \mathcal{A}_n z$ can be written as

$$\begin{pmatrix} \sum_{i=n+1}^{\infty} b_i \sin ix \\ \sum_{i=n+1}^{\infty} \left(-a_i i - \gamma \sum_{j=1}^{\infty} c_j j f_{ij} \right) \sin ix \\ \sum_{i=n+1}^{\infty} \left(-\gamma \sum_{j=1}^{\infty} b_j j f_{ij} - c_i i^2 \right) \sin ix \end{pmatrix} - \begin{pmatrix} 0 \\ \gamma \sum_{i=1}^{n} \left(\sum_{j=n+1}^{\infty} c_j j f_{ij} \right) \sin ix \\ \gamma \sum_{i=1}^{n} \left(\sum_{j=n+1}^{\infty} b_j j f_{ij} \right) \sin ix \end{pmatrix} = I + II. \tag{3.56}$$

It follows from $\mathcal{A}z \in H$ that $\|I\|_H \to 0$ as $n \to 0$. The second entry of II can be estimated as follows:

(3.57)

$$\begin{aligned}\left\| \sum_{i=1}^{n} \left(\sum_{j=n+1}^{\infty} c_j j f_{ij} \right) \sin ix \right\|^2 &= \sum_{i=1}^{n} \left(\sum_{j=n+1}^{\infty} c_j j f_{ij} \right)^2 \\ &\le \sum_{i=1}^{n} \left(\sum_{j=n+1}^{\infty} |c_j|^{2\alpha} j^2 \sum_{j=n+1}^{\infty} |c_j|^{2-2\alpha} |f_{ij}|^2 \right) \\ &\le \left(\sum_{j=n+1}^{\infty} |c_j|^{2\alpha} j^2 \right) \left(\sum_{j=n+1}^{\infty} |c_j|^{2-2\alpha} \sum_{i=1}^{\infty} |f_{ij}|^2 \right) \\ &= \frac{\pi}{2} \left(\sum_{j=n+1}^{\infty} |c_j|^{2\alpha} j^2 \right) \left(\sum_{j=n+1}^{\infty} |c_j|^{2-2\alpha} \right).\end{aligned}$$

Since $\sum_{j=n+1}^{\infty} |c_j|^2 j^6$ is a convergent series, then $\sum_{j=1}^{\infty} |c_j|^{2\alpha} j^2$ and $\sum_{j=n+1}^{\infty} |c_j|^{2-2\alpha}$ are also convergent as long as $\frac{5}{6} > \alpha > \frac{1}{2}$. Therefore, by (3.57), we obtain

$$\left\| \sum_{i=1}^{n} \left(\sum_{j=n+1}^{\infty} c_j j f_{ij} \right) \sin ix \right\|^2 \to 0. \tag{3.58}$$

Similarly, we can obtain

$$\left\| \sum_{i=1}^{n} \left(\sum_{j=n+1}^{\infty} b_j j f_{ij} \right) \sin ix \right\| \to 0. \tag{3.59}$$

Thus, we have proved

$$\lim_{n\to\infty} \|\mathcal{A}z - \mathcal{A}_n z\|_H = 0, \qquad \forall\, z \in \mathcal{D}. \tag{3.60}$$

The convergence of approximate adjoint semigroups can be verified in a similar way since $\mathcal{A}$ and $\mathcal{A}^*$ only differ by the sign in front of the coupling coefficient γ (see [H]). □

Remark 3.2. For the cases of Dirichlet–Neumann and Neumann–Dirichlet boundary conditions, the convergence of $T_n(t)$ and $T_n^*(t)$ is obvious from the above analysis since there is no need to expand cos jx in terms of sin ix in (3.54) and (3.55).

4. Numerical studies. As has been demonstrated in §§2 and 3, preserving uniform exponential stability for the general approximation scheme of thermoelastic system (1.4) can be a complicated problem, due to the structure of the matrix A_n defined in (3.8). On the other hand, for any given approximation scheme, we can compute the eigenvalues of A_n to observe the trends in their location. If the eigenvalues are approaching the imaginary axis, then the uniform exponential stability is unlikely to be preserved. In other words, condition (2.1) will be violated.

In this section, we present three approximation schemes to system (1.4). For each of them, the matrix A_n is constructed and its eigenvalues are computed. In all the following examples, we take $\gamma = 0.1$. Since the real eigenvalues of the matrix A_n are much smaller than the real part of the complex eigenvalues, it is enough to observe the complex one only.

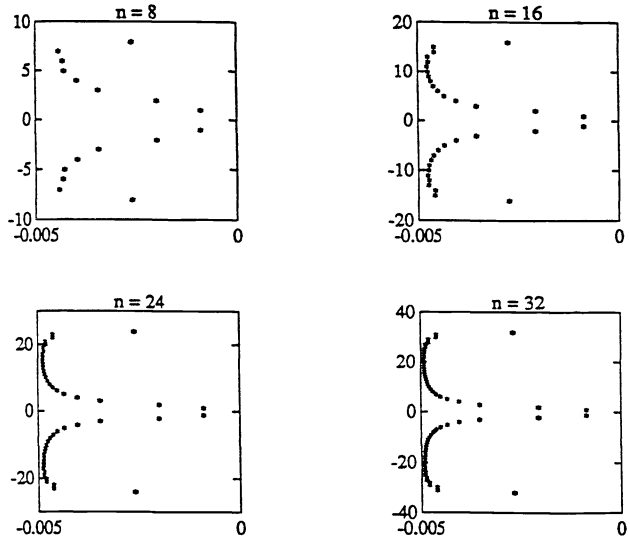


FIG. 4.1. *Location of the complex eigenvalues of the matrix A_n for the modal method in the case of Dirichlet–Dirichlet boundary conditions.*

TABLE 4.1

Distance between $\sigma(A_n)$ and the imaginary axis for the modal method in the case of Dirichlet–Dirichlet boundary conditions.

n	$\min\{-\mathrm{Re}\,\lambda,\ \lambda \in \sigma(A_n)\}$
8	8.9227×10^{-4}
16	8.9383×10^{-4}
24	8.9402×10^{-4}
32	8.9407×10^{-4}

4.1. Modal method. The modal method presented in §3 is implemented. The location of the eigenvalues is shown in Fig. 4.1. We can see that a uniform distance between the eigenvalues and the imaginary axis is preserved, which is consistent with (2.1). In Table 4.1, these distances for $n = 8, 16, 24, 32$ are reported. Comparing the results for the wave equation with boundary damping [BIW], we know that the damping due to heat conduction is much weaker. Another observation is that for fixed n, the eigenvalues of higher frequency modes, in particular, the one of the nth mode, are closer to the imaginary axis. However, as the number of modes increases, these eigenvalues bend back towards the vertical line $\lambda = -\gamma^2/2$. Hansen [H] has proved that, in the cases of Dirichlet–Neumann and Neumann–Dirichlet boundary conditions, the real part of the eigenvalues tend to $-\gamma^2/2$ asymptotically. The corresponding modal approximation scheme mentioned in Remark 3.1 preserves this property as shown in Fig. 4.2. It seems from Fig. 4.1 that in the case of Dirichlet–Dirichlet boundary conditions, the eigenvalues have the same asymptotic behavior. We list them in Table 4.2 for certain modes with $n = 24, 32, 64$, which clearly suggests a fast convergence.

4.2. Finite element method. The classical finite element method is to divide the domain $\Omega = [0, \pi]$ into subintervals, usually in equal length, and use spline functions for the approximation. Here, we choose ϕ_j, ψ_j, and ξ_j to be the normalization of the linear spline functions

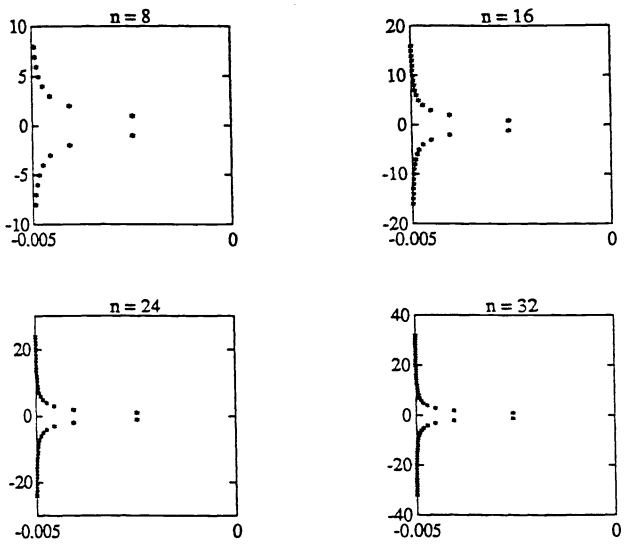


FIG. 4.2. *Location of the complex eigenvalues of the matrix A_n for the modal method in the case of Dirichlet–Neumann boundary conditions.*

TABLE 4.2
Eigenvalues of A_n for the modal method in the case of Dirichlet–Dirichlet boundary conditions.

mode	n = 24	n = 32	n = 64
1	$-8.9402 \times 10^{-4} + i1.0002$	$-8.9407 \times 10^{-4} + i1.0002$	$-8.9410 \times 10^{-4} + i1.0002$
2	$-2.0563 \times 10^{-3} + i2.0017$	$-2.0565 \times 10^{-3} + i2.0017$	$-2.0567 \times 10^{-3} + i2.0017$
11	$-4.8033 \times 10^{-3} + i11.001$	$-4.8116 \times 10^{-3} + i11.001$	$-4.8162 \times 10^{-3} + i11.001$
12	$-4.8224 \times 10^{-3} + i12.001$	$-4.8350 \times 10^{-3} + i12.001$	$-4.8413 \times 10^{-3} + i12.001$
22	$-4.5899 \times 10^{-3} + i22.000$	$-4.9003 \times 10^{-3} + i22.000$	$-4.9396 \times 10^{-3} + i22.000$
23	$-4.5960 \times 10^{-3} + i23.000$	$-4.9033 \times 10^{-3} + i23.000$	$-4.9435 \times 10^{-3} + i23.000$
24	$-2.6492 \times 10^{-3} + i24.000$	$-4.8881 \times 10^{-3} + i24.000$	$-4.9466 \times 10^{-3} + i24.000$

$$(4.1)\quad h_j(x) = \begin{cases} 1 - \dfrac{1}{\Delta}|x - j\Delta|, & x \in [(j-1)\Delta, (j+1)\Delta], \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, \ldots, n-1,$$

with $\Delta = \frac{\pi}{n}$.

It was pointed out in [BLM] that the eigenvalues of the matrix A_n derived from this approximation scheme approach the imaginary axis as n increases. Thus the uniform exponential stability is unlikely to be preserved. The location of the eigenvalues are shown in Fig. 4.3. The difference between Figs. 4.1 and 4.3 is due to the slow convergence of the linear spline approximation. For large n, the eigenvalues of the lower frequency modes do move toward the corresponding locations in Fig. 4.1. We also tested for the cubic spline approximation. For $n = 32$, the eigenvalues of the lower frequency modes are virtually the same as in Fig. 4.1, but the higher frequency one again approaches the imaginary axis. We include this method here to compare with the mixed finite element method, which will be presented in what follows.

4.3. Mixed finite element method. In the most common implementation of the finite element method for thermoelastic system (1.4), the approximation spaces H_1^n and H_2^n are

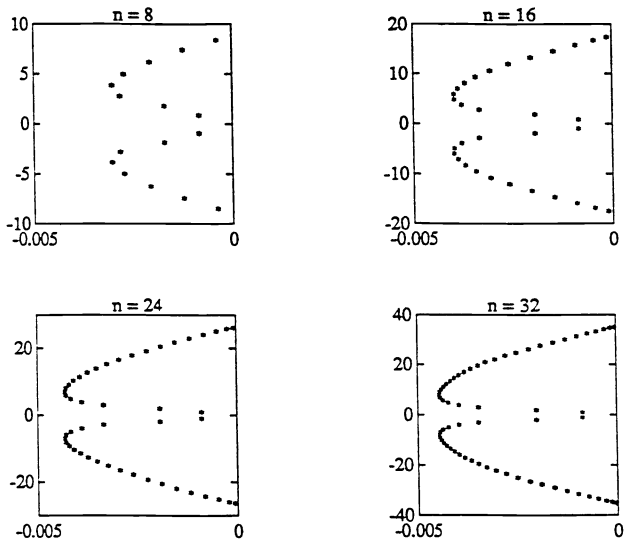


FIG. 4.3. *Location of the complex eigenvalues of the matrix A_n for the finite element method in the case of Dirichlet–Dirichlet boundary conditions.*

often chosen to be identical. However, this neglects the fact that u and v have different smoothness in spacial variable x. A general approximation scheme, called the mixed finite element method, was proposed by Ito and Kappel [IK] using different approximation spaces H_1^n and H_2^n. It was applied to the weakly damped wave equation [BIW], and was proved to be able to preserve the uniform exponential stability for the one-dimensional case. To apply this method here, we choose ψ_j to be the normalization of the piecewise constant function

$$f_j(x) = \begin{cases} 1, & x \in [(j-1)\Delta, (j+1)\Delta], \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, \dots, n-1 \tag{4.2}$$

with $\Delta = \pi/n$, and choose ϕ_j and ξ_j the same as in the above finite element method.

From (3.2), the three components of the approximate solution z_n are

$$z_n^{(1)} = \sum_{j=1}^{n-1} \tilde{z}_j \phi_j, \quad z_n^{(2)} = \sum_{j=1}^{n-1} \tilde{z}_{n-1+j} \psi_j, \quad z_n^{(3)} = \sum_{j=1}^{n-1} \tilde{z}_{2n-2+j} \xi_j. \tag{4.3}$$

They are required to satisfy the following variational system, which is analogous to (3.4):

$$(\dot{z}_n^{(1)}, \phi_j)_{H_0^1} = (q_n(z_n^{(2)}), \phi_j)_{H_0^1}, \tag{4.4}$$

$$(\dot{z}_n^{(2)}, \psi_j)_{L^2} = -(z_n^{(1)}, q_n(\psi_j))_{H_0^1} - \gamma(D_x z_n^{(3)}, p_n(\psi_j))_{L^2}, \tag{4.5}$$

$$(\dot{z}_n^{(3)}, \xi_j)_{L^2} = -\gamma(D_x p_n(z_n^{(2)}), \xi_j)_{L^2} - (z_n^{(3)}, \xi_j)_{H_0^1}, \tag{4.6}$$

where the mapping $q_n : H_2^n \to H_1^n$ is given by $q_n(\psi_j) = \phi_j$, and $p_n : H_2^n \to H_3^n$ is given by $p_n(\psi_j) = \xi_i$ for $j = 1, \dots, n-1$.

Then we have system (3.4), (3.5) again except that at this time

$$(D_n)_{ij} = (D_x \phi_i, D_x \phi_j)_{L^2}, \qquad (F_n)_{ij} = (D_x \xi_i, \xi_j)_{L^2}. \tag{4.7}$$

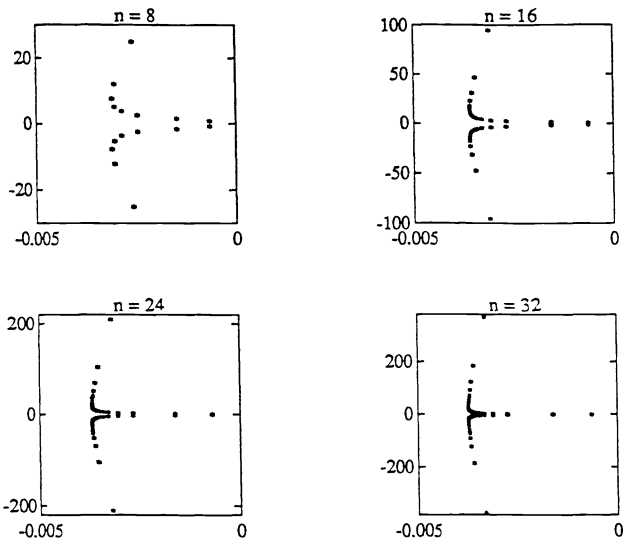


FIG. 4.4. *Location of the complex eigenvalues of the matrix A_n for the mixed finite element method in the case of Dirichlet–Dirichlet boundary conditions.*

We point out here that the only difference between the approximate system (3.4) for the finite element method and the mixed finite element method used in this section is the minor diagonal elements of matrix $M_n^{(2)}$. More precisely,

$$M^{(2)}_{n(\text{finite element})} = \begin{bmatrix} 1 & \frac{1}{4} & & & \\ \frac{1}{4} & 1 & \ddots & & \\ & \ddots & \ddots & & \\ & & \ddots & 1 & \frac{1}{4} \\ & & & \frac{1}{4} & 1 \end{bmatrix}, \tag{4.8}$$

and

$$M^{(2)}_{n(\text{mixed finite element})} = \begin{bmatrix} 1 & \frac{1}{2} & & & \\ \frac{1}{2} & 1 & \ddots & & \\ & \ddots & \ddots & & \\ & & \ddots & 1 & \frac{1}{2} \\ & & & \frac{1}{2} & 1 \end{bmatrix}. \tag{4.9}$$

However, this difference dramatically changes the location of the eigenvalues that are shown in Fig. 4.4. We can see that a uniform distance between the eigenvalues and imaginary axis is maintained. Thus this approximation scheme might be uniform exponentially stable. We tried to use the same contradiction argument as we did in §3 to prove it, but this question still remains open. The obvious disadvantage of this approximation scheme is the slow convergence.

Comparing Figs. 4.1, 4.3, and 4.4, we observe a common phenomenon. The eigenvalues of A_n converge to the eigenvalues of $\mathcal{A}$ in such a fashion that the lower frequency ones are

always more accurate than the higher frequency ones for each n. Among the three approximation schemes, the modal method is most favorable since it not only preserves the exponential stability uniformly, but also provides a rather fast convergence.

REFERENCES

[Ba] A. V. Balakrishnan, *Applied Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1981.

[BIW] H. T. Banks, K. Ito, and C. Wang, *Exponentially Stable Approximations of Weakly Damped Wave Equations*, preprint, 1991.

[BLM] J. A. Burns, Z. Y. Liu, and R. E. Miller, *Approximations of thermoelastic and viscoelastic control systems*, Numer. Functional Anal. Optim., 12 (1991), pp. 79–136.

[C] G. Chen, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures. Appl., 5 (1979), pp. 249–274.

[CP] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear System Theory*, Springer-Verlag, New York, 1978.

[D] A. Day, *Heat Conduction with Linear Thermoelasticity*, Springer-Verlag, New York, 1985.

[G] J. S. Gibson, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.

[GA] J. S. Gibson and A. Adamian, *Approximation theory for* LQG *optimal control of flexible structures*, SIAM J. Control Optim., 29 (1991), pp. 1–37.

[GR] J. S. Gibson and I. G. Rosen, *Numerical approximation for the infinite-dimensional discrete-time optimal Linear-Quadratic-Regulator problem*, SIAM J. Control Optim., 26 (1988), pp. 428–451.

[GRT] J. S. Gibson, I. G. Rosen, and G. Tao, *Approximation in control of thermoelastic systems*, SIAM J. Control Optim., 30 (1992), pp. 1163–1189.

[Ha] S. W. Hansen, *Exponential energy decay in a linear thermoelastic rod*, J. Math. Anal. Appl., 167 (1992), pp. 429–442.

[Hu] F. L. Huang, *Characteristic condition for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–56.

[HS] S. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.

[IK] K. Ito and F. Kappel, *On variational formulations of the Trotter–Kato theorem*, CAMS Rep. #91-7, Department of Mathematics, Univ. of Southern California, April, 1991.

[K] J. U. Kim, *On the energy decay of a linear thermoelastic bar and plate*, SIAM J. Math. Anal., 23 (1992), pp. 889–899.

[L1] J. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

[L2] ———, *Boundary stabilization of linear elastodynamics*, SIAM J. Control Optim., 21 (1983), pp. 968–983.

[LZ] Z. Y. Liu and S. Zheng, *Exponential stability of semigroup associated with thermoelastic system*, Quart. Appl. Math, 51 (1993), pp. 535–545.

[P] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[R] J. E. M. Rivera, *Energy decay rate in linear thermoelasticity*, Funkcial Ekvac, 35 (1992), pp. 19–30.

[S] M. Slemrod, *Global existence, uniqueness, and asymptotic stability of classical smooth solutions in one-dimensional nonlinear thermoelasticity*, Arch. Rat. Mech. Anal., 76 (1981), pp. 97–133.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1247–1265, September 1994

003

THE ASYMPTOTIC BEHAVIOR OF SIMULATED ANNEALING PROCESSES WITH ABSORPTION*

TZUU-SHUH CHIANG† AND YUNSHYONG CHOW†

Abstract. For a large class (satisfying a Perron–Frobenius property) of simulated annealing processes with an absorbing state a and arbitrary cost function, it is shown that there exist constants $h(i) \geq 0$, $\beta_{ij} > 0$ and $\delta > 0$, $N \geq 0$, independent of the starting points such that, for nonabsorbing states i and j,

$$\lim_{t\to\infty} \{P(X_t = i)\lambda^{h(j)}(t)\}/\{P(X_t = j)\lambda^{h(i)}(t)\} = \beta_{ij} \quad \text{and}$$

$$P(X_t \neq a) = \exp\left[-\int_0^t \delta\lambda^N(s) + O(\lambda^{N+1}(s))ds\right] \quad \text{for } t \text{ large.}$$

Here, $\lambda(t) = \exp(-1/T(t))$, $T(t) \to 0$, is the temperature function. As an application, the asymptotic behavior of the expected time of hitting a state (in particular, a global minimum) of a simulated annealing process without absorbing states can be determined.

Key words. simulated annealing process, forward equations, cycle method, Perron–Frobenius theorem

AMS subject classifications. primary 60J27, 60J99; secondary 15A18, 15A51, 90B40

0. Introduction.

For a finite set $S = \{1, 2, \ldots, n\}$, consider an inhomogeneous Markov process X_t on S with transition rates of the following type:

$$q_{ij}(t) = \begin{cases} p_{ij}\lambda(t)^{U(i,j)} & \text{for } j \neq i, \\ -\sum_{k\neq i} q_{ik}(t) & \text{for } j = i, \end{cases} \tag{0.0}$$

where $\lambda(t) = \exp(-1/T(t))$, $T(t) \to 0$, is a suitable temperature function, $P = (p_{ij})_{i,j\in S}$ is the neighborhood choosing-matrix with nonnegative entries, and $U : S \times S \to [0, \infty]$ is a cost function. We assume that $p_{ij} = 0$ if and only if $U(i, j) = \infty$. Processes of this type arise naturally in combinatorial optimization problems and are generally referred to as simulated annealing processes with cost U. Readers are referred to [3], [7], [8] for more details regarding the motivations, their physical meanings, and applications. If there is no absorbing state and (p_{ij}) forms an irreducible matrix, we call it a *regular simulated annealing process*. If there is a nontrivial absorbing state $a \in S$, that is, $p_{aj} = 0$ for all $j \neq a$ and $p_{ja} > 0$ for some $j \neq a$, and $(p_{jk})_{j,k\neq a}$ is irreducible, we call it a *singular simulated annealing process*. For the special case where $U(i, j) = (U(j) - U(i))^+$, a regular simulated annealing process has been used in practice to find the global minima of U [3], [7], [8]. Under some mild conditions on $\lambda(t)$ (sometimes necessary), it is proved in [1] that there exist constants $\beta_i > 0$ such that, with $d(i) = U(i) - \min U$,

$$\lim_{t\to\infty} P(X_t = i)/\lambda(t)^{d(i)} = \beta_i, \quad \forall i \in S, \tag{0.1}$$

independent of the starting points. The method used in [1] was to consider the following forward equation associated with X_t:

$$\vec{F}'(t) = Q^T(t) \cdot \vec{F}(t), \tag{0.2}$$

* Received by the editors May 31, 1989; accepted for publication (in revised form) April 1, 1993. This research was partially supported by the National Science Council of the Republic of China.

† Institute of Mathematics, Academia Sinica, Taipei, Taiwan 11529.

where $Q(t) = (q_{ij}(t))$ is as in (0.0), $\vec{F}(t) = (F_i(t);\ i \in S)^T$, and $F_i(t) = P(X_t = i)$. Thus with probability close to 1, the process will concentrate on the global minima set for large time [4], [5] and (0.1) gives the exact convergence rate. Similar results hold for regular simulated annealing processes with general cost functions ([2], [6], [10]), and a complicated cycle method was used to describe its "global minima" set, which is denoted as $\underline{S}$. Thus one of the important questions that remains to be answered is the expected time of hitting a global minimum. Let τ be the first hitting time of $\underline{S}$ and $F_i(t) = P(X_t = i,\ \tau > t)$. Then

$$E\tau = \int_0^\infty P(\tau > t)dt = \int_0^\infty \sum_{i \notin \underline{S}} F_i(t)dt, \tag{0.3}$$

and an easy calculation shows that $\{F_i(t)\}$ satisfies the following differential equations:

$$F_i'(t) = \sum_{j \notin \underline{S}} q_{ji}(t)F_j(t), \quad \forall\, i \notin \underline{S}. \tag{0.4}$$

Note that (0.4) is the forward equation associated with the singular-simulated annealing process Y_t, where X_t and Y_t have the same transitions except on $\underline{S}$, which is taken to be the absorbing state for Y_t. This motivates us to study the asymptotic behavior of singular cases. The cycle method mentioned above for treating regular processes cannot be directly applied here for (0.4) because there will eventually be no cycle in a singular system. A new condition (0.12) will be developed later to suit our purposes. To precisely describe the results, we first introduce some notations and definitions.

Let S, U be, as above, the state space and the nonnegative cost function on $S \times S$, respectively. We assume that U is *integer-valued*. It should be apparent from the discussion that this is only for technical reasons. In the singular case, which is the main concern in this paper, there is a (unique) absorbing state $a \in S$ such that $U(a, i) = \infty$ for any $i \neq a$ and $U(i, a) < \infty$ for some $i \neq a$. For any two states $i,\ j \in S$, a path connecting i to j is a sequence $i = i_0, i_1, \ldots, i_{k+1} = j$ such that $U(i_r, i_{r+1}) < \infty$ for $0 \leq r \leq k$. We say that $i \geq j$ (relative to U) if there exists a path $i = i_0, i_1, \ldots, i_{k+1} = j$ connecting i to j such that $U(i_r, i_{r+1}) = \min_{i_r \neq z \in S} U(i_r, z) < \infty$ for each $0 \leq r \leq k$. Thus $i \geq j$ if and only if there is a path connecting i to j such that all the costs in the intermediate steps are minimal. We say that $i \geq j$ at level h if, in addition to the above, $U(i_r, i_{r+1}) \leq h$ for all $0 \leq r \leq k$. A state i is said to be minimal if no state j can satisfy $i \geq j$ unless $j \geq i$. Two states i and j are equivalent (at level h), in notation $i \sim j (i \overset{h}{\sim} j)$, if $i \geq j$ (at level h) and $j \geq i$ (at level h). An hth order cycle is defined to be an equivalence class of minimal states under the relation "$\overset{h}{\sim}$." (We assume $i \overset{h}{\sim} i$ for any $i \in S$.) We remark that the absorbing state is only equivalent to itself for any equivalence relation described above. We define a hierarchy of states in S. First, let $(S^0, U^0) = (S, U)$ and $V^0(i) = V(i) = \min_{i \neq j \in S} U(i, j)$ for $i \in S$. Having defined (S^{n-1}, U^{n-1}) and V^{n-1}, let $S^n = \{c^n : c^n$ is an $(n-1)$th order cycle of $(S^{n-1}, U^{n-1})\}$ and define

$$d_{n-1}(c^n) = \max_{c^{n-1} \in c^n} V^{n-1}(c^{n-1}) \quad \text{as the "depth" of } c^n, \tag{0.5}$$

$$U^n(c^n, \tilde{c}^n) = d_{n-1}(c^n) + \min_{\substack{c^{n-1} \in c^n \\ \tilde{c}^{n-1} \in \tilde{c}^n}} \{U^{n-1}(c^{n-1}, \tilde{c}^{n-1}) - V^{n-1}(c^{n-1})\}, \tag{0.6}$$

$$V^n(c^n) = \min_{\tilde{c}^n \neq c^n} U^n(c^n, \tilde{c}^n), \tag{0.7}$$

$$h^n(c^n) = d_n(c^{n+1}) - V^n(c^n). \tag{0.8}$$

The new cost functions U^n appeared in combinatories in solving the minimal spanning tree problem, and the quantity $V^n(c^n)$ is simply the minimal cost coming out of c^n. The symbols c^n, $\tilde{c}^n$ are usually reserved for cycles in S^n. A cycle $c^n \in S^n$ is called nontrivial if c^n consists of more than one state in S^{n-1}. In that case $d_{n-1}(c^n) = n - 1$ and $V^n(c^n) \geq n$ by (0.5) and (0.7). A pair (S^n, U^n) is called indecomposable if no more nontrivial nth order cycle can be formed in S^{n+1} and $V^n(c^n) \leq n$ for every $c^n \in S^n$. If there is no absorbing state, then the sequence (S^0, U^0), $(S^1, U^1), \ldots, (S^n U^n) \cdots$ will eventually become trivial in the sense that S^{N+1} is a singleton from some N on. This case has been studied extensively in many papers ([1]–[8]). Because we are particularly interested in singular simulated annealing processes, the procedure of forming new pairs (S^n, U^n) generally becomes impossible at certain point N and we end up with an indecomposable pair (S^N, U^N), where S^N has more than one state and no more cycles can be formed.

For the first N such that (S^N, U^N) is indecomposable, we impose the following conditions on $\lambda(t)$:

$$\lambda(t) \to 0 \quad \text{and} \quad \lambda'(t)/\lambda(t) = o(\lambda^N(t)) \quad \text{as } t \to \infty. \tag{0.9}$$

It follows from (0.9) that

$$\int^{\infty} \lambda^N(t)dt = \infty. \tag{0.10}$$

Condition (0.9) requires the annealing rate $\lambda(t)$ decrease to 0 slowly, and (0.10) guarantees that the process is not trapped at any particular state forever. For the commonly used $\lambda(t) = t^{-1/c}$, (0.9) holds if and only if $c > N$. If $c < N$, it is well known that the process will be trapped at local minima and therefore no general statements can be made regarding its limiting behavior. See [5, Thm. 1].

The following notion of a transition rate matrix is used repeatedly.

DEFINITION 0.1. *An $m \times m$ matrix $A = (a_{ij})$ is called a transition rate matrix if* (i) *the column sums of A are nonpositive, that is, $\sum_{i=1}^{m} a_{ij} \leq 0$ for each j,* (ii) *$a_{ij} \geq 0$ for all $i \neq j$ and $a_{ii} \leq 0$ for each i.*

Note that the transpose of the matrix $Q(t) = (q_{ij}(t))$ in (0.0) is a transition rate matrix. However, we sometimes call $Q(t)$ a transition rate matrix for the sake of brevity.

We next define a Perron–Frobenius property and some quantities associated with (S^N, U^N) to describe the results. Let $S^N(N)$ denote the set of all states in S^N of which the minimal cost coming out is N, that is, $S^N(N) = \{c^N \in S^N : V^N(c^N) = N\}$. For each $c^N \in S^N(N)$, let $E(c^N)$ be the equivalence class containing c^N in $S^N(N)$. Note that S^N has no cycles by the assumption that (S^N, U^N) is indecomposable, but equivalence classes can still be defined. The forward equation of states in $E(c^N)$ again takes the form (0.2) and can be written as the following:

$$(F'_{\tilde{c}^N};\ \tilde{c}^N \in E(c^N))^T = Q^T_{E(c^N)} \cdot (\lambda^N F_{\tilde{c}^N};\ \tilde{c}^N \in E(c^N))^T + \text{higher order terms}, \tag{0.11}$$

where $Q^T_{E(c^N)}$ is a transition rate matrix with strictly negative eigenvalues. Let $\sigma(E(c^N))$ be the largest eigenvalue of $Q_{E(c^N)}$. The following condition is called Perron–Frobenius property.

(0.12)

There exists a unique $E(c^N) \subseteq S^N(N)$ such that $\sigma(E(c^N)) = \max_{\tilde{c}^N \in S^N(N)} \sigma(E(\tilde{c}^N))$.

Condition (0.12) holds automatically if $S^N(N)$ forms a single equivalence class. In any case it is a mild condition because for a fixed energy landscape U, $\sigma(E(c^N))$ depends only on the neighborhood choosing matrix $(p_{ij})_{i,j\in S}$. Because $(p_{ij})_{i,j\in S}$ is chosen arbitrarily, it happens with Lebesgue measure 0 that two matrices $Q_{E(c^N)}$ and $Q_{E(\tilde{c}^N)}$ can have the same largest eigenvalue.

Let $c^N_{\max}$ be any state in the unique solution of (0.12). The height of a state in S^N is defined as follows

$$h^N(c^N) = \min \sum_j U^N(c^N_j, c^N_{j+1}) - V^N(c^N_{j+1}), \tag{0.13}$$

where the minimum is taken over all paths $c^N_1, c^N_2, \cdots$ connecting $c^N_{\max}$ to c^N in S^N.

The following quantities are crucial to describing our results:

$$-\delta = \sigma(E(c^N_{\max})), \tag{0.14}$$

$$\text{For each } i \in S, \quad h(i) = \sum_{j=0}^{N} h^j(c^j), \tag{0.15}$$

where $i = c^0 \in c^1 \in \cdots \in c^N$ is the unique sequence of cycles containing i successively formed in (0.5)–(0.6), $h^j(c^j)$, $j = 1, \ldots, N-1$, are defined in (0.8) and $h^N(c^N)$ in (0.13). Some examples of δ and h^N are in Examples 2 and 3 in §4. We remark that unlike the regular case that S^N forms a cycle in S^{N+1}, in which case there is only one equivalence class in $S^N(N)$ and (0.12) is always satisfied with $\delta = 0$ in (0.14), there might be several different equivalence classes with the same (maximal, nonzero) largest eigenvalue. In this completely general case with no assumption made regarding $S^N(N)$, our analysis fails and there seems no natural way to generalize (0.8) as an appropriate height function. We therefore impose a mild condition (0.12) so that the natural generalization (0.13) can be defined to replace (0.8). Indeed, for any cycle C and states $c_1, c_2 \in C$, we always have $\lim_{t\to\infty} P(X_t \in c_1)\lambda(t)^{V(c_1)}/\{P(X_t \in c_2)\lambda^{V(c_2)}(t)\}$ exists ([2], Thm. 1.1). Hence the function V characterizes the relative weights among all the states in a cycle. The usual height function (0.8) thus describes the "inverse weight" of c with respect to C. ("c has height k with respect to C" means that $\lim_{t\to\infty} P(X_t \in c)/\{P(X_t \in C)\lambda^k(t)\}$ exists and is positive.) It is therefore clear that states $c \in C$ with maximal V, that is,

$$V(c) = \max_{\tilde{c}\in C} V(\tilde{c}), \tag{0.16}$$

have the largest probability in C and the probability of the process being at a different state $\tilde{c}$ is a factor $\lambda^{h(\tilde{c})}$ of it. In case that (S^N, U^N) does not form a cycle, which is crucial for the above argument to work, we must select a state in S^N that plays the same role in S^N as c does in (0.16). The proper choice we found in our analysis is a state satisfying the Perron–Frobenius property (0.12). The new function (0.13) is then tailor made as a height function relative to such a choice. Details are in §3. In the interesting case where $U(i,j) = [U(j) - U(i)]^+$, if i and j are neighbors and the absorbing state a is taken to be the global minima set of U, the forward equation of X_t will eventually have the form

$$dP(\tau > t)/dt = \left(\sum_{i\neq a} F_i(t)\right)' = (-\delta\lambda^N(t) + O(\lambda^{N+1}(t))) \cdot \sum_{i\neq a} F_i(t), \tag{0.17}$$

where τ is the hitting time of a. Here N has the geometric meaning as the greatest depth of all local minima as was first defined in [5]. The quantity $\delta \cdot \lambda^N$ can be interpreted as the

rate at which the whole system is attracted to a. The constant δ depends on the neighborhood choosing matrix (p_{ij}) and is not directly related to the geometric structure of U. Actually, for a finite set S with a given potential function U and fixed neighborhood system, the constant δ can take any (positive) value by manipulating p_{ij}'s. Examples 2 and 3 in §4 illustrate this point.

The main result in this paper can now be stated as follows.

THEOREM 0.2. *Let $\{X_t;\ t \geq 0\}$ be a singular-simulated annealing process satisfying* (0.9) *and the Perron–Frobenius property* (0.12). *Then there exist constants $\beta_{ij} > 0$ independent of the starting points such that for nonabsorbing states i and j,*

$$\begin{aligned} &\lim P(X_t = i)\lambda^{h(j)}(t)/\{P(X_t = j)\lambda^{h(i)}(t)\} = \beta_{ij} \quad \textit{and} \\ &P(X_t \neq \textit{ absorbing state}) = \exp\left[-\int_0^t \delta\lambda^N(s) + O(\lambda^{N+1}(s))ds\right] \quad \textit{for } t \textit{ large}, \end{aligned} \tag{0.18}$$

where $h(i)$ and δ are given in (0.15), (0.14), *respectively.*

Remark. For the case where $(p_{ij})_{i,j\neq a}$ is not irreducible, we need to break up the state space into irreducible classes. Relations (0.18) still hold in each class.

Remark. The Perron–Frobenius property is used to define the height function $h(i)$, $i \in S$ in (0.18). If it fails to hold, that is, there are more than one equivalence classes in $S^N(N)$ with the same maximal eigenvalues, we can conclude that (Corollary 2.4)

$$P(X_t = i)/\lambda^{h(i)}(t) = O\left(\exp\left[-\delta\int_0^t \lambda^N + O(\lambda^{N+1})ds\right]\right), \tag{0.19}$$

where $h(i) = \sum_{j=0}^N h^j(c^j)$ as in (0.15), but now h^N is the smallest height function computed relative to all the equivalence classes with the same maximal largest eigenvalue. This result is slightly weaker than (0.18), and the height function $h(i)$ can fail to represent the exact convergence rate of $P(X_t = i)$ for some i. See Example 4.2. From some examples that we have computed, it seems that (0.18) should hold even without the Perron–Frobenius property, but the correct height function h is very difficult to describe. We hope to address this point further in the future.

Next we mention a simple consequence of the Perron–Frobenius theorem [9] and list all the notations that are used throughout the paper.

LEMMA 0.3 [1], [2]. *Let $A = (a_{ij})$ be a transition rate matrix of order m. If $A^{-1} = (b_{ij})$ exists, then we have the following:*

(i) *All the eigenvalues of A have negative real parts.*

(ii) $b_{ii} \leq (\min_i a_{ii})^{-1}$ *and* $b_{ii} \leq b_{ij} \leq 0$ *for all i, j.*

(iii) $b_{ij} < 0$ *if and only if i is reachable from j, that is, there exist $i_0 = i, i_1, \ldots, i_k = j$ such that $a_{i_n,i_{n+1}} > 0$ for each $0 \leq n < k$. On the other hand, if A is noninvertible but irreducible, then we have the following:*

(iv) *Zero is an eigenvalue with multiplicity one and all other eigenvalues of A have negative real parts.*

(v) $\sum_{i=1}^m a_{ij} = 0$ *for each $1 \leq j \leq m$.*

(vi) *For any proper subset B of $\{1, 2, \ldots, m\}$, the principal minor $A_B = (a_{ij};\ i, j \in B)$ is an invertible transition rate matrix.*

$F_i(t) = P(X_t = i)$ for $i \in S$. $Q(t) = (q_{ij}(t))$ always denotes a transition rate matrix as in (0.0) for some forward equation. $A_{A,B}(k) = (r_{ij};\ i \in A,\ j \in B)$, where $r_{ij} = p_{ij}$ if $U(i, j) = k$ and 0 otherwise. Note that $Q_{A,B}(k)$ is a constant matrix. It is selected from $Q(t)$ for obvious reasons. $Q_A(k) = Q_{A,A}(k)$. $\vec{e} = (1, 1, \ldots, 1)$ of any dimension. A vector $\vec{v} = (v_1, \ldots, v_m)$ is said nonnegative (positive) if all $v_i \geq 0$ ($v_i > 0$). For the

pair (S^n, U^n), $S^n(k) = \{c^n \in S^n : V^n(c^n) = k\}$, $S^n(i, i+1, \ldots, j) = \cup_{k=i}^{j} S^n(k)$ and $S^n(i, \infty) = \cup_{k \geq i} S^n(k)$. For $c^n \in S^n$, $F_{c^n}(t) = \sum_{c^{n-1} \in c^n} F_{c^{n-1}}(t)$. Note that $F_{c^n}(t) = P(X_t \in c^n)$. For $A \subseteq S^n$, $\vec{F}_A(t) = (F_{c^n}(t);\ c^n \in A)^T$ and $\vec{F}'_A(t) = (F'_{c^n}(t);\ c^n \in A)^T$ are column vectors. The superscript T always means "transpose". For $A \subseteq S^n$, $F_A(t) = \vec{e} \cdot \vec{F}_A(t) = \sum_{c^n \in A} F_{c^n}(t)$. Note that $F_A(t) = P(X_t \in A)$.

We outline our proof and discuss the details in §§1–3.

Step 1. Beginning with (0.2), we establish (by a "boosting and merging" technique) that for each $n = 1, 2, \ldots, N$, the forward equation of states in (S^n, U^n) always preserve a form similar to (0.2) (Lems. 1.1 and 1.2),

$$F'_{c^n}(t) = \sum_{\tilde{c}^n} q_{\tilde{c}^n, c^n}(t) \cdot F_{\tilde{c}^n}(t) + \text{ higher order terms.} \tag{0.20}$$

The exact form of "higher order terms" can be found later. They are neglected for the time being.

Step 2. From (0.20), we show that states in a single cycle have comparable probabilities and their relation can be given as follows (Theorem 1.3). For $c^n \in c^{n+1}$,

$$F_{c^n} = \theta \lambda^{h^n(c^n)}(t) \cdot F_{c^{n+1}}(t) + \text{ higher order terms.}$$

Step 3. Let $S^N(k) = \{c^n \in S^N;\ V^N(c^N) = k\}$. We show that the states in $S^N(k)$, $k \leq N-1$, can be replaced by states in $S^N(N)$ and that the forward equation in $S^N(N)$ takes a form similar to (0.20).

Step 4. Suppose the Perron–Frobenius property (0.12) is satisfied. Then we establish the following comparison of F_{c^N} with $F_{c^N_{\max}}$:

$$F_{c^N} = \theta \cdot \lambda^{h^N(c^N)} \cdot F_{c^N_{\max}} + \text{ higher order terms} \quad \text{and}$$
$$F'_{c^N_{\max}}(t) = \left[-\delta \lambda^N(t) + O(\lambda^{N+1}(t))\right] \cdot F_{c^N_{\max}} \quad \text{(Lems. 3.1 and 3.2).}$$

In §1 we rederive all the estimates used in [1], [2], but omit unnecessary repetition of some technical details. Section 2 treats the indecomposable pair (S^N, U^N) and determines $h(i)$ and β_i for some states in S. Section 3 continues to work in §2 and determines $h(i)$ and β_i for all nonabsorbing states in S. Some representive examples are given in §4 to demonstrate the techniques we use and the application for the expected hitting time of regular simulated annealing processes.

1. Preliminary estimates. Let $\{X_t\}$ be a singular-simulated annealing process satisfying (0.9). The forward equation of nonabsorbing states associated with such a process assumes the following form:

$$\vec{F}'_{S\setminus\{a\}}(t) = Q^T(t) \cdot \vec{F}_{S\setminus\{a\}}(t),$$

where $Q(t) = (q_{ij}(t))_{i,j \neq a}$ is the transition rate matrix in (0.0). Because the absorbing state does not play a role in the analysis, we *abuse* the notation S for $S\setminus\{a\}$. In this section, two techniques called "boosting" and "merging" are used repeatedly to establish some preliminary estimates of $F_i(t)$ and single-out a unique class of states with maximal probability under a "Perron–Frobenius property" assumption. These estimates would be of correct order and yield the desired result if we had a regular simulated annealing process. Conceptually this section is equivalent to [1] or [2]. However, the notation and techniques are more complicated because the error terms are handled more delicately. This complication is of course expected because we have a more complicated process and its necessity becomes clear in the next section.

We begin with a "boosting" lemma that prepares the way for a "merging" in (S^0, U^0). The word "boosting" refers to the power of λ.

LEMMA 1.1 (Boosting). *Let (S^0, U^0) be a decomposable pair. Then for any $c^1 \in S^1(0)$, we have $F_{c^1} = O(\lambda F_{S^1(1,\infty)})$.*

Proof. Let $A = \{i \in S^0 : V(i) = 0$ and i is not contained in any nontrivial zeroth-order cycle in $S^1\}$. Note that $A = S^1(0)$. Then the forward equation of states in A assumes the following form:

$$\vec{F}'_A = Q_A^T(0) \cdot \vec{F}_A + O(\lambda F_S). \tag{1.1}$$

The fact that $Q_A^T(0)$ is a transition rate matrix implies that $\vec{e} \cdot Q_A(0) \leq 0$. Because the states in A cannot form any cycle in S^1, $Q_A^T(0)$ is invertible. Thus we can find a vector $\vec{v} > 0$ in the neighborhood of $\vec{e}$ such that $\vec{v} \cdot Q_A^T(0) < 0$ by the open mapping theorem. Therefore,

$$\vec{v} \cdot \vec{F}'_A = \vec{v} \cdot Q_A^T(0) \cdot \vec{F}_A + O(\lambda F_S) \leq -\alpha \vec{v} \cdot \vec{F}_A + O(\lambda F_{S \setminus A})$$

for some positive α when t is large. Note that the change of error term is justified by combining λF_A with $\vec{v} \cdot \vec{F}_A$. Let $f = \vec{v} \cdot \vec{F}_A$. Then $f' \leq -\alpha f + O(\lambda F_{S \setminus A})$. As in [1, Lem. 1.1], for some $M > 0$

$$\begin{aligned} \limsup_{t \to \infty} f(t)/(\lambda F_{S \setminus A}) &\leq \limsup_{t \to \infty} \int_0^t [(\exp(\alpha s)) \cdot M \lambda F_{S \setminus A}] ds / [(\exp(\alpha t)) \cdot \lambda F_{S \setminus A}] \\ &\leq \limsup_{t \to \infty} (M \lambda F_{S \setminus A})/(\alpha \lambda F_{S \setminus A} + O(\lambda^2 F_{S \setminus A})) \leq M/\alpha. \end{aligned}$$

Because $\vec{v} > 0$ and $F_{S \setminus A} = F_{S^1(1,\infty)}$, we therefore have $F_i = O(\lambda F_{S \setminus A}) = O(\lambda F_{S^1(1,\infty)})$ for each $i \in A$.

We remark that the l'Hôpital's rule was used in the above derivations. This is justified because of condition (0.10). Also, the term λ' did not appear when taking the derivative of the denominator because of (0.9). □

The following lemma indicates that the states forming a nontrivial cycle in S^1 can be merged into a "single" state and thus simplifies the forward equations in S^1. This explains its name.

LEMMA 1.2 (Merging). *Let $c^1 \in S^1$ be a nontrivial zeroth-order cycle. Then for any state $i \in c^1$, there exists a positive constant θ_i such that $F_i = \theta_i F_{c^1} + O(\lambda F_{S^1(1,\infty)})$. Thus the forward equation of any $c^1 \in S^1$ has the form*

$$F'_{c^1} = -p_{c^1} \lambda^{V^1(c^1)} + \sum_{\tilde{c}^1 \neq c^1} p_{\tilde{c}^1, c^1} \lambda^{U^1(\tilde{c}^1, c^1)} F_{\tilde{c}^1} + O(\lambda^2 F_{S^1(1,\infty)}),$$

where $p_{c^1} > 0$ and $p_{\tilde{c}^1, c^1} \geq 0$ are constants.

Proof. Because the states in c^1 form a cycle, their forward equations assume the following form because of Lemma 1.1:

$$\vec{F}'_{c^1} = Q_{c^1}^T(0) \cdot \vec{F}_{c^1} + O(\lambda F_{S^1(1,\infty)}),$$

where $Q_{c^1}^T(0)$ is a singular, irreducible transition rate matrix. Let $-\alpha$ be an eigenvalue of $Q_{c^1}^T(0)$ and $\vec{v}$ a corresponding eigenvector. By Lemma 2.2(iv), $\alpha = 0$ or Re $\alpha > 0$. Let $f = \vec{v} \cdot \vec{F}_{c^1}$. Then $f' = -\alpha f + O(\lambda F_{S^1(1,\infty)})$. If $\alpha = 0$, then $f' = O(\lambda F_{S^1(1,\infty)})$, trivially. If Re $\alpha > 0$, then

$$\begin{aligned}&\limsup_{t\to\infty}|f(t)|/(\lambda F_{S^1(1,\infty)})\\&\quad\le \limsup_{t\to\infty}\left\{\int_0^t(\exp[(\operatorname{Re}\alpha)s])\cdot M\lambda F_{S^1(1,\infty)}\right\}/\{(\exp[(\operatorname{Re}\alpha)t])\cdot\lambda F_{S^1(1,\infty)}\}\\&\quad\le \limsup_{t\to\infty} M\lambda F_{S^1(1,\infty)}/[(\operatorname{Re}\alpha)\lambda F_{S^1(1,\infty)}+O(\lambda^2F_{S^1(1,\infty)})]=M/(\operatorname{Re}\alpha).\end{aligned}$$

Thus $\vec{v}\cdot\vec{F}_{c^1}=O(\lambda F_{S^1(1,\infty)})$ and therefore $\vec{v}\cdot\vec{F}'_{c^1}=O(\lambda F_{S^1(1,\infty)})$. By Jordan's decomposition (see the proof of Lemma 2.2 in [2] for details), we can then find a basis $(\vec{v}_1,\vec{v}_2,\ldots)$ of $\mathbb{C}^{|c^1|}$ such that $\vec{v}_k\cdot\vec{F}'_{c^1}=O(\lambda F_{S^1(1,\infty)})$ for each k. Thus $\vec{F}'_{c^1}=O(\lambda F_{S^1(1,\infty)})$, and then so does $Q^T_{c^1}(0)\cdot\vec{F}_{c^1}$. The assertions in the lemma now follow by first noting $Q^T_{c^1}(0)$ is a singular irreducible transition rate matrix with rank $|c^1|-1$ and then solving linear equations. □

We remark that the forward equation in Lemma 1.2 makes sense only for the cycles c^1 and $\tilde{c}^1$ where $U^1(\tilde{c}^1,c^1)$ equals 0 or 1. For cycles $c^1\in S^1$ with $V^1(c^1)>1$, their forward equations are neglected for now and will be resumed one by one through the following induction.

Lemmas 1.1 and 1.2 are the induction foundation of the following theorem.

THEOREM 1.3 (Boosting and merging). *If (S^n,U^n) is decomposable, then for any $c^{n+1}\in S^{n+1}(0,1,\ldots,n)$, we have $\lambda^{V^{n+1}(c^{n+1})}F_{c^{n+1}}=O(\lambda^{n+1}F_{S^{n+1}(n+1,\infty)})$. Moreover, for a nontrivial nth-order cycle $c^{n+1}\in S^{n+1}$ and $c^n\in c^{n+1}$, there exists a positive constant θ_{c^n} such that*

$$\lambda^{V^n(c^n)}F_{c^n}=\theta_{c^n}\lambda^nF_{c^{n+1}}+O(\lambda^{n+1}F_{S^{n+1}(n+1,\infty)}). \tag{1.2}$$

The forward equation of any $c^{n+1}\in S^{n+1}$ has the following form:

$$\begin{aligned}F'_{c^{n+1}}=&-p_{c^{n+1}}\lambda^{V^{n+1}(c^{n+1})}F_{c^{n+1}}+\sum_{\tilde{c}^{n+1}\neq c^{n+1}}p_{\tilde{c}^{n+1},c^{n+1}}\lambda^{U^{n+1}(\tilde{c}^{n+1},c^{n+1})}F_{\tilde{c}^{n+1}}\\&+O(\lambda^{n+2}F_{S^{n+1}(n+1,\infty)}).\end{aligned} \tag{1.3}$$

Proof. We proceed by induction. When $n=0$, the theorem reduces to Lemmas 1.1 and 1.2. Suppose the theorem holds for (S^{n-1},U^{n-1}). Consider the forward equation of states in $S^n(T,k)=\{c^n\in S^n: V^n(c^n)=k,\ c^n$ is not contained in a nontrivial nth-order cycle in $S^{n+1}\}$. (Note that $S^n(T,k)=S^{n+1}(k),\ k=0,\ldots,n$.) By induction hypothesis,

$$\vec{F}'_{S^n(T,k)}=\sum_{r=0}^{n}Q^T_{S^n(T,r),S^n(T,k)}(r)\cdot(\lambda^r\vec{F}_{S^n(T,r)})+O(\lambda^{n+1}F_{S^n(n,\infty)}). \tag{1.4;k}$$

The following boosting and merging techniques play an essential role in this paper and will be used repeatedly in the sequel.

First, consider (1.4;k), $k=0$. Because $Q^T_{S^n(T,0)}(0)$ is an invertible transition rate matrix, we can find a basis $\vec{v}_1,\ldots,\vec{v}_{|S^n(T,0)|}$ so that $\vec{v}_k\cdot\vec{F}'_{S^n(T,0)}=O(\lambda^{n+1}F_{S^n(n,\infty)})$ by Jordan's decomposition. (See [2, Lems. 2.2 and 2.3] for details.) We thus can express $\vec{F}_{S^n(T,0)}$ in terms of $\{F_{c^n};\ c^n\in\cup_{i=1}^n S^n(T,i)\}$,

$$\begin{aligned}\vec{F}_{S^n(T,0)}=&-(Q^T_{S^n(T,0)}(0))^{-1}\cdot\left(\sum_{r=1}^{n}Q^T_{S^n(T,r),S^n(T,0)}(r)\right)\cdot(\lambda^r\vec{F}_{S^n(T,r)})\\&+O(\lambda^{n+1}F_{S^n(n,\infty)}).\end{aligned} \tag{1.5;0}$$

Substituting (1.5;0) into (1.4;k), $k = 1, 2, \ldots, n$, we obtain

$$
\begin{aligned}
\vec{F}'_{S^n(T,k)} = \sum_{r=1}^{n} \Big[& Q^T_{S^n(T,r),S^n(T,k)}(r) - Q^T_{S^n(T,0),S^n(T,k)}(0) \cdot \\
& (Q^T_{S^n(T,0)}(0))^{-1} \cdot Q^T_{S^n(T,r),S^n(T,0)}(r) \Big] \\
& \cdot (\lambda^r \vec{F}_{S^n(T,r)}) + O(\lambda^{n+1} F_{S^n(n,\infty)}) \\
= \sum_{r=1}^{n} & Q(1)^T_{S^n(T,r),S^n(T,k)}(r) \cdot (\lambda^r \vec{F}_{S^n(T,r)}) + O(\lambda^{n+1} F_{S^n(n,\infty)}).
\end{aligned} \tag{1.5;k}
$$

Having eliminated $S^n(T,0)$ from $S^n(T,k)$, $k = 1, \ldots, n$, the following lemma asserts that (1.5;k), $k = 1, \ldots, n$, again constitute a forward equation and thus enable us to eliminate $S^n(T,1), \ldots, S^n(T,n-1)$, successively.

LEMMA 1.4. *The system $Q(1)^T$, viewed as a transition among states in $\cup_{i=1}^n S^n(T,i)$, forms a forward equation, that is, for each $k = 1, \ldots, n$,*

$$
\begin{aligned}
\vec{e} \cdot \sum_{r=1}^{n} Q(1)^T_{S^n(T,k),S^n(T,r)}(k) = \vec{e} \cdot \Big[& Q^T_{S^n(T,k)}(k) - Q^T_{S^n(T,0)S^n(T,k)}(0) \\
& \cdot (Q^T_{S^n(T,0)}(0))^{-1} \cdot Q^T_{S^n(T,k),S^n(T,0)}(k) \Big] \\
+ \vec{e} \cdot \Bigg[& \sum_{\substack{r=1 \\ r \neq k}}^{n} Q^T_{S^n(T,k),S^n(T,r)}(k) - Q^T_{S^n(T,0),S^n(T,r)}(0) \\
& \cdot (Q^T_{S^n(T,0)}(0))^{-1} \cdot Q^T_{S^n(T,k),S^n(T,0)}(k) \Bigg] \\
\leq 0. &
\end{aligned}
$$

Proof. We only prove $k = 1$; the proof for the other k is the same. Consider the following identity:

$$
\left[\begin{array}{c|c}
 & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\
\hline
Q^T_{S^n(T,0)}(0) & \\
\hline
\sum_{r=1}^{n} \vec{e} \cdot Q^T_{S^n(T,0),S^n(T,r)}(0) & -1
\end{array}\right]
$$

$$
\cdot \left[\begin{array}{c|c}
 & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\
\hline
(Q^T_{S^n(T,0)}(0))^{-1} & \\
\hline
\sum_{r=1}^{n} \vec{e} \cdot Q^T_{S^n(T,0),S^n(T,r)} \cdot (Q^T_{S^n(T,0)}(0))^{-1} & -1
\end{array}\right] = I,
$$

where I is the identity matrix. Because the first term in the product is a transition rate matrix, all the entries of $\sum_{r=1}^{n} [\vec{e} \cdot Q^T_{S^n(T,0),S^n(T,r)}(0) \cdot (Q^T_{S^n(T,0)}(0))^{-1}]$ are nonpositive with absolute value less than or equal to one, by Lemma 0.2. Thus

$$\begin{aligned}
&\vec{e}\cdot\left[Q^T_{S^n(T,1)}(1)-Q^T_{S^n(T,0),S^n(T,1)}(0)\cdot\left(Q^T_{S^n(T,0)}(0)\right)^{-1}\cdot Q^T_{S^n(T,1),S^n(T,0)}(1)\right]\\
&\quad+\vec{e}\cdot\left(\sum_{r=2}^{n}Q^T_{S^n(T,1),S^n(T,r)}(1)-Q^T_{S^n(T,0),S^n(T,r)}(0)\right.\\
&\qquad\left.\cdot\left(Q^T_{S^n(T,0)}(0)\right)^{-1}\cdot Q^T_{S^n(T,1),S^n(T,0)}(1)\right)\\
&=\vec{e}\cdot Q^T_{S^n(T,1)}(1)+\vec{e}\cdot\sum_{r=2}^{n}Q^T_{S^n(T,1),S^n(T,r)}(1)\\
&\quad-\sum_{r=1}^{n}\vec{e}\cdot Q^T_{S^n(T,0),S^n(T,r)}(0)\cdot\left(Q^T_{S^n(T,0)}(0)\right)^{-1}\cdot Q^T_{S^n(T,1),S^n(T,0)}(1)\\
&\le\vec{e}\cdot Q^T_{S^n(T,1)}(1)+\vec{e}\cdot\sum_{r=2}^{n}Q^T_{S^n(T,1),S^n(T,r)}(1)+\vec{e}\cdot Q^T_{S^n(T,1),S^n(T,0)}(1)\le 0,
\end{aligned}$$

because

$$Q^T_{\cup_{i=1}^{n}S^n(T,i)}(1)$$

is a transition rate matrix. This proves the lemma. □

Having eliminated $S^k(T,k)$, $k=0,\dots,n-1$, we have

$$\vec{F}'_{S^n(T,n)}=Q(n)^T_{S^n(T,n)}(n)\cdot(\lambda^n\vec{F}_{S^n(T,n)})+O(\lambda^{n+1}F_{S^n(n,\infty)}),\tag{1.6}$$

where $Q(n)^T_{S^n(T,n)}(n)$ is an invertible transition rate matrix. We are thus in the same situation as that of Lemma 1.1. Therefore, $\vec{F}_{S^n(T,n)}=O(\lambda F_{S^{n+1}(n+1,\infty)})$. (Note the change in the error term from that in (1.6).) Having boosted the power of λ by one, we then substitute this back to obtain $\lambda^k\vec{F}_{S^n(T,n)}=O(\lambda^{n+1}F_{S^{n+1}(n+1,\infty)})$. This completes the boosting part of the lemma. Let $c^{n+1}\in S^{n+1}$ be a nontrivial nth-order cycle. We next consider the merging of states in c^{n+1}. The forward equation of states in $c^{n+1}(k)=\{c^n\in c^{n+1}: V^n(c^n)=k\}$ assumes the following form:

$$\vec{F}'_{c^{n+1}(k)}=\sum_{i=0}^{n}Q^T_{c^{n+1}(i),c^{n+1}(k)}(i)\cdot(\lambda^i\vec{F}_{c^{n+1}(i)})+O(\lambda^{n+1}F_{S^{n+1}(n+1,\infty)}),$$

for $k=0,1,\dots,n$. Note that the error term has the present form because of the order estimates just obtained in the proof. Applying the boosting technique $(n-1)$ times, we obtain (similar to (1.6)) the following:

$$\vec{F}'_{c^{n+1}(n)}=Q(n)^T_{c^{n+1}(n)}(n)\cdot(\lambda^n\vec{F}_{c^{n+1}(n)})+O(\lambda^{n+1}F_{S^{n+1}(n+1,\infty)}),$$

where Q is a singular transition rate matrix. Thus we are back to the situation in Lemma 1.2, and (1.2) thus follows. We are now ready to prove (1.3). To boost the order of the error term from λ^{n+1} to λ^{n+2}, we have to boost the order of all error terms in previous steps. First, let $c^1\in S^{n+1}(1)\subseteq S^1(1)$. If c^1 is a nontrivial zeroth-order cycle in $S^1(1)$, then the forward equation of states in c^1 assumes the following form:

$$\vec{F}'_{c^1}=Q^T_c(0)\cdot\vec{F}_{c^1}+O(\lambda^{n+1}F_{S^{n+1}(n+1,\infty)}).\tag{1.7}$$

The error term in (1.7) differs from that in (1.1) by an nth power of λ. This is because we have now updated the error estimates. The method in Lemma 1.2 then implies that $F_i = \theta_i F_{c^1} + O(\lambda^{n+1} F_{S^{n+1}(n+1,\infty)})$ for each $i \in c^1$. Suppose $c^2 \in S^{n+1}(2) \subseteq S^2(2)$ and c^2 is a nontrivial first-order cycle. Then the forward equation of states in $c^2(k) = \{c^1 \in c^2 : V^1(c^1) = k\}$ assumes the following form:

$$(1.8;k) \quad \vec{F}'_{c^2(k)} = \sum_{i=0}^{1} Q^T_{c^2(i),c^2(k)}(i) \cdot (\lambda^i \vec{F}_{c^2(i)}) + O(\lambda^{n+1} F_{S^{n+1}(n+1,\infty)}), \qquad k = 0, 1.$$

The boosting and merging techniques thus imply that for $c^1 \in c^2$,

$$\lambda^{V^1(c^1)} F_{c^1} = \theta_{c^1} \cdot \lambda F_{c^2} + O(\lambda^{n+1} F_{S^{n+1}(n+1,\infty)}).$$

We inductively obtain that for nontrivial $(k-1)$th-order cycle $c^k \in S^{n+1}(k) \subseteq S^k(k)$, $k = 1, 2, \ldots, n$,

$$\lambda^{V^{k-1}(c^{k-1})} F_{c^{k-1}} = \theta_{c^{k-1}} \cdot \lambda^{k-1} F_{c^k} + O(\lambda^{n+1} F_{S^{n+1}(n+1,\infty)})$$

holds for $c^{k-1} \in c^k$. Thus the forward equation of a state $c^{n+1} \in S^{n+1}$ has the following form:

$$F'_{c^{n+1}} = -p_{c^{n+1}} \lambda^{V^{n+1}(c^{n+1})} F_{c^{n+1}} + \sum_{k=0}^{n+1} \sum_{\substack{V^{n+1}(\tilde{c}^{n+1})=k \\ \tilde{c}^{n+1} \neq c^{n+1}}} p_{\tilde{c}^{n+1},c^{n+1}} \lambda^k F_{\tilde{c}^{n+1}}$$

$$+ O(\lambda^{n+2} F_{S^{n+1}(n+1,\infty)}).$$

This completes the proof. □

Remark. From the above proof, it is clear that for a cycle C with forward equation $(F'_i;\ i \in C) = Q \cdot (\lambda^{V(i)} F_i;\ i \in C)+$ higher order terms, (1.2) can be obtained by solving $Q \cdot (\lambda^{V(i)} F_i;\ i \in C) = 0$. Indeed, because the rank of Q is one less than its dimension, we have $\lambda^{V(i)} F_i = \beta_{ij} \lambda^{V(j)} F_j$ for some $\beta_{ij} > 0$.

2. Indecomposable pair and Perron–Frobenius property. In this section we define a Perron–Frobenius property of a singular-simulated annealing process, which guarantees the existence of a class of dominant (but equivalent) states and then establish an intermediate theorem.

Let $\{X_t\}$ be a singular-simulated annealing process on S with cost function U, and N the smallest number such that (S^N, U^N) is indecomposable. By Theorem 1.3, the forward equation of states in $S^N(k)$, $0 \leq k \leq N$, assumes the following form:

$$(2.0;k) \qquad \vec{F}'_{S^N(k)} = \sum_{i=0}^{N} Q^T_{S^N(i),S^N(k)}(i) \cdot (\lambda^i \vec{F}_{S^N(i)}) + O(\lambda^{N+1} F_{S^N(N)}),$$

where Q is the transition rate matrix on S^N.

Following the boosting and merging technique once, we can eliminate the term $\vec{F}_{S^N(0)}$ from $\{\vec{F}_{S^N(k)};\ k > 0\}$ and obtain for $k = 1, \ldots, N$,

$$\vec{F}'_{S^N(k)} = \sum_{i=1}^{N} \Big[Q^T_{S^N(i),S^N(k)}(i) - Q^T_{S^N(0),S^N(k)}(0) \cdot (Q^T_{S^N(0)}(0))^{-1} \cdot$$

$$(2.1;k) \qquad Q^T_{S^N(i),S^N(0)}(i) \Big] \cdot (\lambda^i \vec{F}_{S^N(i)}) + O(\lambda^{N+1} F_{S^N(N)})$$

$$= \sum_{i=1}^{N} Q(1)^T_{S^N(i),S^N(k)}(i) \cdot (\lambda^i \vec{F}_{S^N(i)}) + O(\lambda^{N+1} F_{S^N(N)}).$$

After applying the techniques of boosting and merging N times, we obtain the following:

$$(2.2;N) \qquad \vec{F}'_{S^N(N)} = Q(N)^T_{S^N(N)}(N) \cdot (\lambda^N \vec{F}_{S^N(N)}) + O(\lambda^{N+1} F_{S^N(N)}),$$

where $Q(N)^T_{S^N(N)}(N)$ is an invertible transition rate matrix. Recall that two states c_1^N and c_2^N in $S^N(N)$ are said to be equivalent if $c_1^N \geq c_2^N$ and $c_2^N \geq c_1^N$ (see the Introduction). Denote the set of all the equivalence classes of $S^N(N)$ by $E(S^N(N)) = \{S^N(N,i) : S^N(N,i)$ is an equivalence class of $(S^N(N), \geq)\}$. We remark that the equivalence classes can also be defined from $(S^N(N), Q(N)^T_{S^N(N)}(N))$, where $Q(N)^T_{S^N(N)}(N)$ is viewed as an ordinary transition rate matrix. Indeed, the following identity and Lemma 0.3 show that an element $[-Q^T_{S^N(0),S^N(k)}(0) \cdot (Q^T_{S^N(0)}(0))^{-1}](c_1^N, c_2^N) > 0$, where $c_1^N \in S^N(k)$ and $c_2^N \in S^N(0)$ if and only if $c_2^N \geq c_1^N$:

$$\left[\begin{array}{c|ccc} Q^T_{S^N(0)}(0) & 0 & & \\ \hline Q^T_{S^N(0),S^N(k)}(0) & \begin{matrix}-1\\ \cdot\\ 0\end{matrix} & \begin{matrix}0\\ \cdot\\ -1\end{matrix} & \end{array}\right] \cdot$$

$$\left[\begin{array}{c|ccc} (Q^T_{S^N(0)}(0))^{-1} & 0 & & \\ \hline Q^T_{S^N(0),S^N(k)}(0) \cdot (Q^T_{S^N(0)}(0))^{-1} & \begin{matrix}-1\\ \cdot\\ 0\end{matrix} & \begin{matrix}0\\ \cdot\\ -1\end{matrix} & \end{array}\right] = I,$$

where I is the identity matrix. From this fact, it follows that $[Q(1)^T_{S^N(i),S^N(k)}(i)]\,(c_1^N, c_3^N) > 0$ where $c_1^N \in S^N(k)$ and $c_3^N \in S^N(i)$ if and only if $U^N(c_3^N, c_1^N) = N$ or $c_3^N \geq c_2^N \geq c_1^N$ for some $c_2^N \in S^N(0)$. In either case $c_3^N \geq c_1^N$. We thus conclude inductively that in (2.2;N), $[Q(N)^T_{S^N(N)}(N)]\,(c_1^N, c_2^N) > 0$ for $c_1^N, c_2^N \in S^N(N)$ if and only if $c_2^N \geq c_1^N$ in (S^N, U^N). Let $\sigma(S^N(N,i)) < 0$ be the largest eigenvalue of the transition rate matrix among states of $Q(N)^T_{S^N(N,i)}(N)$. The Perron–Frobenius property (0.12) that we need guarantees the existence of a unique equivalence class $S^N(N, \max)$ in $E(S^N(N))$ such that $\sigma(S^N(N, \max)) = -\delta$, the maximum of $\sigma(S^N(N,i))$ over $E(S^N(N))$.

We remark that $-\delta$ can be obtained as follows. We first equate the right-hand side of (2.0;k) to 0 for $k = 0, 1, \ldots, N-1$. Because this is a linear system with more unknowns than equations, we can solve $\lambda^k F_{S^N(k)}$ in terms of $\lambda^N F_{S^N(N)}$ for $k = 0, 1, \ldots, N-1$. Substituting $\lambda^k F_{S^N(k)}$ into the forward equation of $S^N(N)$, we obtain (2.2;N) and $-\delta$ is the largest eigenvalue of the coefficient matrix $Q(N)_{S^N(N)}$.

The next lemma shows that $S^N(N, \max)$ is the dominant term among all equivalence classes in $S^N(N)$. First, we say that $S^N(N,i) \leq S^N(N,j)$ if there are states $c_1^N \in S^N(N,i)$ and $c_2^N \in S^N(N,j)$ such that $c_1^N \leq c_2^N$.

LEMMA 2.1. *Let $S^N(N,i)$ be an equivalence class in $E(S^N(N))$. Then $F_{S^N(N,i)} = O(F_{S^N(N,\max)})$. To be more precise, if $S^N(N,i) \not\leq S^N(N, \max)$, then $F_{S^N(N,i)} =$*

$O(\lambda F_{S^N(N,\max)})$, *and if* $c^N \in S^N(N,i) \leq S^N(N,\max)$, *then there exists a constant* θ_{c^N} *such that* $F_{c^N} = \theta_{c^N} \cdot F_{S^N(N,\max)} + O(\lambda F_{S^N(N,\max)})$.

Proof. Let $A = \cup_{S^N(N,i) \not\leq S^N(N,\max)} S^N(N,i)$, $B = \cup_{S^N(N,i) \leq S^N(N,\max)} S^N(N,i)$, and $D = S^N(N) \backslash S^N(N,\max)$. The forward equation of states in D assumes the following form by (2.2;N):

(2.3)

$$\vec{F}'_D = Q(N)^T_D(N) \cdot (\lambda^N \vec{F}_D) + Q(N)^T_{S^N(N,\max),D}(N) \cdot (\lambda^N \vec{F}_{S^N(N,\max)}) + O(\lambda^{N+1} F_{S^N(N)}).$$

Let $\vec{v} > 0$ be a vector such that $\vec{v} \cdot Q(N)^T_D(N) < 0$ and let $f = \vec{v} \cdot \vec{F}_D$. Then

$$f' \leq -\alpha\lambda^N f + \vec{v} \cdot Q(N)^T_{S^N(N,\max),D}(N) \cdot (\lambda^N \vec{F}_{S^N(N,\max)}) + O(\lambda^{N+1} F_{S^N(N,\max)}),$$

where the constant $-\alpha$ can be chosen so that $-\alpha < -\alpha_1 = -\delta = \sigma(S^N(N,\max))$ by the Perron–Frobenius theorem. Let $\vec{v}_1 > 0$ be an eigenvector corresponding to $(-\delta)$ of the matrix $Q(N)^T_{S^N(N,\max)}(N)$. Then

$$\begin{aligned}
&\limsup_{t\to\infty} f(t)/(\vec{v}_1 \cdot \vec{F}_{S^N(N,\max)}) \\
&\quad \leq \limsup_{t\to\infty} \left[\int_0^t \left(\exp \int_0^s \alpha\lambda^N\right) \cdot (\vec{v}_1 \cdot \vec{F}_{S^N(N,\max)})\right] \Big/ \left[\left(\exp \int_0^t \alpha\lambda^N\right) \cdot (\vec{v}_1 \cdot \vec{F}_{S^N(N,\max)})\right] \\
&\quad \leq \limsup_{t\to\infty} (\vec{v}_1 \cdot \vec{F}_{S^N(N,\max)})/[(\alpha - \delta) \cdot (\vec{v}_1 \cdot \vec{F}_{S^N(N,\max)})] = 1/(\alpha - \delta).
\end{aligned}$$

(The last inequality holds because of l'Hôpital's rule and is the main reason why we need the P–F property.) Thus we have $F_{S^N(N,i)} = O(F_{S^N(N,\max)})$ for any $S^N(N,i) \in E(S^N(N))$. Then consider the forward equation of states in A,

$$\vec{F}'_A = Q(N)^T_A(N) \cdot (\lambda^N \vec{F}_A) + O(\lambda^{N+1} F_{S^N(N,\max)}).$$

The same technique applied to (2.3) can now be applied again to yield $F_{S^N(N,i)} = O(\lambda F_{S^N(N,\max)})$ for $S^N(N,i) \not\leq S^N(N,\max)$. Having obtained this estimate, the forward equation of states in B can be written as follows:

$$\begin{aligned}
\vec{F}'_B &= Q(N)^T_B(N) \cdot (\lambda^N \vec{F}_B) + Q(N)^T_{A,B}(N) \cdot (\lambda^N \vec{F}_A) + O(\lambda^{N+1} F_{S^N(N,\max)}) \\
&= Q(N)^T_B(N) \cdot (\lambda^N \vec{F}_B) + O(\lambda^{N+1} F_{S^N(N,\max)}).
\end{aligned}$$

Let $\vec{v}_k$ be an eigenvector of $Q(N)^T_B(N)$ corresponding to the eigenvalue $(-\alpha_k)$. If $-(\operatorname{Re}\ \alpha_k) < -\delta$, then, similar to the above, we obtain $\vec{v}_k \cdot \vec{F}_B = O(\lambda F_{S^N(N,\max)})$, and therefore

(2.4)
$$\begin{bmatrix} \vec{v}_1 \\ \vec{v}_2 \\ \vdots \\ \vec{v}_{|B|} \end{bmatrix} \cdot \vec{F}_B = \begin{bmatrix} \vec{v}_1 \cdot \vec{F}_{S^N(N,\max)} \\ O(\lambda F_{S^N(N,\max)}) \\ \vdots \\ O(\lambda F_{S^N(N,\max)}) \end{bmatrix}.$$

Note that $\vec{v}_1$ is the eigenvector corresponding to the largest eigenvalue $-\delta = \sigma(S^N(N,\max))$ and $\vec{v}_1 \cdot \vec{F}_B = \vec{v}_1 \cdot \vec{F}_{S^N(N,\max)}$. Let L be the matrix on the left side of (2.4). Multiplying both

sides of (2.4) by L^{-1} and observing that the first column of L^{-1} is strictly positive by Lemma 0.3(iii), we then conclude the proof of the lemma. □

We can rephrase Lemma 2.1 by saying that for a state $c^N \in S^N(N)$, if $h^N(c^N)$ (relative to any state in $S^N(N, \max)$) is zero, then $F_{c^N} = \theta_{c^N} \cdot F_{S^N(N,\max)} + O(\lambda F_{S^N(N,\max)})$ and $F_{c^N} = O(\lambda F_{S^N(N,\max)})$ if $h^N(c^N) \geq 1$. Now consider states in $S^N(N-1)$. The forward equation of states in $S^N(N-1)$ assumes the following form:

$$\begin{aligned}\vec{F}'_{S^N(N-1)} &= Q(N-1)^T_{S^N(N-1)}(N-1) \cdot (\lambda^{N-1}\vec{F}_{S^N(N-1)}) \\ &\quad + Q(N-1)^T_{S^N(N),S^N(N-1)}(N) \cdot (\lambda^N \vec{F}_{S^N(N)}) + O(\lambda^{N+1}F_{S^N(N,\max)}).\end{aligned}$$

Let $A = \{c^N : c^N \in S^N(N-1) \text{ and } h^N(c^N) \geq 2\}$. Then the forward equation of states in A has the following simpler expression

$$\begin{aligned}\vec{F}'_A &= Q(N-1)^T_A(N-1) \cdot (\lambda^{N-1}\vec{F}_A) \\ &\quad + Q(N-1)^T_{S^N(N),A}(N) \cdot (\lambda^N \vec{F}_{S^N(N)}) + O(\lambda^{N+1}F_{S^N(N,\max)}) \\ &= Q(N-1)^T_A(N-1) \cdot (\lambda^{N-1}\vec{F}_A) + O(\lambda^{N+1}F_{S^N(N,\max)}).\end{aligned}$$

Thus $\vec{F}_A = O(\lambda^2 F_{S^N(N,\max)})$. On the other hand, for states in $B = \{c^N : c^N \in S^N(N-1)$ and $h^N(c^N) = 1\}$, we have

$$\begin{aligned}\vec{F}'_B &= Q(N-1)^T_B(N-1) \cdot (\lambda^{N-1}\vec{F}_B) \\ &\quad + Q(N-1)^T_{S^N(N),B}(N) \cdot (\lambda^N \vec{F}_{S^N(N)}) + O(\lambda^{N+1}F_{S^N(N,\max)}).\end{aligned}$$

We thus obtain

$$\vec{F}_B = [Q(N-1)^T_B(N-1)]^{-1} \cdot Q(N-1)^T_{S^N(N),B}(N) \cdot (\lambda \vec{F}_{S^N(N)}) + O(\lambda^2 F_{S^N(N,\max)})$$

and $F_{c^N} = \theta_{c^N} \cdot \lambda F_{S^N(N,\max)} + O(\lambda^2 F_{S^N(N,\max)})$ for $c^N \in B$. Inductively, we conclude that the following lemma holds.

LEMMA 2.2. *Let $c^N \in S^N(k)$ with $h^N(c^N) = N - k$. Then there exists a constant $\theta_{c^N} > 0$ such that $F_{c^N} = \theta_{c^N} \cdot \lambda^k F_{S^N(N,\max)} + O(\lambda^{k+1}F_{S^N(N,\max)})$. If $c^N \in S^N(k)$ but $h^N(c^N) > N - k$, then $F_{c^N} = O(\lambda^{k+1}F_{S^N(N,\max)})$.*

By using Lemma 2.1 and (2.4), we have $F'_{S^N(N,\max)} = -\delta\lambda^N F_{S^N(N,\max)} + O(\lambda^{N+1}F_{S^N(N,\max)})$. It is trivial to solve this differential equation, which yields

$$F_{S^N(N,\max)}(t) = c_0 \exp\left[-\delta \int_{t_0}^{t} (\lambda^N(s) + O(\lambda^{N+1}))ds\right]$$

for t large. We now state an intermediate theorem regarding the order estimates of some states in S.

THEOREM 2.3. *Let $\{X_t\}$ be a singular-simulated annealing process satisfying* (0.9) *and* (0.12). *Let N be the smallest integer such that (S^N, U^N) is indecomposable. Then there exist constants $\beta_{ij} > 0$ and a function $\beta(t) = O(\lambda^{N+1}(t))$ such that* (0.18) *holds for any $i, j \in \cup c^N$, where $c^N \in S^N(k)$ with $h^N(c^N) = N - k$ for some k. If $i \in c^N \in S^N(k)$ with $h^N(c^N) > N - k$, then*

$$P(X_t = i) = O\left(\lambda^{d(i)+N-k+1} \cdot \exp\left[-\delta \int_0^t (\lambda^N(s) + \beta(s))ds\right]\right),$$

where $d(i) = \sum_{j=0}^{N-1} h^j(c^j)$.

Proof. The proof is a combination of Lemmas 2.1, 2.2, and Theorem 1.3. □

From the proof above, we can conclude, without assuming the Perron–Frobenius property (0.12), the following weaker form of (0.18).

COROLLARY 2.4. *Let* $\{X_t\}$ *be a singular-simulated annealing process satisfying* (0.9). *Let* N *be the smallest integer such that* (S^N, U^N) *is indecomposable. Then for some function* $\beta(t) = O(\lambda^{N+1}(t))$,

$$P(X_t = i) = O\left(\lambda^{h(i)} \exp\left[-\delta \int_0^t \lambda^N(s) + \beta(s)ds\right]\right) \quad \textit{for } t \textit{ large.}$$

Note that in (0.19), $h(i)$ is defined to be $\sum_{j=0}^{N} h^j(c^j)$ as in (0.15). Only $h^N(c^N)$ now needs an explanation because of lack of the Perron–Frobenius property. In (0.13), the defining equation of $h^N(c^N)$, we now take the minimum over all paths connecting $\tilde{c}^N$ to c^N, where $\tilde{c}^N$ is in any equivalence class of $S^N(N)$ with the maximal largest eigenvalue.

3. Exact asymptotic behavior. We continue to seek the exact asymptotic behavior of each state in S. Again, we have to rederive all our previous estimates with the updated information contained in Theorem 2.3.

LEMMA 3.1. *Let* $c^N \in S^N(k) \subseteq S^k(k)$ *be such that* $h^N(c^N) \geq N - k + 1$. *Then for any* $c^{k-1} \in c^N$, *there exists a positive constant* $\theta_{c^{k-1}}$ *such that*

$$\lambda^{V^{k-1}(c^{k-1})} F_{c^{k-1}} = \theta_{c^{k-1}} \cdot \lambda^{k-1} F_{c^N} + O(\lambda^{N+1} F_{S^N(N,\max)}).$$

Proof. Let c^N be a nontrivial zeroth-order cycle in $S^1(1)$. If $h^N(c^N) \geq N$, its forward equation has the following form:

$$\vec{F}'_{c^N} = Q^T_{c^N}(0) \cdot \vec{F}_{c^N} + O(\lambda^{N+1} F_{S^N(N,\max)}).$$

The error term is of order $O(\lambda^{N+1} F_{S^N(N,\max)})$ because $h^N(c^N) \geq N$ and the contribution from states other than c^N is of order $O(\lambda^{N+1} F_{S^N(N,\max)})$ by Lemma 2.2. Thus, by the same technique as in Lemma 1.2, we obtain $F_i = \theta_i \cdot F_{c^N} + O(\lambda^{N+1} F_{S^N(N,\max)})$ for $i \in c^N$. Suppose Lemma 3.1 is true for $n \leq r$ and let $c^N \in S^N(r+1)$ be such that $h^N(c^N) \geq N - r$. The forward equation of states in $c^N(j) = \{c^r \in c^N : V^r(c^r) = j\}$ has the following form:

$$(3.1;j) \qquad \vec{F}'_{c^N(j)} = \sum_{k=0}^{r} Q^T_{c^N(k), c^N(j)}(k) \cdot (\lambda^k \vec{F}_{c^N(k)}) + O(\lambda^{N+1} F_{S^N(N,\max)}).$$

The boosting and merging technique can now be applied to (3.1;j), $j = 1, \ldots, r$, to obtain $\lambda^{V^r(c^r)} F_{c^r} = \theta_{c^r} \lambda^r F_{c^N} + O(\lambda^{N+1} F_{S^N(N,\max)})$ for any $c^r \in c^N$. This completes the proof. □

Lemma 3.1 implies that the forward equation of any state c^N in $A = \{c^N \in S^N : h^N(c^N) \geq N - V^N(c^N) + 1\}$ assumes the following form:

$$F'_{c^N} = -p_{c^N} \lambda^{V^N(c^N)} F_{c^N} + \sum_{c^N \neq \tilde{c}^N \in A} p_{\tilde{c}^N, c^N} \lambda^{V^N(\tilde{c}^N)} F_{\tilde{c}^N} + \sum_{\tilde{c}^N \in S^N \setminus A} p_{\tilde{c}^N, c^N} \lambda^{V^N(\tilde{c}^N)+1} F_{\tilde{c}^N} + O(\lambda^{N+2} F_{S^N(N,\max)}).$$

Let $B = S^N \backslash A$, $B(k) = \{c^N \in B : V^N(c^N) = k\}$, and $A(k) = \{c^N \in A : V^N(c^N) = k\}$, $0 \le k \le N$. Then in matrix form,

$$\begin{aligned}\vec{F}'_{A(k)} &= \sum_{j=0}^{N} Q^T_{A(j),A(k)}(j) \cdot (\lambda^j \vec{F}_{A(j)}) \\ &\quad + \sum_{j=0}^{N} Q^T_{B(j),A(k)}(j+1) \cdot (\lambda^{j+1} \vec{F}_{B(j)}) + O(\lambda^{N+2} F_{S^N(N,\max)}).\end{aligned} \tag{3.2;k}$$

The purpose of Lemma 3.1 is to generate the second term on the right-hand side of (3.2;k), which has not played any role in previous sections but is of vital importance now.

LEMMA 3.2. *Let $c^N \in S^N(N)$ with $h^N(c^N) \ge 2$. Then $F_{c^N} = O(\lambda^2 F_{S^N(N,\max)})$. Moreover, for any $c^N \in S^N(k)$ with $h^N(c^N) \ge N-k+2$, we have $F_{c^N} = O(\lambda^{n-k+2} F_{S^N(N,\max)})$.*

Proof. The proof is basically a repetition of that of Theorem 2.3. Having successively applied the boosting and merging techniques n times, we have

$$\begin{aligned}\vec{F}'_{A(N)} &= Q(N)^T_{A(N)}(N) \cdot (\lambda^N \vec{F}_{A(N)}) \\ &\quad + \sum_{j=0}^{N} Q(N)^T_{B(j),A(N)}(j+1) \cdot (\lambda^{j+1} \vec{F}_{B(j)}) + O(\lambda^{N+2} F_{S^N(N,\max)}).\end{aligned} \tag{3.3}$$

Let $D = \{c^N \in A(N) : h^N(c^N) \ge 2\}$. Because $Q(N)^T_{A(N)\backslash D,D}(N) = 0$ and $Q(N)^T_{B(N),D}(N+1) = 0$,

$$\vec{F}'_D = Q(N)^T_D(N) \cdot (\lambda^N \vec{F}_D) + O(\lambda^{N+2} F_{S^N(N,\max)}).$$

This implies that $\vec{F}_D = O(\lambda^2 F_{S^N(N,\max)})$ as in Lemma 1.1. Similarly, we consider

$$\begin{aligned}\vec{F}'_{A(N-1)} &= Q(N-1)^T_{A(N-1)}(N-1) \cdot (\lambda^{N-1} \vec{F}_{A(N-1)}) \\ &\quad + \sum_{j=0}^{N} Q(N-1)^T_{B(j),A(N)}(j+1) \cdot (\lambda^{j+1} \vec{F}_{B(j)}) + O(\lambda^{N+2} F_{S^N(N,\max)})\end{aligned}$$

and obtain, with $G = A(N-1) \cap \{c^N : h^n(c^N) \ge 3\}$,

$$\vec{F}'_G = Q(N-1)^T_G(N-1) \cdot (\lambda^{N-1} \vec{F}_G) + O(\lambda^{N+2} F_{S^N(N,\max)}).$$

Thus $\vec{F}_G = O(\lambda^3 F_{S^N(N,\max)})$. The lemma follows now by an obvious induction. □

We are now ready to prove the main theorem of this paper.

Proof of Theorem 0.2. Let $i \in c^N \in S^N(k) \subseteq S^k(k)$ with $h^N(c^N) = N-k+1$. From (3.3), the forward equation of states in $A(N) \cap \{c^N : h^N(c^N) = 1\} = A(N,1)$ assumes the following form:

$$\begin{aligned}\vec{F}'_{A(N,1)} &= Q(N)^T_{A(N,1)}(N) \cdot (\lambda^N \vec{F}_{A(N,1)}) \\ &\quad + \sum_{j=0}^{N} Q(N)^T_{B(j),A(N,1)}(j+1) \cdot (\lambda^{j+1} \vec{F}_{B(j)}) + O(\lambda^{N+2} F_{S^N(N,\max)}).\end{aligned}$$

Note that all terms of the form $\{\lambda^N F_{c^N} : c^N \in A(N), h^N(c^N) \ge 2\}$ have been absorbed into the error term by Lemma 3.2 and therefore do not appear in (3.3). The same proof

of Lemma 2.2 can now be applied to yield $F_{c^N} = \theta_{c^N} \cdot \lambda F_{S^N(N,\max)} + O(\lambda^2 F_{S^N(N,\max)})$ for each $c^N \in A(N,1)$ and θ_{c^N} is a positive constant. The forward equation of states in $A(N-1) \cap \{c^N : h^N(c^N) = 2\} = A(N-1,2)$ can be written as

$$\begin{aligned}\vec{F}'_{A(N-1,2)} &= Q(N-1)^T_{A(N-1,2)}(N-1) \cdot (\lambda^{N-1}\vec{F}_{A(N-1,2)}) \\ &\quad + Q(N-1)^T_{A(N,1),A(N-1,2)}(N) \cdot (\lambda^N \vec{F}_{A(N,1)}) \\ &\quad + \sum_{j=0}^{N} Q(N-1)^T_{B(j),A(N-1,2)}(j+1) \cdot (\lambda^{j+1}\vec{F}_{B(j)}) + O(\lambda^{N+2}F_{S^N(N,\max)}).\end{aligned}$$

This, again, will yield that $F_{c^N} = \theta_{c^N} \cdot \lambda^2 F_{S^N(N,\max)} + O(\lambda^3 F_{S^N(N,\max)})$ for each $c^N \in A(N-1,2)$. The theorem is now completed for any state $i \in H_r = \{i : i \in c^N \in S^N(k)$ with $h^N(c^N) = N - k + r$ for some $k\}$, where $r = 0$ or 1. A similar induction on r can obviously complete the proof of the theorem. □

4. Examples.

Example 1. $S = \{1,2,3,4,a\}$ and has a linear neighborhood structure. The cost function $U(i,j) = (U(j) - U(i))^+$ is determined by a potential function $U(\cdot)$ with $U(1) = U(3) = 1$ and $U(2) = U(4) = 3$. It is easy to compute that $U(1,2) = U(3,2) = U(3,4) = 2$ and $U(2,1) = U(2,3) = U(4,3) = U(4,a) = 0$. The forward equation of $\{1,2,3,4\}$ is as follows:

$$\begin{aligned}F_1' &= -p_1\lambda^2 F_1 + p_{21}F_2, \\ F_2' &= -p_2 F_2 + p_1\lambda^2 F_1 + p_{32}\lambda^2 F_3, \\ F_3' &= -p_3\lambda^2 F_3 + p_{23}F_2 + p_{43}F_4, \\ F_4' &= -p_4 F_4 + p_{34}\lambda^2 F_3,\end{aligned} \tag{4.1}$$

where $p_1 = p_{12}$, $p_2 = p_{21} + p_{23}$, $p_3 = p_{32} + p_{34}$, and $p_4 = p_{43} + p_{4a}$. In matrix form, the forward equation of $\{2, 4\}$ is as follows:

$$\begin{pmatrix}F_2\\F_4\end{pmatrix}' = \begin{pmatrix}-p_2 & 0\\0 & -p_4\end{pmatrix}\begin{pmatrix}F_2\\F_4\end{pmatrix} + \begin{pmatrix}p_1 & p_{32}\\0 & p_{34}\end{pmatrix}\begin{pmatrix}\lambda^2 F_1\\\lambda^2 F_3\end{pmatrix}.$$

The boosting and merging technique (Theorem 1.3) implies that (see (1.5;0))

$$\begin{pmatrix}F_2\\F_4\end{pmatrix} = \begin{pmatrix}p_2 & 0\\0 & p_4\end{pmatrix}^{-1}\begin{pmatrix}p_1 & p_{32}\\0 & p_{34}\end{pmatrix}\begin{pmatrix}\lambda^2 F_1\\\lambda^2 F_3\end{pmatrix} + O(\lambda^3 F_1 + \lambda^3 F_3). \tag{4.2}$$

By eliminating F_2, F_4 from the forward equation of $\{1,3\}$,

$$\begin{aligned}\begin{pmatrix}F_1\\F_3\end{pmatrix}' &= \left\{\begin{pmatrix}-p_1 & 0\\0 & -p_3\end{pmatrix} + \begin{pmatrix}p_{21} & 0\\p_{23} & p_{43}\end{pmatrix}\begin{pmatrix}p_1 p_2^{-1} & p_{32}p_2^{-1}\\0 & p_{34}p_4^{-1}\end{pmatrix}\right\}\begin{pmatrix}\lambda^2 F_1\\\lambda^2 F_3\end{pmatrix} \\ &\quad + O(\lambda^3 F_1 + \lambda^3 F_3) \\ &= Q(2)^T_{S^2(2)}(2)\begin{pmatrix}\lambda^2 F_1\\\lambda^2 F_3\end{pmatrix} + O(\lambda^3 F_1 + \lambda^3 F_3).\end{aligned} \tag{4.3}$$

(See (2.2;N).) Because $Q(2)^T_{S^2(2)}(2)$ is irreducible, the Perron–Frobenius property is thus satisfied. Lemma 2.1 then implies the existence of positive constants θ_1 and θ_3 such that

$$F_i = \theta_i(F_1 + F_3) + O(\lambda(F_1 + F_3)), \qquad i = 1,3. \tag{4.4}$$

The relative orders among all F_i's are specified by (4.2) and (4.4). To find their asymptotic behaviors we multiply (4.3) by an eigenvector $\vec{v} = (v_1, v_3)$ corresponding to the largest eigenvalue, say $-\delta$, of $Q(2)^T_{S^2(2)}(2)$. By the Perron–Frobenius theorem, $\delta > 0$ and v_1, v_3 can be assumed positive. This leads to

$$g' = -\delta(\lambda^2 + O(\lambda^3)) \cdot g, \quad \text{where } g = v_1 F_1 + v_3 F_3.$$

An integration shows

$$g(t) \approx \exp\left[-\delta \int_0^t (\lambda^2(s) + O(\lambda^3))ds\right].$$

The asymptotic behaviors of F_i's then follow immediately. Because $h^2(2) = h^2(4) = 2$, this result is in accordance with Theorem 0.2.

Example 2. $S = \{1, 2, a\}$. Let $U(1) = 1$, $U(2) = 2$, $U(i, j) = [U(j) - U(i)]^+$, and $U(i, a) = 0$ for $i, j = 1, 2$. The forward equation of $\{1, 2\}$ is as follows:

$$F_1' = -(p_1 + p_{12}\lambda)F_1 + p_{21}F_2,$$
$$F_2' = -p_2 F_2 + p_{12}\lambda F_1.$$

In this case, $N = 0$ and the Perron–Frobenius property is satisfied if and only if $p_1 \neq p_2$. Let $-\delta = \max(-p_1, -p_2)$. If $-\delta = -p_1 > -p_2$, then

$$F_1 = \exp\left(-\delta t + \int_0^t O(\lambda)ds\right)$$

and $F_2 = O(\lambda) \cdot F_1$. If $-\delta = -p_2 > -p_1$, then

$$F_1 = \exp\left(-\delta t + \int_0^t O(\lambda)ds\right)$$

and $F_2 = O(1) \cdot F_1$.

On the other hand, suppose $p_1 = p_2 (= 1$ for brevity). Then $(F_2 - \lambda^{1/2}F_1)' = -(F_2 - \lambda^{1/2}F_1) + O(\lambda^{1/2}(F_2 + \lambda^{1/2}F_1))$. Hence

$$F_2 - \lambda^{1/2}F_1 = e^{-t}\left(c + \int_0^t e^s\, O(\lambda^{1/2}(F_2 + \lambda^{1/2}F_1))ds\right)$$

and by l'Hôpital's rule,

$$F_2 - \lambda^{1/2}F_1 = o(F_2 + \lambda^{1/2}F_1) \quad \text{if } \int^\infty \lambda^{1/2} = \infty.$$

Thus $\lim P(X_t = 1) \cdot \lambda^{1/2}(t)/P(X_t = 2) = \beta > 0$. Note that the height function h defined in the second remark after Theorem 0.2 is 0 and does not reflect the exact convergence rate for state 2.

Example 3. $S = \{1, 2, a\}$ and $U(2, 1) = U(2, a) = 0$, $U(1, 2) = 1$. This example intends to show that the expected time of hitting a global minimum can be infinity for a regular-simulated annealing process. The forward equation of $\{1, 2\}$ is as follows:

$$F_1' = -p_1\lambda F_1 + p_{21}F_2,$$
$$F_2' = -p_2 F_2 + p_1\lambda F_1,$$

where $p_1 = p_{12}$ and $p_2 = p_{21} + p_{2a}$. Take $T(t) = c/(\log t)$. Then $\lambda(t) = \exp(-1/T(t)) = t^{-1/c}$ and $\int^{\infty} \lambda(t)dt = \infty$ if and only if $c \geq 1$. If $0 < c < 1$, then

$$F_1(t) \geq F_1(t_0) \exp\left(-p_1 \int_{t_0}^{\infty} \lambda(s)ds\right) > 0$$

and thus $P(\tau = \infty) > 0$, where τ is the hitting time of state a for a regular-simulated annealing process with $\{a\} = \{\text{global minimum}\}$. If $c \geq 1$, then [1, Cor. 0.3] and Theorem 0.2 are applicable. In particular, $P(\tau < \infty) = 1$. From Theorem 1.3, we have $F_2 = p_1 p_2^{-1} \lambda F_1 + O(\lambda^2 F_1)$. Thus $F_1' = -p_1(1 - p_{21}p_2^{-1})\lambda F_1 + O(\lambda^2 F_1)$. Obviously,

$$F_1(t) \approx \exp\left[(-p_1(1 - p_{21}p_2^{-1})) \int_0^t (\lambda(s) + O(\lambda^2))ds\right]$$

asymptotically. By (0.3), $E\tau \approx \int_0^{\infty} F_1(t)dt$. It is thus easy to see that $E\tau < \infty$ if and only if (i) $c > 1$ or (ii) $c = 1$ and $p_1(1 - p_{21}p_2^{-1}) > 1$.

Acknowledgment. We thank the referee for his suggestions for improving the presentation of this paper.

REFERENCES

[1] T. S. Chiang and Y. Chow, *On the convergence rate of annealing processes*, SIAM J. Control Optim., 26 (1988), pp. 1455–1470.

[2] ———, *On the asymptotic behavior of some inhomogeneous Markov processes*, Ann. Probab., 17 (1989), pp. 1483–1502.

[3] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intelligence, 6 (1984), pp. 721–741.

[4] B. Gidas, *Global optimization via the Langevin equation*, in Proc. 24th IEEE Conf. Decision and Control, Fort Lauderdale, FL, 1985, pp. 774–778.

[5] B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.

[6] C. R. Hwang and S. J. Sheu, *Singular perturbed Markov chains and exact behaviors of simulated annealing process*, J. Theoret. Probab., 5 (1992), pp. 223–249.

[7] S. Kirkpatrick, C. Gelatt, and M. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[8] P. J. M. Van Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht, 1987.

[9] E. Seneta, *Nonnegative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, New York, 1981.

[10] A. Trouve, *Problèmes de convergence et d'ergodicité pour les algorithmes de recuit parallélisés*, C. R. Acad. Sci. Paris Ser. I Math., 307 (1988), pp. 161–164.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1266–1288, September 1994

004

ASYMPTOTIC FIRST HITTING-TIME DISTRIBUTION OF ANNEALING PROCESSES*

CHRISTIAN MAZZA†

Abstract. This paper presents asymptotics for the distribution of the first hitting time τ of $E_{\min}$ for the continuous-time version of simulated annealing processes. The method considers the backward equation associated with the process. It is shown that under certain assumptions, it is possible to characterize the eigenvectors of the transition matrix with the help of polynomials that are related to some families of graphs.

Key words. simulated annealing, continuous time, first hitting time, digraph, eigenvector

AMS subject classifications. primary 60J27, 60J99; secondary 15A51, 15A18, 90B40

1. Introduction.

Let S be a finite set and let $E : S \to \mathbb{R}$ be an arbitrary real-valued function. Simulated annealing is a Monte Carlo method for locating the minima of $E(\cdot)$ on S, i.e., $E_{\min} := \{i \in S;\ E(i) = \min_{j\in S} E(j)\}$. This probabilistic algorithm is defined to be a time nonhomogeneous Markov chain $X(t)$, $t \in \mathbb{N}$, on S with transition matrix

$$P_{ij}(T(t)) = q_{ij} \exp(-[E(j) - E(i)]^+ T(t)^{-1}), \qquad i \neq j, \tag{1}$$

where $[\cdot]^+$ denotes the positive part, and $T(t)$ is a positive function converging to 0 as $t \to \infty$, called temperature function. (q_{ij}) is the transition matrix of an irreducible Markov chain on S. The question of convergence for this kind of processes has been studied by many authors [Hajek (1988)], [Chiang and Chow (1988b)], [Holley and Stroock (1988)]. It has been proved [Hajek (1988)] that there exists a constant $D \geq 0$ such that $\lim_{t\to\infty} P(X(t) \in E_{\min}) = 1$ if and only if $\sum_{t=1}^{\infty} \exp(-DT(t)^{-1}) = +\infty$. Define τ to be the first hitting time of $E_{\min}$. This article presents asymptotics for the distribution of τ for the continuous-time version of the Markov chain $X(t)$. The process runs as follows (see, e.g., [Feller (1950)]). Let $q_i(t) := \sum_{j\neq i} P_{ij}(T(t))$ and $\Gamma(i,j,t) := P_{ij}(T(t))q_i(t)^{-1}$, for $i \neq j$; $\Gamma(i,i,t) := 0$ for all t, for all i. If at epoch t the process is in state i, the probability that, between t and $t+h$, a move occurs is $q_i(t)h + o(h)$. $\Gamma(i,j,t)$ is interpreted as the conditional probability that, if a move from i occurs between t and $t+h$, the process jumps to j. Intuitively the process can be described as follows: The process being at time t at location x waits for an exponential time Δ with mean 1 before it selects at random a neighbor y with probability q_{xy}. When $E(y) \leq E(x)$, the process moves to y. However, when $E(y) > E(x)$, it moves to y with probability $\exp(-T(t+\Delta)^{-1}(E(y)-E(x)))$, and with probability $1-\exp(-T(t+\Delta)^{-1}(E(y)-E(x)))$ it stays at x. The sample functions of $X(\cdot)$ are S-valued right-continuous step functions. Mathematically, the process is described via its transition semigroup, which is expressed by a system of differential equations.

1.1. The Fokker–Planck equation.

Let $P_{ij}(t',t)$ be the transition function associated with the annealing process. Consider the operator

$$L_T\Phi(i) := \sum_{j\in S} (\Phi(j) - \Phi(i))P_{ij}(T), \qquad i \in S, \tag{2}$$

for $\Phi : S \to \mathbb{R}$. Transition probabilities are determined by the Fokker–Planck, or forward equation,

$$\frac{\partial}{\partial t} P_{ij}(t',t) = [L^*_{T(t)} P_{i,\cdot}(t',t)](j), \qquad t \geq t', \tag{3}$$

* Received by the editors March 20, 1991; accepted for publication (in revised form) March 5, 1993. This work was partially supported by the Swiss National Science Foundation.

† Institut de Mathématiques, Université de Fribourg, CH-1700 Fribourg, Switzerland.

$$P_{ij}(t',t') = \delta_{ij}, \tag{4}$$

where (3) means

$$\frac{\partial}{\partial t}[P(t',t)\Phi](i) = [P(t',t)L_{T(t)}\Phi](i), \qquad t \geq t',$$

for $\Phi : S \to \mathbb{R}$, and where

$$[P(t',t)\Phi](i) := \sum_{j\in S} \Phi(j)P_{ij}(t',t). \tag{5}$$

$P_{ij}(t',t)$ satisfies the Chapman–Kolmogorov equation

$$P_{ij}(t',t) = \sum_{k\in S} P_{ik}(t',t'')P_{kj}(t'',t), \qquad 0 \leq t' < t'' < t,$$

which, in addition to the Fokker–Planck equation, yields the backward equation

$$\frac{\partial}{\partial t'}[P(t',t)\Phi](i) = -[L_{T(t')}P(t',t)\Phi](i), \qquad 0 \leq t' \leq t. \tag{6}$$

Once the initial distribution ν has been specified, the transition kernel $P(t',t)$ completely determines the annealing process.

1.2. First hitting-time distribution and backward equation. Let $\tau := \inf\{t \geq 0;\ X(t) \in E_{\min}\}$ be the first hitting time of $E_{\min}$. Consider the process $\tilde{X}(\cdot)$ on S with

$$\tilde{P}_{ij}(T(t)) := \begin{cases} P_{ij}(T(t)) & \text{if } i \in S - E_{\min}, \\ 0 & \text{if } i \in E_{\min} \text{ and } i \neq j, \\ 1 & \text{if } i = j \text{ and } i \in E_{\min}, \end{cases}$$

as infinitesimal generator. Start $X(\cdot)$ and $\tilde{X}(\cdot)$ at time t_0 at location $i \in S - E_{\min}$. Up to τ, the distribution of $\tilde{X}(\cdot)$ is then the same as the distribution $X(\cdot)$. Let $\Phi : S \to \mathbb{R}$. The backward equation for the absorbing chain $\tilde{X}(\cdot)$,

$$\frac{\partial}{\partial t'}[\tilde{P}(t',t)\Phi](i) = -[\tilde{L}_{T(t')}\tilde{P}(t',t)\Phi](i), \qquad 0 \leq t' \leq t, \tag{7}$$

permits a characterization of the distribution of the first hitting time τ. Indeed, choose the indicator function of $S - E_{\min}$ for Φ. Then, for $i \in E_{\min}^c$,

$$[\tilde{P}(t',t)\Phi](i) = \sum_{j\in S-E_{\min}} P(\tilde{X}(t) = j|\tilde{X}(t') = i) \tag{8}$$

$$= P(\tilde{X}(t) \in E_{\min}^c|\tilde{X}(t') = i) = P(\tau > t|\tilde{X}(t') = i). \tag{9}$$

Let $y(i,t',t) := P(\tau > t|\tilde{X}(t') = i)$. As $[\tilde{P}(t',t)\Phi](i) = y(i,t',t)$, we obtain, by (7),

$$\begin{aligned} \frac{\partial}{\partial t'}y(i,t',t) &= -\sum_{k\in S}[\tilde{P}(t',t)\Phi](k)\tilde{P}_{ik}(T(t')) + [\tilde{P}(t',t)\Phi](i) \\ &= y(i,t',t) - \sum_{k\in S-E_{\min}} y(k,t',t)\tilde{P}_{ik}(T(t')). \end{aligned}$$

In matrix form, if $Y(t',t) := (y(i,t',t))_{i\in S-E_{\min}}$, we have

$$\frac{\partial Y(t',t)}{\partial t'} \equiv A(t')Y(t',t), \qquad 0 \le t' \le t, \tag{10}$$

$$Y(t,t) = (1,\dots,1)^{\mathrm{Tr}}, \tag{11}$$

where Tr denotes the matrix transpose, $A(t') = Id - B(T(t'))$, and $B(T(t'))$ is the matrix obtained from $P(T(t'))$ by deleting the rows and the columns corresponding to states of $E_{\min}$. An alternative approach could be to consider the forward equation for the absorbing chain $\tilde{X}(\cdot)$ (as in [Lawler and Sokal (1988)]) to obtain asymptotics for the probabilities $P(\tilde{X}(t) = j|\tilde{X}(t') = i)$, $i, j \in E_{\min}^c$, and therefore for (8). More precisely, set $\Phi := 1_{\{j\}}$, $j \in E_{\min}^c$ in the forward equation for $\tilde{X}(\cdot)$; using the fact that $\tilde{X}(\cdot)$ is absorbed on $E_{\min}$, we obtain

$$\frac{\partial}{\partial t}\bar{Y}(t',t) \equiv -\bar{Y}(t',t)A(t), \qquad t' \le t, \tag{12}$$

where $\bar{Y}(t',t)$ is the row vector

$$\bar{Y}(t',t) \equiv (\tilde{P}_{ij}(t',t))_{j\in E_{\min}^c}$$

and the matrices $A(t)$, $t \in \mathbb{R}$ are as above. In both approaches we are confronted with linear differential systems related to the matrices $A(t)$. In the special case of time-homogeneous Markov chains, i.e., $A(t) \equiv A$, the matrices A and $\exp(\int_{t'}^{t} A(u)du) \equiv \exp((t-t')A)$ commute, and the relations $y(i,t',t) \equiv \bar{Y}(t',t)(1,\dots,1)^{\mathrm{Tr}}$, and (12) yield

$$\frac{\partial}{\partial t}Y(t',t) \equiv -AY(t',t), \qquad t \ge t',$$

$$Y(t',t') = (1,\dots,1)^{\mathrm{Tr}}.$$

A powerful result of [Levinson (1948)] (see, e.g., [Eastham (1989)]) yields explicit formulas for the solutions of (10) in terms of the eigenvalues and eigenvectors of the generator. A study of the spectrum can be done along ideas of [Freidlin and Wentzell (1984)] on large deviations. We characterize the eigenvectors with the help of polynomials related to some families of digraphs. Under certain assumptions it is then possible to obtain asymptotics of the form

$$P(\tau \ge t|X(t_0) = k) = (p_k + o(1))\exp\left(-b_k\int_{t_0}^{t}\exp(-D(k)T(v)^{-1})dv\right), \tag{13}$$

as $t \to \infty$, for $k \in S - E_{\min}$, where $p_k \ge 0$, $b_k > 0$, and $D(k) \ge 0$. Although the results are obtained under certain hypotheses, the method has the advantage of not being limited to the simulated annealing case, and could be useful for studying Markov chains with transition probabilities of order $\varepsilon(t)^{V_{ij}}$, where $\varepsilon(t) \to 0$ as $t \to \infty$, and $V_{ij} \le +\infty$ are nonnegative constants (see, e.g., [Connors and Kumar (1989)]). As stated above, the first ingredient of the approach is a theorem of Levinson on systems of differential equations.

THEOREM 1.2.1 [Levinson (1948)]. *Let $\Lambda(t)$ be a diagonal matrix*

$$\Lambda(t) = \mathrm{Diag}(\lambda_1(t),\dots,\lambda_N(t)),$$

which verifies condition L. For all integer pairs (i,j) in $[1,N]$, $i \ne j$ and for all t', t such that $t_0 \le t' \le t$, either

$$\int_{t'}^{t}\Re(\lambda_i(v) - \lambda_j(v))dv \le K_1 \tag{14}$$

or

$$\int_{t'}^{t} \Re(\lambda_i(v) - \lambda_j(v))dv \geq K_2, \tag{15}$$

where K_1 and K_2 are two real constants. Let $R(t)$ be an $N \times N$ matrix satisfying

$$\int_0^{\infty} |R(t)|dt < \infty, \tag{16}$$

where $|R(t)|$ is the matrix maximum norm. Then, when $t \to \infty$, the differential system

$$\frac{dx}{dt} = \{\Lambda(t) + R(t)\}x(t) \tag{17}$$

has solutions $x_u(t)$, $1 \leq u \leq N$, such that

$$x_u(t) = (e_u + o(1)) \exp\left(\int_{t_0}^{t} \lambda_u(v)dv\right), \tag{18}$$

where e_u is the uth unit vector.

2. Spectral analysis. To obtain information on solutions of (10), we use Levinson's theorem, which applies to differential systems in the so called "Levinson form" (17); knowledge about the spectrum and about the eigenvectors of $B(T)$ is then necessary to put the system in Levinson form by diagonalization.

2.1. Wentzell approximations. Wentzell approximations concerning large deviations for Markov processes can be used in our situation [Wentzell (1972)]. Let $P(T) = (P_{ij}(T))$ be a family of stochastic matrices of the form

$$P_{ij}(T) = l_{ij} \exp(-V_{ij}T^{-1}), \qquad i \neq j, \tag{19}$$

where l_{ij}, $V_{ij} \leq \infty$ are nonnegative constants.

DEFINITION 2.1.1. *Let $W \subset S$. We associate with each $W \subset S$ a set of digraphs $G(W)$, called Wentzell graphs, as follows. By definition, a digraph $g \in G(W)$ with node set S is a family of arrows $(i \to j)$, $i, j \in S$, $i \neq j$, satisfying the following conditions*: (1) *g does not contain any cycle*, (2) *for each $i \in S - W$, g contains a unique arrow starting at i*; (3) *for all $i \in W$, g contains no arrow starting at i. ($G(S)$ consists of just the empty graph). $G^{(u)}$, $1 \leq u \leq |S|$, is the set of the W-graphs for all $W \subset S$, $|W| = u$. Let $g \in G(W)$, $W \subset S$, and $G := \cup_{W \subset S} G(W)$. Consider the function*

$$E(\cdot) : G \to \mathbb{R} \tag{20}$$

defined by

$$E(i \to j) := V_{ij}, \qquad E(g) := \sum_{(i \to j) \in g} V_{ij}. \tag{21}$$

Set

$$V^{(u)} := \min_{g \in G^{(u)}} E(g) \tag{22}$$

for $u = 1, \ldots, |S|$ ($V^{(|S|)} := 0$).

LEMMA 2.1.2 [Wentzell (1972)]. *Let* $A := \mathrm{Diag}(1-a_1, 1-a_2, \ldots, 1-a_N)$; *let* $(P_{ij}) = P \in \mathbb{R}^{N\times N}$ *be a stochastic matrix. Then* $\det(A-P)$ *is a multinomial in the indeterminates* a_i; *the coefficient of* $(-a_{i_1})\cdots(-a_{i_n})$ *is the sum of products* $\pi(g) := \prod_{(i\to j)\in g} P_{ij}$, *where the sum is taken over all the graphs of* $G\{i_1,\ldots,i_n\}$. *Let* $a = 1-\lambda$, *where* λ *is an eigenvalue of* P. *Then*

$$A^1(-a) + A^2(-a)^2 + \cdots + A^N(-a)^N = 0, \tag{23}$$

where

$$A^u := \sum_{g\in G^u} \pi(g). \tag{24}$$

THEOREM 2.1.3 [Wentzell (1972)]. *Let* $\lambda_1(T) = 1$, *and let* $\lambda_2(T),\ldots,\lambda_N(T)$ *denote the eigenvalues of the matrix* $P(T)$, *arranged in order of decreasing real part. Then, for* $u = 2,\ldots,N$,

$$\Re(1-\lambda_u(T)) \asymp \exp(-(V^{(u-1)} - V^{(u)})T^{-1}), \tag{25}$$

where $\Re(\cdot)$ *denotes the real part and* $\asymp$ *indicates that* $\lim_{T\to 0} T\log(1-\Re(\lambda_u(T))) = -(V^{(u-1)} - V^{(u)})$, *for all* u. *Moreover, the broken line through the points* $(u, V^{(u)})$ *is convex downward,*

$$V^{(1)} - V^{(2)} \geq V^{(2)} - V^{(3)} \geq \cdots \geq V^{(N-1)} - V^{(N)}. \tag{26}$$

Originally, (25) yields for stochastic matrices with entries satisfying

$$\lim_{T\to 0} T\log(P_{ij}(T)) = -V_{ij}.$$

2.2. Spectral analysis of $B(T)$. Let $E_{\min} = \{s+1, s+2, \ldots, s+m\}$, where $s := |S - E_{\min}|$ and $N = s+m$. Define $B(T)$ to be the substochastic matrix obtained from $P(T)$ by deleting the rows and columns corresponding to the states of $E_{\min}$. Complete $B(T)$ to a stochastic matrix by defining the states of $E_{\min}$ as absorbing

$$B'_{ij}(T) := 0, \quad \forall\, i \in E_{\min}, \quad j \neq i,$$
$$B'_{ij}(T) := P_{ij}(T(t)), \quad \forall\, i \in S - E_{\min}, \quad j \neq i.$$

All the coefficients of $B'(T)$ have exponential order. In particular, $B'_{ij}(T) = \exp(-\infty T)$ for all $i \in E_{\min}$, for all $T > 0$. Define the constants

$$V_{ij} := \begin{cases} [E(j) - E(i)]^+ & \text{if } q_{ij} > 0 \text{ and } i \notin E_{\min}, \\ +\infty & \text{if } i \in E_{\min} \text{ or if } q_{ij} = 0. \end{cases} \tag{27}$$

Let $-\lambda(T)$ be a nonzero eigenvalue of $B'(T) - Id$. By (25), there exist constants $V^{(u)}$ for $m \leq u \leq s+m$ such that

$$\lambda(T) \asymp \exp(-(V^{(u)} - V^{(u+1)})T^{-1}) \tag{28}$$

as $T \to 0$. States of $E_{\min}$ are absorbing. Assume $W \subset S$ is such that $|W| < m = |E_{\min}|$. Each $g \in G(W)$ then has one arrow $(i \to j)$ with $i \in E_{\min}$; as $E(i\to j) = \infty$, $E(g) = \infty$ for all $g \in G(W)$. Thus

$$V^{(u)} = \infty, \quad A^u = 0, \quad \text{for } 0 \leq u < m \tag{29}$$

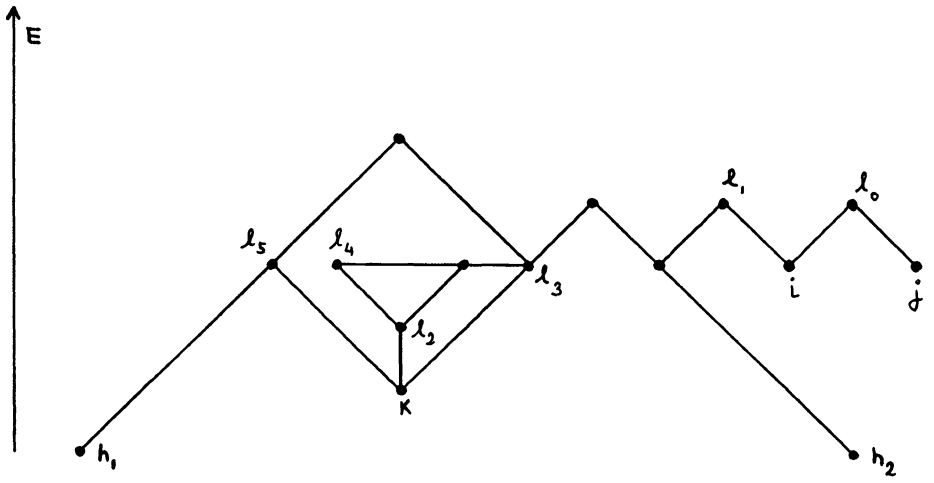


FIG. 1. $E_{\min} = \{h_1, h_2\}$, $E_{\text{locmin}} = \{k, i\}$, $d(k) = E(l_5) - E(k)$, $d(i) = E(l_1) - E(i)$, $d(j) = E(l_0) - E(j)$, $[k] = \{k, l_2, l_3, l_4\}$ *and* $[i] = \{i\}$, $[j] = \{j\}$, $[j] \subset [i]$.

(the m zero eigenvalues corresponding to states of $E_{\min}$).

DEFINITION 2.2.1. *Assume that* $q_{ij} > 0$ *if and only if* $q_{ji} > 0$. *Two states* i *and* j *are neighbors if and only if* $q_{ij} > 0$. *We denote by* E_{locmin} *the set of nodes* $i \in S - E_{\min}$ *for which* $E(j) \geq E(i)$ *for all* j *in the neighborhood of* i. *Two states* i *and* j *are said to communicate at height* h *if either* $i = j$ *and* $E(i) \leq h$, *or if there is a path*

$$i = x_0 \to x_1 \to \cdots \to x_n = j,$$

such that $E(x_u) \leq h$ *for* $0 \leq u \leq n$ *and* $q_{x_{u-1}x_u} > 0$ *for* $1 \leq u \leq n$. *We then have an equivalence relation* $\cong_h$ *on the set* $S_h := \{p \in S;\ E(p) \leq h\}$. *For* $i \in S - E_{\min}$, *let* $d(i)$ *be the depth of* i *defined by the relation*

$$E(i) + d(i) := \inf\{h > 0;\ \exists j \in S_h \text{ with } i \underset{h}{\cong} j \text{ and } E(j) < E(i)\} \tag{30}$$

($d(i) = 0$ *if* $i \notin E_{\text{locmin}}$), *and define the set*

$$[i] := \{p \in S;\ \exists h < E(i) + d(i) \text{ with } i \underset{h}{\cong} p\} \tag{31}$$

(see Fig. 1).

DEFINITION 2.2.2. *We say we are in the generic case if the following conditions are satisfied*: (1) $V^{(u)} - V^{(u+1)} > V^{(u+1)} - V^{(u+2)}$ *if* $V^{(u)} - V^{(u+1)} \neq 0$, *for* $m \leq u$; (2) *if* $i, j \in S - E_{\min}$ *are neighbors, then* $E(i) \neq E(j)$; (3) *for all* $(i, j) \in (S - E_{\min} - E_{\text{locmin}})^2$, $i \neq j$, $\lim_{T \to 0} P_{ii}(T) \neq \lim_{T \to 0} P_{jj}(T)$.

Remark 2.2.3. By (25), Definition 2.2.2(1) implies that there exists $t_0 \geq 0$ such that the eigenvalues $\lambda(t)$ of $B(T(t))$, which converge to 1, are all different for $t \geq t_0$. Definition 2.2.2(1) is hard to check without knowledge about the constants $V^{(u)}$; we will see in §2.3 that they are strongly related to the depths associated with the local minima, and therefore Definition 2.2.2(1) will be much easier to check. Assume the states are arranged in order of increasing values of $E(\cdot)$ and that we are in the generic case. Let $\alpha := |E_{\text{locmin}}|$. Considering the form of the transition matrix (1) and the hypothesis of Definition 2.2.2(2), we see that the limiting matrix $B(0^+)$ is triangular, and has α eigenvalues equal to 1 and $s - \alpha$ different

eigenvalues $1-\lambda<1$ (by Definition 2.2.2(3)). Let k be a node of $S-E_{\min}-E_{\text{locmin}}$; we associate with k the simple eigenvalue $-\lambda_{\sigma(k)}(T)$ of $B(T)-Id$, which converges to $-\lambda_{\sigma(k)}(0^+):=P_{kk}(0^+)-1\neq 0$. Define $\mathcal{N}(k)^-:=\{j\neq k;\ E(j)\leq E(k) \text{ and } q_{jk}>0\}$. As $1-P_{kk}(0^+)=\sum_{i\neq k}P_{ki}(0^+)$, Definition 2.2.2(3) becomes

$$\sum_{i\in\mathcal{N}(k)^-} q_{ki}\neq \sum_{i\in\mathcal{N}(j)^-} q_{ji}$$

for all $k\neq j$, $k,\ j\in S-E_{\min}-E_{\text{locmin}}$.

Hypothesis 2.2.4. Assume that Definition 2.2.2 holds and that

$$[k]\cap E_{\text{locmin}}=\{k\},\qquad \forall\, k\in E_{\text{locmin}}. \tag{32}$$

Note that the depths of the local minima are arranged in order of increasing value,

$$0\leq d_1\leq d_2\leq\cdots\leq d_{\sigma(k)-1}\leq d_{\sigma(k)}\leq d_{\sigma(k)+1}\leq\cdots\leq d_\alpha, \tag{33}$$

where $\alpha=|E_{\text{locmin}}|$, $d_{\sigma(k)}$ is the depth of the local minimum k ($d_{\sigma(k)}=d(k)$), and σ is a permutation of $\{1,\dots,s\}$. We set $d_{\sigma(j)}:=0$ for $j\notin E_{\text{locmin}}$.

2.3. Construction of optimal graphs. The constants $V^{(u)}$ appearing in (22) have been computed in [Chiang and Chow (1988a)]. For our purpose, we construct digraphs $g\in G^u$ for which $E(g)=V^{(u)}$. This constructive method will be useful in the remaining argument. Assume that (32) is satisfied. Let (S,Γ) be the graph with node set S and edge set Γ defined as follows: $(i,j)\in\Gamma$ if and only if $q_{ij}>0$. We identify, for each $i\in E_{\text{locmin}}$, all nodes $p\in[i]$ in a class $\langle i\rangle$. If $k\in S-\cup_{j\in E_{\text{locmin}}}[j]$, define $\langle k\rangle:=\{k\}$. Let $\langle k\rangle$ and $\langle k'\rangle$ be two classes of this quotient set $\tilde{S}$. Now define a new graph $(\tilde{S},\tilde{\Gamma})$ with node set $\tilde{S}$ and edge set $\tilde{\Gamma}$: if $k,\ k'\in S-\cup_{j\in E_{\text{locmin}}}[j]$ and $(k,k')\in\Gamma$, then $(k,k')=(\langle k\rangle,\langle k'\rangle)\in\tilde{\Gamma}$. If $k\in S-\cup_{j\in E_{\text{locmin}}}[j]$, $k'\in[i]$ for $i\in E_{\text{locmin}}$, and $(k,k')\in\Gamma$, then $(k,\langle i\rangle)\in\tilde{\Gamma}$. This quotient graph $(\tilde{S},\tilde{\Gamma})$ is in fact a multigraph. Consider the function $\tilde{E}$, which is defined on $\tilde{S}$,

$$\tilde{E}(\langle k\rangle):=\begin{cases} E(i)+d(i) & \text{if } k\in[i] \text{ for some } i\in E_{\text{locmin}},\\ E(k) & \text{otherwise,}\end{cases}$$

and extend $\tilde{E}(\cdot)$ to families of Wentzell graphs on $\tilde{S}$, as in Definition 2.1.1 (see Fig. 2).

Let $\tilde{\gamma}$ be a path in $(\tilde{S},\tilde{\Gamma})$. We say that $\tilde{\gamma}$ is monotonically decreasing (respectively, increasing) if $\tilde{E}$ is decreasing (respectively, increasing) along $\tilde{\gamma}$. The quotient graph $(\tilde{S},\tilde{\Gamma})$ contains no strict local minimum for $\tilde{E}$. Let $\langle j_1\rangle$ be an arbitrary node of $\tilde{S}$. Choose any neighbor $\langle l\rangle$ of $\langle j_1\rangle$, with $(\langle j_1\rangle,\langle l\rangle)\in\tilde{\Gamma}$ and $\tilde{E}(\langle l\rangle)\leq\tilde{E}(\langle j_1\rangle)$. Proceeding in this way, we obtain a nonincreasing path $\tilde{\gamma}_1$ that starts at $\langle j_1\rangle$ and ends in $E_{\min}$. Then choose an arbitrary element j_2 in the complement of $\tilde{S}-\tilde{\gamma}_1$. Working in the same way, we obtain a nonincreasing path $\tilde{\gamma}_2$ that starts at j_2 and stops the first time $\tilde{\gamma}_2$ meets $\tilde{\gamma}_1\cup E_{\min}$. We then obtain a directed forest $\tilde{g}\in\tilde{G}(E_{\min})$, with roots in $E_{\min}$, which is the union of directed trees, each of them being pointed in $E_{\min}$ (cf. Fig. 3).

Each path $\tilde{\gamma}\in\tilde{g}$ is monotonically decreasing; therefore $\tilde{E}(\tilde{g})=0$. Let $i\in E_{\text{locmin}}$. There exists a node p_2 in $\tilde{g}$ such that $(\langle i\rangle\to p_2)\in\tilde{g}$ with $\tilde{E}(\langle i\rangle)=\tilde{E}(p_2)$, and possibly a node p_1 with $(p_1\to\langle i\rangle)\in\tilde{g}$ with $\tilde{E}(p_1)\geq\tilde{E}(\langle i\rangle)$ (see Fig. 4).

Let q_1 and q_2 be two nodes of $[i]$, neighbors of p_1 and p_2, respectively. Choose any node j_1 of $[i]$. If $j_1\neq i$, (32) implies that there exists a node j_2 in the neighborhood of j_1 such that $E(j_2)<E(j_1)$. Then there exists a decreasing path γ_1 that starts at j_1 and ends at i, since by hypothesis, $[i]\cap E_{\text{locmin}}=\{i\}$. Then choose any node k_1 in $[i]\cap\gamma_1^c$. Working in the same

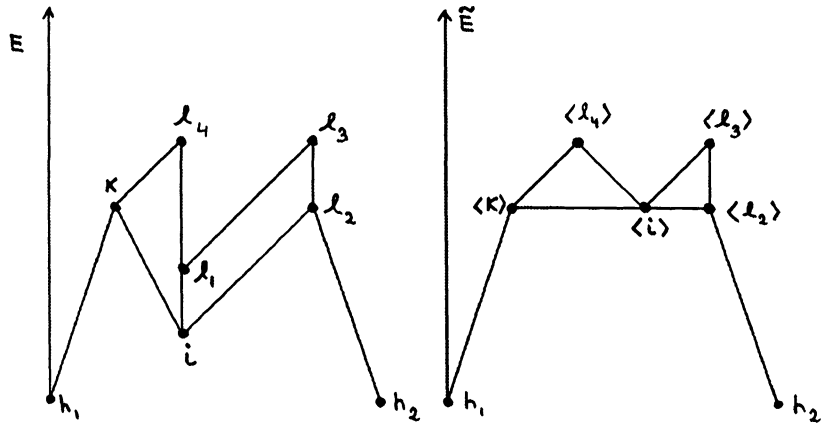


FIG. 2. *Graphical representation of* (S, Γ) *and* $(\tilde{S}, \tilde{\Gamma})$. $E_{\min} = \{h_1, h_2\}$. *The line between two nodes indicates they are neighbors; for example,* $(i, k) \in \Gamma$. $E_{\text{locmin}} = \{i\}$, $d(i) = E(k) - E(i)$, $[i] = \{i, l_5, l_1\}$, $\langle i \rangle = [i]$, $\langle l_1 \rangle = [i]$, $(\langle l_3 \rangle, \langle i \rangle) \in \tilde{\Gamma}$.

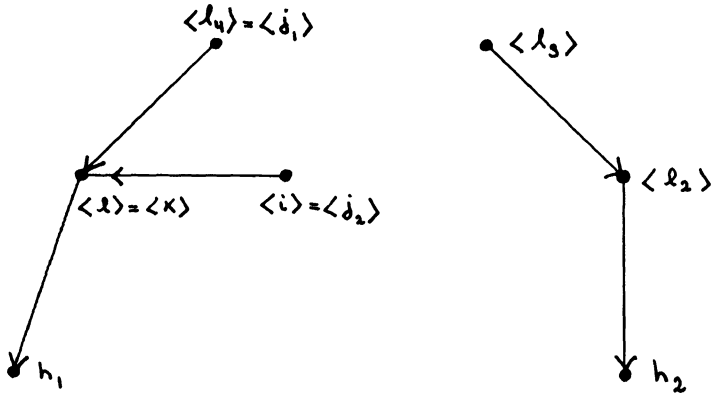


FIG. 3. *Decreasing paths in the quotient graph.*

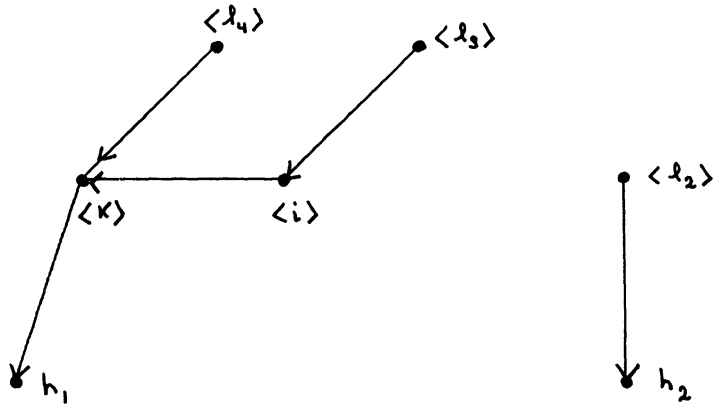


FIG. 4. $p_2 = \langle k \rangle, p_1 = \langle l_3 \rangle$.

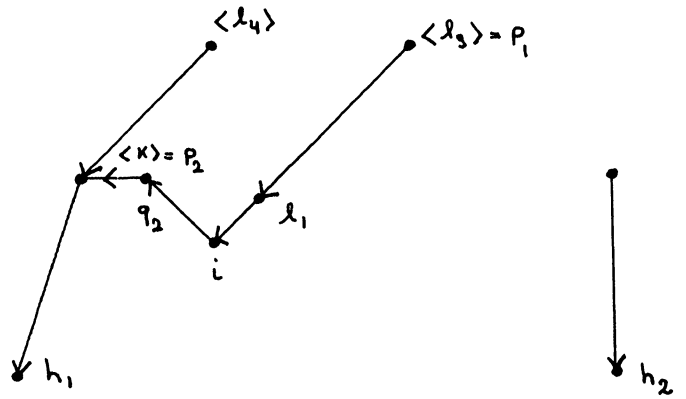


FIG. 5. *Example of a graph g of* $G(E_{\min})$ *for which* $E(g) = \sum_{i \in E_{\text{locmin}}} d(i)$. $q_1 = l_1$, $q_2 = l_5$.

way, we obtain a decreasing path γ_2 that starts at k_1 and ends the first time γ_2 meets γ_1. In this way, we obtain a maximal directed tree h_i in $[i]$. Let $p \in [i]$. The unique geodesic of h_i that starts at p and ends at i is then monotonically decreasing. We add to $\tilde{g}$: the monotonically increasing geodesic of h_i from i to q_2, all the other decreasing geodesics of h_i with terminal nodes i, and the arrow $(q_2 \to p_2)$, $(p_1 \to q_1)$. Then we obtain a graph $g \in G(E_{\min})$ with $E(g) = \sum_{i \in E_{\text{locmin}}} d(i)$ (see Fig. 5).

Let h be an arbitrary graph of $G(E_{\min})$ and let $i \in E_{\text{locmin}}$. Let γ be the path of h that starts at i and ends in $E_{\min}$. For each step $(k \to l) \in \gamma$, $E(k \to l) = [E(l) - E(k)]^+$ and the contribution of $E(\cdot)$ on this path in $[i]$ is at least equal to the depth of the local minimum, since γ comes out of $[i]$. Thus $E(h) \geq \sum_{i \in E_{\text{locmin}}} d(i)$ since, by (32), $[i] \cap E_{\text{locmin}} = \{i\}$ for all $i \in E_{\text{locmin}}$. It follows then that $E(g) = \min_{h \in G(E_{\min})} E(h)$ and $V^{(m)} = \sum_{i \in E_{\text{locmin}}} d(i)$, $m = |E_{\min}|$. To compute $V^{(m+1)}$, we must choose a node $p \in S - E_{\min}$ and evaluate E on $G(E_{\min} \cup \{p\})$. If $p \in S - \cup_{i \in E_{\text{locmin}}}[i]$, E is at least equal to $V^{(m)}$; otherwise it is easy to see that we must choose $p \in E_{\text{locmin}}$ with $d_{\sigma(p)} = d_\alpha$. We have $V^{(m+1)} = V^{(m)} - d_\alpha$. Proceeding in the same way, we obtain $V^{(u)} = +\infty$ for $1 \leq u < m$, $V^{(m)} - V^{(m+1)} = d_\alpha$, $V^{(m+1)} - V^{(m+2)} = d_{\alpha-1}, \ldots, V^{(\alpha+m-1)} - V^{(\alpha+m)} = d_1$, $V^{(u)} = 0$ for $u \geq m + \alpha$.

Remark 2.3.1. The constructions of this section can be generalized for matrices with entries of the form (19) where $V_{ij} = [E(j) - E(i)]^+$ for arbitrary graphs (S, Γ) [Mazza (1990)].

If (32) holds we have seen that

$$\{V^{(u)} - V^{(u+1)}\}_{m \leq u \leq |S|-1} \equiv \{d(k)\}_{k \in S - E_{\min}}.$$

Therefore, under (32), Definition 2.2.2(1) becomes $d_1 < d_2 < \cdots < d_\alpha$, i.e., all the local minima have different depths. It is not difficult to see that under Hypothesis 2.2.4,

$$\Re(1 - \lambda_u(T)) = (c_u + o(1))\exp(-(V^{(u-1)} - V^{(u)})T^{-1}), \qquad u = 2, \ldots, N, \tag{34}$$

for positive constants c_u.

2.4. Eigenvectors of $B(T)$. Let $-\lambda(T)$ be a nonzero eigenvalue of $B'(T) - Id$, which is solution of the equation (see (23) and (29))

$$A^m + A^{m+1}(-\lambda)^1 + \cdots + A^{m+s}(-\lambda)^s = 0. \tag{35}$$

Consider the *cofactor matrix* $\mathrm{Cof}(B(T) - (1 - \lambda(T))Id)$ associated with the matrix $B(T) - (1 - \lambda(T))Id$. We have

$$(B(T) - (1 - \lambda(T))Id) \circ \mathrm{Cof}(B(T) - (1 - \lambda(T))Id)^{\mathrm{Tr}} = 0, \tag{36}$$

where $\circ$ denotes the matrix product. Any row $C(T)$ of the cofactor matrix could be a candidate for eigenvector, but $C(T)$ might vanish. Note that even if $C(T) \neq 0$ for all $T > 0$, it might happen that $C(T) \to 0$ as $T \to 0$. Nevertheless, we establish a lemma that permits the computation of the cofactor matrix associated with the transition matrix of an arbitrary Markov chain.

DEFINITION 2.4.1. *Let $X(t)$ be a Markov chain on a finite set S with transition matrix P. $|S| := N$. Let Λ be a subset of S. Define the restricted chain on $S - \Lambda$ as the chain obtained by making the states of Λ absorbing. Define the family of digraphs for the restricted chain on $S - \Lambda$ and G_Λ, and set*

$$G_\Lambda^u := \{g \in G_\Lambda(U);\ U \subset S,\ |U| = u + |\Lambda|\}, \tag{37}$$

$$A_\Lambda^u := \sum_{g \in G_\Lambda^u} \pi(g), \qquad 0 \leq u \leq |S - \Lambda|. \tag{38}$$

LEMMA 2.4.2. *Consider a Markov chain on a finite set S with transition matrix $P \in \mathbb{R}^{N \times N}$ ($N := |S|$). Let $\hat{P}(\lambda) := P - (1 - \lambda)Id$. Let $C_{ij}(\lambda)$, $j \neq i$, be the term of* $\mathrm{Cof}(\hat{P}(\lambda))$ *associated with the* ith *row and the* jth *column. Then*

$$C_{ij}(\lambda) = (-1)^{N+1} \sum_{\gamma : j \to i} P(\gamma) P_\gamma(\lambda), \tag{39}$$

where the sum is taken over all simple paths γ (no-cycle) from j to i, $P(\gamma)$ is the probability associated with γ, $P_\gamma(\lambda)$ is the polynomial

$$A^0_{\mathcal{N}(\gamma)} + A^1_{\mathcal{N}(\gamma)}(-\lambda) + \cdots + A^{N-r}_{\mathcal{N}(\gamma)}(-\lambda)^{N-r}, \tag{40}$$

$r := |\mathcal{N}(\gamma)|$, and $\mathcal{N}(\gamma)$ is the "trace" of γ (nodes situated on γ).

Remark 2.4.3. If we set $\lambda = 0$ in (39), we obtain a lemma in [Freidlin and Wentzell (1984), p. 177] on the invariant measure of irreducible Markov chains. For the matrix B, $C_{ij}(\lambda)$ becomes

$$(-1)^{|S - E_{\min}|+1} \sum_{\substack{\gamma : j \to i \\ \mathcal{N}(\gamma) \subset S - E_{\min}}} P(\gamma)[A^0_{\mathcal{N}(\gamma) \cup E_{\min}} + A^1_{\mathcal{N}(\gamma) \cup E_{\min}}(-\lambda) + \cdots + A^{s-r}_{\mathcal{N}(\gamma) \cup E_{\min}}(-\lambda)^{s-r}],$$

where $r := |\mathcal{N}(\gamma)|$, $s = |S - E_{\min}|$. See §3.1 for the proof.

2.5. First hit of $E_{\min}$. Let γ be a simple path of (S, Γ) from j to k, for $k \in E_{\mathrm{locmin}}$ and $j \in S - E_{\min}$. Consider the quotient graph: $\hat{\gamma}$ is now a path of $(\tilde{S}, \tilde{\Gamma})$ from $\langle j \rangle$ to $\langle k \rangle$ (from $\langle k \rangle$ to $\langle k \rangle$ if $j \in [k]$).

DEFINITION 2.5.1. *If k is a node of E_{locmin}, define by k^- the set of elements $j \in S - E_{\min}$ such that each path $\hat{\gamma}$ from j to $\langle k \rangle$ has a strictly increasing arrow, and set $k^+ := (k^-)^c$. If k is a node of $S - E_{\min} - E_{\mathrm{locmin}}$, then k^- is defined as the set of nodes such that each path γ from j to k has a strictly increasing arrow and k^+ is defined as $(k^-)^c$. We say that the elements of $S - E_{\min}$ are well ordered if*

$$j \in k^+ \Rightarrow k \in j^-, \quad \forall k, \quad j \in S - E_{\min}. \tag{41}$$

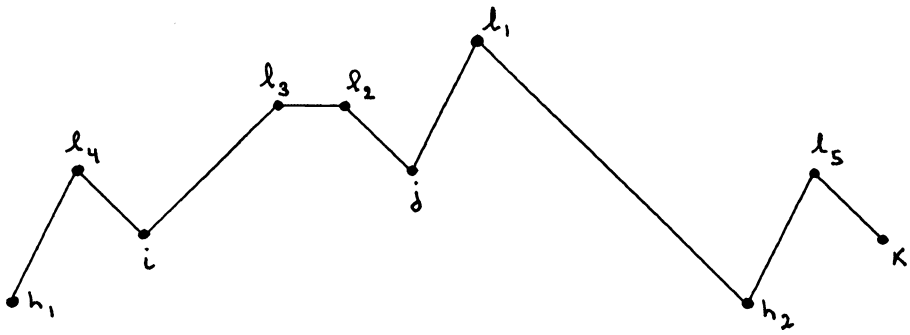


FIG. 6. *In this example,* $d(i) = E(l_4) - E(i)$, $d(j) = E(l_2) - E(j)$, *and* $d(k) = E(l_5) - E(k)$; $i \in j^-$, $j \in i^+$, $j \in k^-$, *and* $k \in j^-$.

Example 1. In Fig. 2, $i \in l_3^-$, $l_3 \in l_2^+$, $l_2 \in l_3^-$, and $l_3 \in i^+$.

Example 2. See Fig. 6.

The elements of $S - E_{\min}$ are not well ordered since $l_3 \in l_2^+$ and $l_2 \in l_3^+$. Roughly speaking, the concept of "well orderedness" permits us to avoid horizontal paths between nodes. If, for example, i and j are two nodes of $S - E_{\min} - E_{\text{locmin}}$, the assumption that the elements are well ordered implies that there is no path γ from i to j such that $E(\gamma) = 0$. As this notion is directly related to (S, Γ) and $E(\cdot)$, it is difficult to check in general. The elements are well ordered when no situation like $j \in k^+$, $k \in j^+$ occurs. If $j \in S - E_{\min} - E_{\text{locmin}}$ and $k \in E_{\text{locmin}}$, then obviously $k \in j^-$; if Definition 2.2.2(2) holds, and this will be the case in the next theorem, it is easy to see that we have only to concentrate on pairs of local minima. Suppose that Definition 2.2.2(2) holds and that $k \in j^+$, $j \in k^+$ for two nodes k and j of $S - E_{\min} - E_{\text{locmin}}$. Then there exists a monotonically decreasing path γ from j to k, and one from k to j; thus $E(j) \geq E(k)$ and $E(k) \geq E(j)$, so that $E(j) = E(k)$. If γ is given by the sequence $x_0 = j \to x_1 \to \cdots \to x_n = k$, then x_0 and x_1 are two neighbors such that $E(x_0) = E(x_1)$, which contradicts Definition 2.2.2(2). This shows that we must check the notion only for pairs of local minima.

Let $k \in S - E_{\min}$. Define $\Lambda(k)$ as the subset of $S - E_{\min}$ consisting of those j for which there exists a path γ from j to k for which $\mathcal{N}(\gamma) \cap E_{\min} = \emptyset$. Let

$$D(k) := \max_{j \in \Lambda(k)} d(j). \tag{42}$$

THEOREM 2.5.2. *Assume that Hypothesis* 2.2.4 *is satisfied and the elements of* $S - E_{\min}$ *are well ordered. Let* $T : \mathbb{R}^+ \to \mathbb{R}^+$ *be a differentiable function such that* $T(t) \to 0$ *as* $t \to \infty$ *and*

$$\int_0^\infty \left| \frac{d}{dt} \exp(-KT^{-1}(t)) \right| dt < +\infty \tag{43}$$

for every nonnegative constant $K \in \mathbb{R}$. *Then, if* t_0 *is large enough,*

$$P(\tau \geq t | X(t_0) = k) = (p_k + o(1)) \exp\left(-b_k \int_{t_0}^t \exp(-D(k)T(v)^{-1}) dv \right), \tag{44}$$

as $t \to \infty$, *for constants* $p_k \geq 0$ *and* $b_k > 0$. *Thus, the first hitting time* τ *is almost surely finite if*

$$\int_0^\infty \exp(-D(k)T(v)^{-1}) dv = +\infty. \tag{45}$$

Remark 2.5.3. Theorem 2.5.2 generalizes to the case where condition $[k] \cap E_{\text{locmin}} = \{k\}$ of (32) is replaced by $\{x \in E_{\text{locmin}};\ x \in [k] \text{ and } E(x) = E(k)\} = \{k\}$ for all $k \in E_{\text{locmin}}$, so that a given set $[k]$ might contain other elements of E_{locmin} [Mazza (1990)].

Let us specify the role of the hypotheses. The main idea consists of using the cofactor expansion (39) to put the differential system (10) in Levinson form (17). It can be proved (see, e.g., [Groebner (1956)]) that $\operatorname{rank}(\operatorname{Cof}(B - (1-\lambda)Id)) \leq 1$ for all $\lambda \in \operatorname{spec}(Id - B)$, and that $\operatorname{Cof}(B - (1-\lambda)Id) \equiv 0$ if λ has multiplicity greater than one. As explained in Remarks 2.2.3 and 2.3.1, Hypothesis 2.2.4 implies that the spectrum of $B(T(t))$ is simple for all $t \geq t_0$. This permits a direct use of the cofactor expansion for diagonalization of system (10). As $\lambda_{\sigma(k)}(T(t)) \to 0$ as $t \to \infty$ for all $k \in E_{\text{locmin}}$, the spectrum is asymptotically degenerate. As we will see in the proof of Theorem 2.5.2, the assumption that the elements of $S - E_{\min}$ are well ordered permits keeping the rank of the eigenvector matrix $W(T(t)) = (w_\lambda(T(t)))_{\lambda \in \operatorname{spec}(B(t))}$ unaltered in the limit $t \to \infty$. This is quite restrictive, but as we mentioned in §1, the method could be applied to various situations. Concerning the problem of relaxing the hypotheses, it might be possible to use the cofactor expansion for the blocks in the Jordan decomposition of $B(T)$, and to relax them in this way.

3. Proofs.

3.1. Proof of Lemma 2.4.2. Let $A(i) := \{1, 2, \ldots, i-1, i+1, \ldots N\}$. Define $B(i,j)$ to be the set of all bijective maps from $A(i)$ to $A(j)$. Assume $i < j$. Let ψ be the map of $B(i,j)$ given by

$$\psi(k) = \begin{cases} k, & \text{if } 1 \leq k \leq i-1, \\ i, & \text{if } k = i+1, \\ k-1, & \text{if } i+2 \leq k \leq j, \\ k, & \text{if } j < k < N. \end{cases}$$

$C_{ij}(\lambda)$ in (39) then becomes

$$(-1)^{i+j} \sum_{\pi \in B(i,i)} \operatorname{sign}(\pi) \hat{P}_{1\psi(\pi(1))} \hat{P}_{2\psi(\pi(2))} \cdots \hat{P}_{(i-1)\psi(\pi(i-1))} \cdot \hat{P}_{(i+1)\psi(\pi(i+1))} \cdots \hat{P}_{j\psi(\pi(j))} \cdots \hat{P}_{N\psi(\pi(N))}.$$

Let ψ' be the map of $B(j,j)$ given by

$$\psi'(k) = \begin{cases} k, & \text{if } 1 \leq k \leq i-1, \\ k+1, & \text{if } i \leq k \leq j-2, \\ i, & \text{if } k = j-1, \\ k, & \text{if } j < k \leq N. \end{cases}$$

Note that

$$\psi' \circ \psi(k) = \begin{cases} k, & \text{if } k \neq j, \\ i, & \text{if } k = j. \end{cases} \tag{46}$$

Then

$$-C_{ij} = \sum_{\pi \in B(i,i)} \operatorname{sign}(\pi) \prod_{k \in A_i} \hat{P}_{k,\psi' \circ \psi \circ \pi(k)}.$$

Consider, for π the cycle in j,

$$j \to \pi(j) \to \pi^2(j) \to \cdots \to \pi^{l-1}(j) \to \pi^l(j) = j,$$

which becomes, by (46),

$$j \to (\psi' \circ \psi \circ \psi)(j) = \pi(j), \quad \text{since } \pi(j) \neq j,$$

$$\pi(j) \to (\psi' \circ \psi \circ \pi)(\pi(j)) = \pi^2(j), \quad \text{since } \pi^2(j) \neq j,$$

$$\cdots$$

$$\pi^{l-1} \to (\psi' \circ \psi \circ \pi)(\pi^{l-1}) = i.$$

The cycle is interpreted as a path from j to i in S, denoted by γ_π, with trace $\mathcal{N}(\gamma_\pi)$. If $P(\gamma_\pi)$ is the probability associated with γ_π, then

$$-C_{ij} = \sum_{\pi \in B(i,i)} \operatorname{sign}(\pi) P(\gamma_\pi) \prod_{k \in A(i), k \notin \mathcal{N}(\gamma_\pi)} \hat{P}_{k,\psi' \circ \psi \circ \pi(k)}.$$

Next factorize the sum over all bijective maps $\pi \in B(i,i)$ so that the associated path corresponds to a given path γ,

$$\sum_{\gamma: j \to i} P(\gamma) \sum_{\pi: \gamma_\pi = \gamma} \operatorname{sign}(\pi) \prod_{k \in A(i), k \notin \mathcal{N}(\gamma_\pi)} \hat{P}_{k,\psi' \circ \psi \circ \pi(k)}.$$

Let π be a permutation such that $\gamma_\pi = \gamma$. The size of the associated cycle $\hat{\gamma}$ is $|\mathcal{N}(\gamma)| - 1$. π is then the composition of a cyclic permutation $\hat{\gamma}$ and of a permutation ν of $B(i,i)$, which leaves the nodes of $\mathcal{N}(\hat{\gamma})$ ($\mathcal{N}(\hat{\gamma})$ = orbit of $\hat{\gamma}$ in j) invariant. $-C_{ij}$ becomes

$$\sum_{\gamma: j \to i} P(\gamma) \operatorname{sign}(\hat{\gamma}) \sum_{\pi: \gamma_\pi = \gamma} \operatorname{sign}(\nu) \prod_{k \in A(i), k \notin \mathcal{N}(\gamma_\pi)} \hat{P}_{k,\nu(k)}.$$

As $\hat{\gamma}$ is cyclic, $\operatorname{sign}(\hat{\gamma}) = (-1)^{|\mathcal{N}(\gamma)|}$. Then

$$\delta(\gamma) := \sum_{\pi: \gamma_\pi = \gamma} \operatorname{sign}(\nu) \prod_{k \in A(i), k \notin \mathcal{N}(\gamma_\pi)} \hat{P}_{k,\nu(k)}$$

is the determinant of the matrix obtained by deleting the rows and the columns of $\hat{P}$ corresponding to states of $\mathcal{N}(\gamma)$. To apply Lemma 23 of [Wentzell (1972)], complete this matrix to a new matrix $\hat{P}'$ by defining the states of $\mathcal{N}(\gamma)$ as absorbing. If $k \in \mathcal{N}(\gamma)$, then $\hat{P}'_{kl} = 0$ if $l \neq k$ and $\hat{P}'_{kk} = 1 - (1 - \lambda) = \lambda$. If $k \notin \mathcal{N}(\gamma)$, then $\hat{P}'_{kl} = \hat{P}_{kl}$. This matrix is in fact the transition matrix of the restricted chain on $S - \mathcal{N}(\gamma)$ from which we subtract $(1 - \lambda)Id$. Then

$$(+\lambda)^{|\mathcal{N}(\gamma)|} \delta(\gamma) = (-1)^N [A^1(-\lambda) + \cdots + A^N(-\lambda)^N].$$

Since the states of $\mathcal{N}(\gamma)$ are absorbing, we have, by (29),

$$A^1 = \cdots = A^{|\mathcal{N}(\gamma)|-1} = 0.$$

Then

$$\delta(\gamma) = (-1)^{\mathcal{N}(\gamma)+N} P_\gamma(\lambda),$$

where $P_\gamma(\lambda)$ is the polynomial $A^0_{\mathcal{N}(\gamma)} + \cdots + A_{\mathcal{N}(\gamma)}(-\lambda)^{N-|\mathcal{N}(\gamma)|}$. $-C_{ij}$ therefore becomes

$$\sum_{\gamma: j \to i} P(\gamma)(-1)^{|\mathcal{N}(\gamma)|}(-1)^{|\mathcal{N}(\gamma)|+N} P_{\mathcal{N}(\gamma)}(\lambda). \qquad \square$$

Proof of Remark 2.4.3. For $B(T)$, consider $B'(T)$ and apply Lemma 2.4.2. We have

$$P'_\gamma(\lambda) = A^0_{\mathcal{N}(\gamma)} + \cdots + A^{s+m-r}_{\mathcal{N}(\gamma)}(-\lambda)^{s+m-r},$$

where $s = |S - E_{\min}|$, $m = |E_{\min}|$, $r = |\mathcal{N}(\gamma)|$. If $\mathcal{N}(\gamma) \cap E_{\min} \neq \emptyset$, then $P'(\gamma) = 0$. It is then sufficient to consider paths γ with $\mathcal{N}(\gamma) \subset S - E_{\min}$. As $E_{\min}$ is absorbing, we have, by (29), $A^0_{\mathcal{N}(\gamma)} = \cdots = A^{m-1}_{\mathcal{N}(\gamma)} = 0$; then the terms of the sum become

$$P(\gamma)(-\lambda)^m [A^m_{\mathcal{N}(\gamma)} + \cdots + A^{s-r}_{\mathcal{N}(\gamma)}(-\lambda)^{s-r}],$$

and the result follows from the equality $A^{m+u}_{\mathcal{N}(\gamma)} = A^u_{\mathcal{N}(\gamma) \cup E_{\min}}$. $\square$

3.2. Proof of Theorem 2.5.2. In the remaining argument, we assume that Hypothesis 2.2.4 is satisfied. Let γ be a simple path from j to k. All nodes of $\mathcal{N}(\gamma)$ are absorbing. Let g be a graph for the restricted chain on $S - E_{\min} - \mathcal{N}(\gamma)$ such that $g \in G_{\mathcal{N}(\gamma) \cup E_{\min}}(\mathcal{N}(\gamma) \cup E_{\min})$. Let g be a graph with node set $\mathcal{N}(g) \subset S$ and arrow set $\mathcal{A}(g)$. For any couple of graphs g_1 and g_2 such that $\mathcal{A}(g_1) \cap \mathcal{A}(g_2) = \emptyset$, consider the graph sum $g_1 \cup g_2$, which is defined as follows: $\mathcal{N}(g_1 \cup g_2) := \mathcal{N}(g_1) \cup \mathcal{N}(g_2)$ and $\mathcal{A}(g_1 \cup g_2) := \mathcal{A}(g_1) \cup \mathcal{A}(g_2)$ (note that g_1 and g_2 may have common nodes). Then $g \cup \gamma \in G_{E_{\min}}(\{k\} \cup E_{\min})$, and

$$E(g \cup \gamma) = E(g) + E(\gamma). \tag{47}$$

For any graph g define $c(g) := \prod_{(i \to j) \in g} q_{ij}$, where the probabilities q_{ij} are those defined in (1). In exponential form,

$$P(\gamma) = c(\gamma)\exp(-E(\gamma)T^{-1})$$

and

$$A^0_{\mathcal{N}(\gamma) \cup E_{\min}} = \sum_{g \in G_{\mathcal{N}(\gamma) \cup E_{\min}}(\mathcal{N}(\gamma) \cup E_{\min})} c(g)\exp(-E(g)T^{-1}). \tag{48}$$

LEMMA 3.2.1. *Let γ be a simple path from j to k. Then*

$$P(\gamma)A^n_{\mathcal{N}(\gamma) \cup E_{\min}} = \sum_U \sum_g c(g)\exp(-E(g)T^{-1}), \tag{49}$$

where the first sum is taken over all subsets U of $S - E_{\min} - \mathcal{N}(\gamma)$ with $|U| = n$, and the second over all graphs g of the family $G(U \cup \{k\} \cup E_{\min})$ containing γ.

Proof. As $\mathcal{N}(\gamma)$ is absorbing, we must consider only graph g of $G_{\mathcal{N}(\gamma) \cup E_{\min}}(U \cup \mathcal{N}(\gamma) \cup E_{\min})$ for $U \subset S - \mathcal{N}(\gamma) - E_{\min}$, $|U| = n$. The result is then a consequence of the relation

$$g \cup \gamma \in G_{E_{\min}}(\{k\} \cup U \cup E_{\min}). \tag{50}$$ $\square$

DEFINITION 3.2.2. *Define $G(k, j)$ to be the unique nonnegative real number such that*

$$C_{kj}(\lambda_{\sigma(k)})(T) \sim \exp(-G(k, j)T^{-1}) \quad \text{as } T \to 0.$$

$G(k,j)$ is well defined since all the terms (49) *appearing in the development* (39) *of* $C_{kj}(\lambda_{\sigma(k)})$ *have exponential form and since, by* (34),

$$\lambda_{\sigma(k)}(T) \sim \exp(-d_{\sigma(k)}T^{-1}) \quad \text{as } T \to 0.$$

LEMMA 3.2.3. *Let k be a local minimum and let $-\lambda_{\sigma(k)}(T)$ be the associated eigenvalue. Then*

$$C_{kk}(\lambda_{\sigma(k)}(T)) \neq 0 \qquad \forall T > 0,$$

$$G(k,k) = d_1 + d_2 + \cdots + d_{\sigma(k)-1} + d_{\sigma(k)}(\alpha - \sigma(k)),$$

$$G(k,j) \geq G(k,k) \qquad \forall j,$$

$$|C_{kj}(\lambda_{\sigma(k)}(T))| < c_{kj}|C_{kk}(\lambda_{\sigma(k)}(T))| \quad \forall j \in S - E_{\min}, \quad c_{kj} > 0,$$

$w_k^j(T) := C_{kj}(\lambda_{\sigma(k)})(C_{kk}(\lambda_{\sigma(k)}))^{-1}$ *converges as* $T \to 0\ \forall j \in S - E_{\min}, \forall k \in E_{\text{locmin}}$.

Proof. Consider the restricted chain on $S - E_{\min} - \{k\}$; $(-1)^s C_{kk}(\lambda_{\sigma(k)})$ becomes, by (39),

$$A^0_{\{k\}\cup E_{\min}} + A^1_{\{k\}\cup E_{\min}}(-\lambda_{\sigma(k)})^1 + \cdots + A^{s-1}_{\{k\}\cup E_{\min}}(-\lambda_{\sigma(k)})^{s-1},$$

where $s := |S - E_{\min}|$. By (34), the eigenvalues $1 - \lambda$ satisfy

$$\lambda(T) \sim \exp(-(V^{(u)} - V^{(u+1)})T^{-1}), \qquad 1 \leq u \leq s-1,$$

where the constants $V^{(u)}$, $1 \leq u \leq s-1$ are the constants associated with the restricted chain on $S - E_{\min} - \{k\}$ (see Definition 2.4.1). In the generic case, $d_{\sigma(k)} \neq V^{(u)} - V^{(u+1)}$ for all u since k is absorbing. $\lambda_{\sigma(k)}$ is then not solution of the equation $C_{kk}(\lambda) = 0$.

Under Hypothesis 2.2.4, it is easy to evaluate the coefficient $A^n_{\{k\}\cup E_{\min}}$. We must consider the depths of all local minima contained in $E_{\text{locmin}} - \{k\}$. ($\alpha := |E_{\text{locmin}}|$). Proceeding as in §2.3, we obtain

(51)

$$\begin{gathered}
A^0_{\{k\}\cup E_{\min}}|(\lambda_{\sigma(k)})|^0 \sim \exp(-(d_1 + \cdots + d_{\sigma(k)-1} + d_{\sigma(k)+1} + \cdots + d_\alpha)T^{-1}),\\
A^1_{\{k\}\cup E_{\min}}|(\lambda_{\sigma(k)})|^1 \sim \exp(-(d_1 + \cdots + d_{\sigma(k)-1} + d_{\sigma(k)+1} + \cdots + d_{\alpha-1} + d_{\sigma(k)})T^{-1})\\
\cdots\\
A^{\alpha-\sigma(k)}_{\{k\}\cup E_{\min}}|(\lambda_{\sigma(k)})|^{\alpha-\sigma(k)} \sim \exp(-(d_1 + \cdots + d_{\sigma(k)-1} + d_{\sigma(k)}(\alpha - \sigma(k)))T^{-1})\\
A^{\alpha-\sigma(k)+1}_{\{k\}\cup E_{\min}}|(\lambda_{\sigma(k)})|^{\alpha-\sigma(k)+1} \sim \exp(-(d_1 + \cdots + d_{\sigma(k)-2} + d_{\sigma(k)}(\alpha - \sigma(k) + 1))T^{-1})\\
\cdots\\
A^{s-1}_{\{k\}\cup E_{\min}}|(\lambda_{\sigma(k)})|^{s-1} \sim \exp(-(d_{\sigma(k)}(s-1))T^{-1}).
\end{gathered}$$

Thus

(52)

$$\max_{0\leq n\leq s-1} A^n_{\{k\}\cup E_{\min}}|(-\lambda_{\sigma(k)})^n| = \exp(-(d_1 + \cdots + d_{\sigma(k)-1} + d_{\sigma(k)}(\alpha - \sigma(k)))T^{-1}).$$

Consider the cofactor $C_{kj}(\lambda_{\sigma(k)})$. By Lemma 2.4.2 we must consider all simple paths γ from j to k. By Lemma 3.2.1 we also have

$$P(\gamma)A^n_{\mathcal{N}(\gamma)\cup E_{\min}} = \sum_U \sum_{g\in G} c(g)\exp(-E(g)T^{-1}),$$

where the first sum to be taken over all subsets $U \subset S - E_{\min} - \mathcal{N}(\gamma)$, $|U| = n$, and the second over all graphs $g \in G(U \cup \{k\} \cup E_{\min})$ containing γ. $P(\gamma)A^n_{\mathcal{N}(\gamma)\cup E_{\min}}$ is then dominated by $A^n_{\{k\}\cup E_{\min}}$. The order of $C_{kk}(\lambda_{\sigma(k)})$ has been obtained by considering each term $A^n_{\{k\}\cup E_{\min}}|(-\lambda_{\sigma(k)})^n|$. We have

$$(53)\quad (A^n_{\{k\}\cup E_{\min}}|(-\lambda_{\sigma(k)})^n|)(A^{\alpha-\sigma(k)}_{\{k\}\cup E_{\min}}|(-\lambda_{\sigma(k)})^{\alpha-\sigma(k)}|)^{-1} \underset{T\to 0}{\longrightarrow} 0, \quad \text{for } n \neq \alpha - \sigma(k).$$

Then we conclude that the order of $C_{kj}(\lambda_{\sigma(k)})$ is greater than the order of $C_{kk}(\lambda_{\sigma(k)})$. The required result follows from the form of the cofactors. □

LEMMA 3.2.4. *Let $k \in E_{\text{locmin}}$ and $j \in S - E_{\min}$. If $j \in k^-$, then*

$$G(k,j) > G(k,k), \qquad w^j_k(T) \underset{T\to 0}{\longrightarrow} 0.$$

DEFINITION 3.2.5. *Let Λ be a subset of S and $\Lambda \supset E_{\min}$. Suppose that the states of Λ are the only absorbing states. If $n < |\Lambda|$, $E(g) = +\infty$ for all g of $G(U)$ and each $U \subset S$ with $|U| = n$. Define the subset*

$$\tilde{\Lambda} := \{i \in E_{\text{locmin}};\ [i] \cap \Lambda \neq \emptyset\},$$

the new depths

$$d^1_{\sigma(i)} := \begin{cases} 0, & \text{if } i \notin \tilde{\Lambda}, \\ d_{\sigma(i)} + E(i) - \min\limits_{p\in\Lambda\cap[i]} E(p), & \text{otherwise}, \end{cases}$$

so that $0 \le d^1_{\sigma(i)} \le d_{\sigma(i)}$ and

$$d^2_{\sigma(i)} := d_{\sigma(i)} - d^1_{\sigma(i)}.$$

Working as in §2.3, we obtain

$$A^0_\Lambda \sim \exp\left(-\sum_{i\in E_{\text{locmin}}} d^2_{\sigma(i)}T^{-1}\right),$$

$$A^1_\Lambda \sim \exp\left(-\left(\sum_{i\in E_{\text{locmin}}} d^2_{\sigma(i)} - \max_i d^2_{\sigma(i)}\right)T^{-1}\right),$$

....

Proof of Lemma 3.2.4. Let γ be a simple path from j to k with $c(\gamma) > 0$. $F > 0$ is defined to be the total contribution of $\tilde{E}$ on $\hat{\gamma}$ (γ in the quotient). Consider Definition 3.2.5 with $\Lambda := \mathcal{N}(\gamma) \cup E_{\min}$ (see Fig. 7). We have

$$P(\gamma) \sim c(\gamma)\exp\left(-\left(F + H + \sum_{\tilde{\Lambda}-\{k\}} d^1_{\sigma(i)}\right)T^{-1}\right),$$

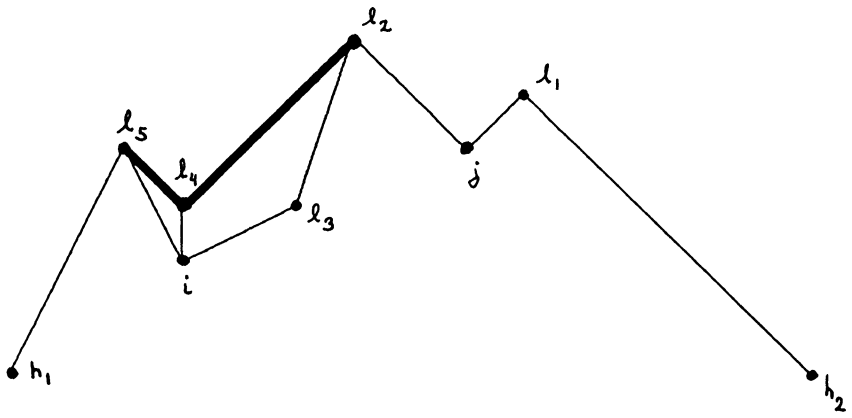


FIG. 7. $\Lambda = \{l_2, l_4, l_5\}$, $[i] = \{i, l_4, l_3\}$, $\tilde{\Lambda} = \{i\}$. $d(i) = E(l_5) - E(i)$, $d(j) = E(l_1) - E(j)$, $d^1_{\sigma(i)} = E(l_5) - E(l_4)$, $d^2_{\sigma(i)} = E(l_4) - E(i)$, $d^1_{\sigma(j)} = 0$, $d^2_{\sigma(j)} = d_{\sigma(j)} = d(j)$.

where H is a nonnegative constant. Consider the polynomial

$$A^0_{\mathcal{N}(\gamma)\cup E_{\min}} + A^1_{\mathcal{N}(\gamma)\cup E_{\min}}(-\lambda_{\sigma(k)})^1 + \cdots + A^{s-r}_{\mathcal{N}(\gamma)\cup E_{\min}}(-\lambda_{\sigma(k)})^{s-r} \tag{54}$$

with $r = |\mathcal{N}(\gamma)|$, and the products

$$P(\gamma)A^p_{\mathcal{N}(\gamma)\cup E_{\min}}|\lambda_{\sigma(k)}|^p, \quad 0 \le p \le s - r.$$

Let $p = 0$. Then

$$A^0_{\mathcal{N}(\gamma)\cup E_{\min}} \sim \exp\left(-\sum_{i\in E_{\text{locmin}}} d^2_{\sigma(i)}T^{-1}\right).$$

To evaluate

$$\max_p A^p_{\mathcal{N}(\gamma)\cup E_{\min}}|\lambda_{\sigma(k)}|^p,$$

arrange the terms of the sum $\sum d^2_{\sigma(i)}$ in order of increasing value and compare them to $d_{\sigma(k)}$. If $d^2_{\sigma(i)} \le d_{\sigma(k)}$ for all $i \in E_{\text{locmin}}$, then $\max A^p_{\mathcal{N}(\gamma)\cup E_{\min}}|\lambda_{\sigma(k)}|^p = A^0_{\mathcal{N}(\gamma)\cup E_{\min}}$ and the term associated with γ in the development of C_{kj} has order

$$\exp\left(-\left(F + H + \sum_{\tilde{\Lambda}-\{k\}} d^1_{\sigma(i)} + \sum_{i\in E_{\text{locmin}}} d^2_{\sigma(i)}\right)T^{-1}\right),$$

$$= \exp\left(-\left(F + H + \sum_{i\in E_{\text{locmin}}-\{k\}} d_{\sigma(i)}\right)T^{-1}\right).$$

Otherwise, let p_1 be element of E_{locmin} such that

$$d^2_{\sigma(p_1)} = \max_{E_{\text{locmin}}-\{k\}} d^2_{\sigma(i)} > d_{\sigma(k)}.$$

We have then

$$A^0_{\mathcal{N}(\gamma)\cup E_{\min}} < A^1_{\mathcal{N}(\gamma)\cup E_{\min}}|\lambda_{\sigma(k)}|^1 \sim \exp\left(-\left(\sum_{E_{\text{locmin}}-\{k\}-\{p_1\}} d^2_{\sigma(i)} + d_{\sigma(k)}\right)T^{-1}\right).$$

Working in the same way, we obtain a sequence $\{p_1,\dots,p_v\}$ such that

$$d^2_{\sigma(p_1)} \geq d^2_{\sigma(p_2)} \geq \cdots \geq d^2_{\sigma(P_{v-1})} > d^2_{\sigma(p_v)},$$

v being the first index for which $d^2_{\sigma(p_v)} \leq d_{\sigma(k)}$. $P(\gamma)\max_p A^p_{\mathcal{N}(\gamma)\cup E_{\min}}|\lambda_{\sigma(k)}|^p$ has order

$$\exp\Bigg(-\Bigg(F+H+\sum_{E_{\text{locmin}}-\{k\}} d^1_{\sigma(i)} + \sum_{E_{\text{locmin}}-\{k\}-\{p_1,\dots,p_{v-1}\}} d^2_{\sigma(i)} + (v-1)d_{\sigma(k)}\Bigg)T^{-1}\Bigg),$$

which is equal to

$$\exp\Bigg(-\Bigg(F+H+\sum_{\{p_1,\dots,p_{v-1}\}}(d^1_{\sigma(i)}+d_{\sigma(k)}) + \sum_{E_{\text{locmin}}-\{k\}-\{p_1,\dots,p_{v-1}\}}(d^2_{\sigma(i)}+d^1_{\sigma(i)})\Bigg)T^{-1}\Bigg).$$

As $d^2_{\sigma(i)} + d^1_{\sigma(i)} = d_{\sigma(i)}$, we obtain

$$\exp\Bigg(-\Bigg(F+H+(v-1)d_{\sigma(k)}+\sum_{i\in E_{\text{locmin}}-\{k\}-I} d_{\sigma(i)}\Bigg)T^{-1}\Bigg) \cdot \exp\Bigg(-\Bigg(\sum_{p_1,\dots,p_{v-1}} d^1_{\sigma(i)}\Bigg)T^{-1}\Bigg),$$

where I is a subset of $E_{\text{locmin}} - \{k\}$ that contains $v-1$ elements. The term associated with γ then has order

$$\exp\Bigg(-\Bigg(F+H+(v-1)d_{\sigma(k)}+\sum_{i\in E_{\text{locmin}}-\{k\}-I} d_{\sigma(i)}\Bigg)T^{-1}\Bigg) \cdot \exp\Bigg(-\Bigg(\sum_{p_1,\dots,p_{v-1}} d^1_{\sigma(i)}\Bigg)T^{-1}\Bigg),$$

and $G_{kj}(\gamma,\lambda_{\sigma(k)}) = F+H+(v-1)d_{\sigma(k)} + \sum_{i\in E_{\text{locmin}}-\{k\}-I} d_{\sigma(i)} + \sum_{p_1,\dots,p_{v-1}} d_{\sigma(i)}$. Since $G(k,k) = d_1+\cdots+d_{\sigma(k)-1}+d_{\sigma(k)}(\alpha-\sigma(k))$, the term associated with γ has order at least equal to $F+G(k,k) > G(k,k)$. This conclusion holds for all paths γ since $j \in k^-$. We conclude therefore that $G(k,j) > G(k,k)$. □

LEMMA 3.2.6. *Let $-\lambda_{\sigma(k)}(T)$ be the eigenvalue of $B(T)-Id$ converging to $P_{kk}(0^+)-1$, $k \in S - E_{\min} - E_{\text{locmin}}$. Then $C_{kk}(\lambda_{\sigma(k)}) \neq 0$ and $C_{kk}(\lambda_{\sigma(k)})$ converges to a nonzero limit. Moreover, if $j \in k^-$, then*

$$G(k,j) - G(k,k) > 0, \qquad w^j_k(T) := C_{kj}(\lambda_{\sigma(k)}(T))C_{kk}(\lambda_{\sigma(k)}(T))^{-1} \underset{T\to 0}{\longrightarrow} 0.$$

Consider the matrix $W := (w_k)_{k \in S - E_{\min}}$.

COROLLARY 3.2.7. *Suppose that the elements of E_{locmin} are well ordered. Then there exists $T_0 > 0$ such that*

$$W(T) \quad \text{is regular} \quad \forall\, T \leq T_0,$$
$$W(T) \to W(0^+), \quad \text{where the latter is regular.}$$

Proof of Lemma 3.2.6. If T is small enough, $-\lambda_{\sigma(k)}$ is simple. $C_{kk}(\lambda)(-\lambda)^{|E_{\min}|+1}$ is the characteristic polynomial associated with the restricted chain on $S - E_{\min} - \{k\}$, which has $|E_{\min}| + 1$ zero eigenvalues and $|S - E_{\min}| - 1$ nonzero eigenvalues, each asymptotically different from $-\lambda_{\sigma(k)}$ (by (34), and Definitions 2.2.3 and 2.3.1). It follows that $C_{kk}(\lambda_{\sigma(k)}) \neq 0$. Let $j \in S - E_{\min}$. We have $C_{kj}(\lambda_{\sigma(k)}) = \sum_{\gamma : j \to k} P(\gamma) P_\gamma(\lambda_{\sigma(k)})$. If $j \in k^-$, then $P(\gamma) \to 0$ as $T \to 0$ for all γ, and it follows that $G(k, j) > 0$. Moreover, as $C_{kk}(\lambda_{\sigma(k)})$ does not converge to 0, $G(k, k) = 0$ and $G(k, j) - G(k, k) > 0$. □

Proof of Corollary 3.2.7. By Definition 2.2.3, the vectors $w_k, k \in S - E_{\min} - E_{\text{locmin}}$ are independent for $T > 0$. In the limit $T \to 0$, the vectors $w_k(0^+)$ are the eigenvectors associated with the eigenvalues $-\lambda_{\sigma(k)}(0^+)$ of the matrix $B(0^+)$, which are all different. Thus they are independent. For $k \in E_{\text{locmin}}$, $\lambda_{\sigma(k)}(T) \to 0$. By Hypothesis 2.2.4, the elements of E_{locmin} are well ordered. Let $k, l \in E_{\text{locmin}}$ such that $l \in k^-$; then $w_k^k(T) \to 1$ and $w_l^k(T) \to 0$ by Lemma 3.2.4. It follows that the vectors $w_k(0^+)$ and $w_l(0^+)$ are independent. Corollary 3.2.7 follows by iterating this argument for all vectors w_k, $k \in E_{\text{locmin}}$. □

LEMMA 3.2.8. *Assume Hypothesis* 2.2.4 *is satisfied. Let $T : \mathbb{R}^+ \to \mathbb{R}^+$ be a differentiable function such that $T(t) \to 0$ as $t \to \infty$ and which verifies* (43),

$$\int_0^\infty \left| \frac{d}{dt} \exp(-KT^{-1}(t)) \right| dt < +\infty,$$

for every nonnegative constant $K \in \mathbb{R}$. Then

(a) $$\frac{d}{dt} \lambda_{\sigma(k)}(T(t)) \sim \lambda_{\sigma(k)}(T(t)) \frac{d}{dt} \left(\frac{-1}{T(t)} \right), \quad \forall\, k \in E_{\text{locmin}},$$

(b) $$\int_0^\infty \left| \frac{d}{dt} w_k^j \right| (t) dt < \infty, \qquad \forall\, k, j \in S - E_{\min}.$$

Proof. Assume first that $k \in S - E_{\min} - E_{\text{locmin}}$. We follow the method given in [Levinson (1948)]. Differentiating the cofactor $C_{kj}(\lambda_{\sigma(k)})$ as a function of t, we obtain an expression that is linear homogeneous in the entries of $(d/dt)B$ and $(d/dt)\lambda_{\sigma(k)}$. We must check the integrability of the terms involving the coefficients $|(d/dt)B_{il}|$. By definition, $B_{il}(T(t)) = q_{il} \exp(-E_{il} T(t)^{-1})$, where we assume that $q_{il} > 0$ (otherwise the result is obvious), and $E_{il} := [E(l) - E(i)]^+$. Thus

$$\left| \frac{d}{dt} B_{il}(T(t)) \right| = q_{il} E_{il} \left| \frac{d}{dt} (T(t)^{-1}) \right| \exp(-E_{il} T(t)^{-1}), \tag{55}$$

which is integrable by (43).

To see (b), we check the integrability of $|(d/dt)\lambda_{\sigma(k)}(T(t))|$. Let $M(\lambda) := B - (1 - \lambda) Id$. As $-\lambda_{\sigma(k)}$ is a characteristic root of $B - Id$, $\det(M(\lambda_{\sigma(k)})) \equiv 0$. As a function of t, $M(\lambda)(t)$ is a function of two functions of t, namely $M(T(t), \lambda(T(t)))$. Define $Q(T(t), \lambda(T(t))) := \det(M(T(t), \lambda(T(t))))$. Then we obtain

$$\frac{\partial Q}{\partial \lambda} \frac{d\lambda(T(t))}{dt} + \frac{\partial Q}{\partial T} \frac{dT}{dt} \equiv 0. \tag{56}$$

As $Q(T(t), \lambda(T(t))) = \det(B(T(t)) - Id + \lambda(T(t)))$, we have

$$\frac{\partial Q}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left(\prod_{u=1}^{s}(\lambda_u - \lambda)\right), \tag{57}$$

where $\{\lambda_u\}_{u=1,\dots,s} = \operatorname{spec}(Id - B)$. Thus

$$\frac{\partial Q}{\partial \lambda} = \sum_{u=1}^{s}\prod_{v\neq u}(\lambda_v - \lambda). \tag{58}$$

As $\lambda = \lambda_{\sigma(k)} \in \operatorname{spec}(Id - B)$,

$$\frac{\partial Q}{\partial \lambda}\bigg|_{\lambda=\lambda_{\sigma(k)}} = \prod_{v\neq\sigma(k)}(\lambda_v - \lambda_{\sigma(k)}). \tag{59}$$

As $\lambda_{\sigma(k)}(T(t))$ is simple for all t (see Remark 2.2.3), $\partial Q/\partial\lambda \neq 0$ for all t. Moreover, Remark 2.2.3 implies that $\lambda_{\sigma(k)}(0^+) \neq 0$ and $\lambda_{\sigma(k)}(0^+) \neq \lambda_{\sigma(k')}(0^+)$ for all $k \neq k'$. We conclude therefore that $(\partial/\partial\lambda)\det(M(\lambda_{\sigma(k)}))$ converges to a nonzero limit as $t \to \infty$. By (56) and (57) we can write

$$\frac{d}{dt}\lambda_{\sigma(k)} = -\frac{\partial}{\partial t}\det(M(\lambda_{\sigma(k)}))\left(\prod_{u\neq\sigma(k)}(\lambda_u - \lambda_{\sigma(k)})\right)^{-1}. \tag{60}$$

The integrability of $|(d/dt)\lambda_{\sigma(k)}(T(t))|$ follows then from the integrability of $|(\partial/\partial t)\det(M(\lambda_{\sigma(k)}))|$.

Let

$$\prod_{i:\pi(i)\neq i}\exp(-E_{i\pi(i)}T^{-1})\prod_{i:\pi(i)=i}(P_{ii}(T(t)) - \lambda(T(t))) \tag{61}$$

be a typical term of Q, where π is a permutation. Its partial derivative with respect to T has the form

$$\lambda(T(t))^n \exp(-KT^{-1}(t)), \tag{62}$$

where $n \in \mathbb{N}$ and $K > 0$. The integrability of $|(\partial Q/\partial T)(\partial T/\partial t)|$ then follows from the integrability of the terms (62), which are integrable (recall the integrability of $|(d/dt)B_{il}(T(t))|$). Let $k \in E_{\text{locmin}}$. We first prove assertion (a). Consider the denominator of (60). By hypothesis, $\lambda_{\sigma(k)}$ is simple for all t, so that the product is different from 0 for all t. Let $\lambda_{\sigma(j)}$ be the eigenvalue associated with the local minimum j, which verifies $\lambda_{\sigma(j)} \sim \exp(-d_{\sigma(j)}T(t)^{-1})$. Let $\hat{\lambda}_{\sigma(j)}$ be the eigenvalue associated with j for the restricted chain on $S - E_{\min} - \{k\}$. Proceeding as in §2.3, we obtain $\lambda_{\sigma(j)} \sim \hat{\lambda}_{\sigma(j)}$. If $j \notin E_{\text{locmin}}$, the equivalence $\lambda_{\sigma(j)} \sim \hat{\lambda}_{\sigma(j)}$ follows from the triangular form of the limiting transition matrix and by hypothesis. Thus

$$\prod_{u\neq\sigma(k)}(\lambda_{\sigma(k)} - \lambda_u) \sim C_{kk}(\lambda_{\sigma(k)}).$$

Consider now the numerator of (60), which is equal to

$$\sum_{n=0}^{s}\frac{\partial A^n_{E_{\min}}}{\partial t}(-\lambda_{\sigma(k)})^n, \tag{63}$$

with

$$A^n_{E_{\min}} = \sum_{g \in G^n_{E_{\min}}} c(g)\exp(-E(g)T(t)^{-1}).$$

Then we have

$$\frac{d}{dt}A^n_{E_{\min}} = \frac{d}{dt}(-T(t)^{-1}) \sum_{g \in G^n_{E_{\min}}} c(g)E(g)\exp(-E(g)T(t)^{-1}).$$

Let us show that the last sum is equivalent to $C_{kk}(\lambda_{\sigma(k)})\lambda_{\sigma(k)}(d/dt)(-T(t)^{-1})$. For that purpose, let $D_n := \sum_{g \in G^n_{E_{\min}}} c(g)E(g)\exp(-E(g)T(t)^{-1})$. Assume first that $n \leq \alpha$. Working as in the proof of Lemma 3.2.3, we obtain

$$D_n \sim \exp(-(d_1 + \cdots + d_{\alpha-n})T^{-1}).$$

If $n > \alpha$, we have $D_n \sim$ const. Then

$$D_n\lambda^n_{\sigma(k)} \sim \exp(-(d_1 + \cdots + d_{\alpha-n} + nd_{\sigma(k)})T^{-1}).$$

In any case, the dominating term of (63) is

$$\exp(-(d_1 + \cdots + d_{\sigma(k)-1} + (\alpha - \sigma(k) + 1)d_{\sigma(k)})T^{-1}),$$

so that Lemma 3.2.3 implies (a).

We now check the integrability of $(d/dt)w^j_k$ for $k \in E_{\text{locmin}}$. We have

$$C_{kk}(\lambda_{\sigma(k)}) = \sum_{n=0}^{s-1} A^n_{E_{\min}\cup\{k\}}(-\lambda_{\sigma(k)})^n$$

and

$$A^n_{E_{\min}\cup\{k\}}(\lambda_{\sigma(k)})^n \sim \exp(-C_nT^{-1}),$$

where C_n is a positive constant. We have seen in the proof of Lemma 3.2.3 that if γ is a simple path from j to k, then

$$P(\gamma)A^n_{E_{\min}\cup\mathcal{N}(\gamma)}(\lambda_{\sigma(k)})^n \sim \exp(-B_nT^{-1}), \qquad B_n > 0,$$

with $\min_n C_n \geq \min_n B_n$. Now normalize w^j_k:

$$w^j_k := (C_{kj}(\lambda_{\sigma(k)})\exp(C_*T^{-1}))(C_{kk}(\lambda_{\sigma(k)})\exp(C_*T^{-1}))^{-1} =: \tilde{C}_{kj}\tilde{C}^{-1}_{kk},$$

where $C_* := \min_n C_n$. $\tilde{C}_{kk}$ converges to a nonzero limit, so that

$$\frac{d}{dt}w^j_k \sim \left(\frac{d}{dt}\tilde{C}_{kj}\right)\tilde{C}_{kk} - \tilde{C}_{kj}\left(\frac{d}{dt}\tilde{C}_{kk}\right).$$

It remains to see that $|(d/dt)\tilde{C}_{kj}|$ and $|(d/dt)\tilde{C}_{kk}|$ are integrable. We have

$$\tilde{C}_{kk} = \sum_{n:C_n>C_*} \exp(C_*T^{-1})A^n_{E_{\min}\cup\{k\}}(-\lambda_{\sigma(k)})^n + \text{ const},$$

$$A^n_{E_{\min}\cup\{k\}} = \sum_{g\in G^n_{E_{\min}}} c(g)\exp(-E(g)T^{-1}).$$

However, by (a),

$$\frac{d}{dt}(\exp(-(E(g)-C_*)T^{-1})\lambda^n_{\sigma(k)}) \sim \frac{d}{dt}(-T^{-1})\exp(-(E(g)-C_*)T^{-1})\lambda^n_{\sigma(k)},$$

so that

$$\left|\frac{d}{dt}\tilde{C}_{kk}\right| \sim \frac{d}{dt}(-T^{-1})\exp(C_*T^{-1})\sum_{n:C_n>C_*}\exp(-C_nT^{-1}),$$

which is integrable (recall the integrability of $|(d/dt)B_{il}(T(t))|$). The integrability of $|(d/dt)\tilde{C}_{kj}|$ is obtained in the same way. □

Proof of Theorem 2.5.2. Assume, without loss of generality, that $\Lambda(k) = S - E_{\min}$. Let $Y(t',t)$ be the solution of the system

$$\frac{\partial Y}{\partial t'} = (-B(t') + Id)Y, \qquad 0 \le t' \le t, \tag{64}$$
$$Y(t,t) = (1,\dots,1)^{\mathrm{Tr}}.$$

Let $W(T(t'))$ be the regular matrix given in Corollary 3.2.7, where we assume $t_0 = 0$. Define $x(t') := W(T(t'))^{-1}Y(t')$. Then we have the new system

$$\frac{\partial x}{\partial t'} = R(t')x(t') + \mathrm{Diag}(\lambda_{\sigma(k)}(t'))Y(t'),$$

where $R(t') = -W^{-1}(d/dt')W(T(t'))$ and $\lambda_{\sigma(k)}(t')$ are the eigenvalues of $B - Id$ associated with the elements k of $S - E_{\min}$. System (64) therefore becomes

$$x(t,t) = W^{-1}(T(t))(1,\dots,1)^{\mathrm{Tr}}.$$

This system has the Levinson form (17). $|R(t)|$ is integrable by Corollary 3.2.7 and Lemma 3.2.8; it is not difficult to see that conditions (14) and (15) of Levinson's theorem are satisfied, so that we know the system has $s := |S - E_{\min}|$ solutions $x_u(t')$ with

$$x_u(t') = (e_u + o(1))\exp\left(\int_0^{t'}\lambda_u(v)dv\right).$$

Let x be an arbitrary solution; then there exist constants $f_1,\dots,f_s \in \mathbb{R}$ such that

$$x(t') = \sum_u f_u(e_u + o(1))\exp\left(\int_0^{t'}\lambda_u(v)dv\right).$$

Set $t' = t$; then we have the equation

$$x(t) = W^{-1}(T(t))(1,\dots,1)^{\mathrm{Tr}},$$

which has a solution since $W(T(t))$ is inversible for all t. The matrix $Z(t) := (e_1 + o(1), \dots, e_s + o(1))$ is regular for all t, so that

$$\left(f_u\exp\left(\int_0^t\lambda_u(v)dv\right)\right)_u = Z^{-1}(t)W^{-1}(T(t))(1,\dots,1)^{\mathrm{Tr}},$$

and then

$$f_u(t) = \exp\left(-\int_0^t \lambda_u(v)dv\right) \langle Z^{-1}(t)W^{-1}(T(t))(1,\ldots,1)^{\mathrm{Tr}},\, e_u\rangle,$$

where $\langle\cdot,\cdot\rangle$ is the scalar product. Then, if $t' = 0$,

$$Y(0,t) = \sum_u W(T(0))Z_u(0) \exp\left(-\int_0^t \lambda_u(v)dv\right) \langle Z^{-1}(t)W^{-1}(T(t))(1,\ldots,1)^{\mathrm{Tr}},\, e_u\rangle.$$

As $t \to \infty$, $Z^{-1}(t)$ converges to the identity matrix and $W^{-1}(T(t))$ to $W^{-1}(0^+)$, so that $Y(0,t)$ follows the behavior of its dominating term

$$\exp\left(-\min_u \int_0^t \lambda_u(v)dv\right) \sim \exp\left(-b_k \int_0^t \exp(-D(k)T^{-1}(v))dv\right),$$

where b_k is one of the constants c_u appearing in (34), and is associated with a node realizing $D(k)$. □

Acknowledgments. I thank my Ph.D. advisor Professor A. Antille for his encouragement and advice. Many thanks to my friends J. P. Berrut, J. P. Gabriel, P. Milasevic, and C. Schwab for valuable remarks and discussions. I am also grateful to the referees for their helpful comments and suggestions that improved the substance as well as the presentation of the article.

REFERENCES

T. S. Chiang and Y. Chow, *On eigenvalues and annealing rates*, Math. Oper. Res., 13 (1988a), pp. 508–511.

T. S. Chiang and Y. Chow, *On the convergence rate of annealing processes*, SIAM J. Control Optim., 26 (1988b), pp. 1455–1470.

D. P. Connors and P. R. Kumar, *Simulated annealing type Markov chains and their order balance equations*, SIAM J. Control Optim., 27 (1989), pp. 1440–1462.

M. S. P. Eastham, *The asymptotic solutions of linear differential systems: Applications of the Levinson Theorem*, Oxford Science Publications, London Mathematical Society Monographs 4, 1989.

R. Feller, *An Introduction to Probability Theory*, Vol. 1, John Wiley, New York, 1950.

M. I. Freidlin and A. D. Wentzell, *Random perturbations of Dynamical Systems*, Springer, Berlin, 1984.

W. Groebner, *Matrizenrechnung*, Verlag von Oldenburg, Muenchen, 1956.

B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.

R. Holley and D. Stroock, *Simulated annealing via Sobolev inequalities*, Comm. Math. Phys., 115 (1988), pp. 553–569.

G. F. Lawler and A. D. Sokal, *Bounds on the L^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger's inequality*, Trans. Amer. Math. Soc., 309 (1988), pp. 557–580.

N. Levinson, *The asymptotic nature of solutions of linear systems of differential equations*, Duke, 15 (1948), pp. 111–126.

C. Mazza, *Algorithmes de recuit simulé et graphes de Ventcel*, Ph.D. thesis, Department of Mathematics, University of Fribourg, Fribourg, Switzerland, 1990.

A. D. Wentzell, *On the asymptotics of eigenvalues of matrices with elements of order* $\exp(-V_{ij}/(2\varepsilon^2))$, Soviet Math. Dokl., 13 (1972), pp. 65–68.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1289–1296, September 1994

005

ON THE SOLUTIONS OF A CLASS OF CONTINUOUS LINEAR PROGRAMS*

EDWARD J. ANDERSON† AND ANDREW B. PHILPOTT‡

Abstract. This paper discusses the form of solutions for a class of continuous linear programs called separated continuous linear programs. It is shown that under certain assumptions on the problem data the optimal solutions can be taken to be piecewise analytic functions. This yields a strong duality result as a corollary.

Key words. continuous linear program, duality, discrete approximation

AMS subject classifications. 49A55, 49B36, 49D99, 90C48

1. Introduction.

In 1953 Bellman [7] introduced a class of optimization problems which he called *bottleneck problems.* These problems are generally referred to in the literature as continuous linear programs since they can be formulated as linear programs having variables which are functions of time as follows:

$$\begin{aligned} \text{CLP} \qquad & \text{maximize} \int_0^T c(t)^T x(t)dt \\ & \text{subject to } B(t)x(t) + \int_0^t K(t,s)x(s)ds \leq b(t), \\ & \qquad x(t) \geq 0, \qquad t \in [0,T]. \end{aligned}$$

The usual approach to solving CLP is to form an approximation by discretizing the time interval $[0,T]$ (see, e.g., Buie and Abrham [9]). A number of authors (see Lehman [12], Drews [10], Hartberger [11], Segers [17], Perold [13], and Anstreicher [6]) have attempted to generalize the simplex method to solve instances of CLP without discretizing. Such a generalization requires the analogues of basic feasible solutions and a pivot operation to be defined. The most comprehensive treatment of these issues is represented by the work of Perold [13] who derives a characterization of the extreme points of the feasible region of CLP in the case where the matrices $B(\cdot)$ and $K(\cdot,\cdot)$ are constant and the components of $x(\cdot)$ are so-called right analytic functions. (A function $g : [0,T] \to R$ is called *right analytic* if for each $t \in [0,T)$ there is some $\epsilon > 0$ and an analytic function $h_t(\cdot) : (t-\epsilon, t+\epsilon) \to R$ such that $g(s) = h_t(s), s \in [t, t+\epsilon)$.) In general, however, there is no guarantee that there will be an extreme-point optimal solution to CLP that is right analytic, even when the matrices B and K are constant and the functions appearing in the problem formulation are all analytic. Observe that a right analytic function may have an infinite number of discontinuities.

We are tempted to conjecture (see [13, p. 111]) that in many instances of CLP there will be optimal solutions that are analytic with a finite number of jump discontinuities. We call such functions *piecewise analytic* and define them formally as follows. First a set $\{t_0, t_1, \ldots, t_m\}$ is said to be a *partition* of $[0,T]$ if

$$0 = t_0 < t_1 < \cdots < t_m = T.$$

* Received by the editors March 9, 1992; accepted for publication (in revised form) January 29, 1993.

† Judge Institute of Management Studies, Fitzwilliam House, 32 Trumpington Street, Cambridge CB2 1QY, England.

‡ Department of Engineering Science, University of Auckland, Auckland, New Zealand.

Now, we call a function $g(\cdot)$ *piecewise analytic* (*piecewise constant*) on $[0,T]$ if there is a partition $\{t_0, t_1, \ldots, t_m\}$, $\epsilon > 0$, and analytic (constant) functions $h_i(\cdot) : (t_{i-1} - \epsilon, t_i + \epsilon) \to R$ such that $g(s) = h_i(s)$, $s \in [t_{i-1}, t_i)$ for $i = 1, 2, \ldots, m$.

The purpose of this paper is to show that for a substantial class of continuous linear programs there is a piecewise analytic optimal solution. The class of problems we consider are called *separated continuous linear programs*, which have the following form:

$$\text{SCLP} \qquad \text{minimize} \int_0^T c(t)^T x(t)dt$$

$$\text{subject to} \int_0^t Gx(s)ds + y(t) = a(t), \tag{1}$$

$$Hx(t) + z(t) = b(t), \tag{2}$$

$$x(t), y(t), z(t) \geq 0, \qquad t \in [0,T]. \tag{3}$$

Here $x(\cdot)$, $z(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are bounded measurable functions; $y(\cdot)$ and $a(\cdot)$ are continuous functions. The dimensions of $x(\cdot)$, $y(\cdot)$, and $z(\cdot)$ are n_1, n_2, and n_3, respectively. We let $\hat{x}^T(\cdot) = (x^T(\cdot), y^T(\cdot), z^T(\cdot))$.

This problem was first studied by Anderson [1] in the context of job-shop scheduling, and is discussed in detail in [2] and [3]. By differentiating the constraints (1), SCLP can be shown to be equivalent to a special type of linear optimal control problem with state variable inequality constraints. These constraints pose difficulties when applying classical optimal control techniques, since in many cases the control ceases to be of bang-bang type. However, with an assumption that the feasible region of SCLP is bounded, the optimal solution can be shown to be an extreme point of this set. The reader is referred to [3] for this result and a characterization of the extreme points of SCLP as appropriately defined basic feasible solutions.

The assumption that the feasible region of SCLP is bounded also allows the derivation of some results defining the form of optimal solutions of SCLP. In this respect, it is shown in [3] that if the components of $b(\cdot)$ are constant and the components of $a(\cdot)$ and $c(\cdot)$ are affine functions then SCLP has an optimal solution $\hat{x}(\cdot)$ that is piecewise affine. In recent work, Pullan [16] has shown under the same boundedness assumption that if the components of $b(\cdot)$ are constant, the components of $a(\cdot)$ are affine, and the components of $c(\cdot)$ are concave analytic functions on a neighbourhood of $[0,T]$, then SCLP has an optimal solution $\hat{x}(\cdot)$ that is piecewise affine. In this paper we demonstrate, under the (stronger) assumption that the set $D(t) = \{\xi : H\xi \leq b(t), \xi \geq 0\}$ is bounded for each t, that SCLP has an optimal solution $\hat{x}(\cdot)$ that is piecewise analytic if the components of $a(\cdot)$ and $b(\cdot)$ are analytic on a neighbourhood of $[0,T]$ and the components of $c(\cdot)$ are constant functions.

2. Solutions for separated continuous linear programs. The proof of our result proceeds according to the following plan. First, by using a dual problem, we construct a lower bound on the value of SCLP. Following the approach of Pullan [15] this lower bound is shown to be the same as the value of a discrete approximation to SCLP. We then define a certain partition and corresponding discrete program DP that has the property that the lower bound is tight. With this partition we may construct from the optimal solution $(\bar{x}, \bar{y}, \bar{z})$ to DP a piecewise analytic feasible solution to SCLP with the same objective function value as $(\bar{x}, \bar{y}, \bar{z})$, thus making it optimal.

Dual problems for SCLP can be defined in various ways. We choose to define the

dual SCLP* as follows:

$$\begin{aligned}\text{SCLP}^* \qquad &\text{maximize } -\int_0^T d\pi(t)^T a(t) - \int_0^T \eta(t)^T b(t)dt \\ &\text{subject to } c(t) - G^T\pi(t) + H^T\eta(t) \geq 0, \\ &\qquad \eta(t) \geq 0, \text{ a.e. on } [0,T], \\ &\qquad \pi(t) \text{ monotonic increasing and right continuous} \\ &\qquad \text{on } [0,T] \text{ with } \pi(T) = 0.\end{aligned}$$

Here the components of $\eta(\cdot)$ are in $L_1[0,T]$. It is straightforward (see [15]) to establish the following result.

LEMMA 2.1 (weak duality). $V(\text{SCLP}^*) \leq V(\text{SCLP})$.

(Here, and in what follows, we write $V(\text{P})$ for the optimal value of program P.)

We now restrict our attention to the case where the components of a and b are analytic on a neighbourhood of $[0,T]$ and the components of c are constant functions. In this case we can obtain a bound on $V(\text{SCLP})$ by considering the solution to a discretized problem. Formally, given a partition $\{t_0, t_1, \ldots, t_m\}$ of $[0,T]$, if we define $\bar{a}_i = a(t_i) - a(t_{i-1})$ and $\bar{b}_i = \int_{t_{i-1}}^{t_i} b(t)dt$, then we are led to the following discrete version of SCLP:

$$\begin{aligned}\text{DP} \qquad &\text{minimize} \quad \sum_{i=1}^m c^T\bar{x}_i \\ &\text{subject to} \quad G\bar{x}_1 + \bar{y}_1 = a(t_0) + \bar{a}_1, \\ &\qquad G\bar{x}_i + \bar{y}_i - \bar{y}_{i-1} = \bar{a}_i, \qquad i = 2,3,\ldots,m, \\ &\qquad H\bar{x}_i + \bar{z}_i = \bar{b}_i, \qquad i = 1,2,\ldots,m, \\ &\qquad \bar{x}_i, \bar{y}_i, \bar{z}_i \geq 0, \qquad i = 1,2,\ldots,m.\end{aligned}$$

The problem DP has a linear programming dual given by

$$\begin{aligned}\text{DP}^* \qquad &\text{maximize} \quad a^T(t_0)\bar{\pi}_1 + \sum_{i=1}^m \bar{a}_i^T\bar{\pi}_i - \sum_{i=1}^m \bar{\eta}_i^T\bar{b}_i \\ &\text{subject to} \quad c - G^T\bar{\pi}_i + H^T\bar{\eta}_i \geq 0, \quad i = 1,2,\ldots,m, \\ &\qquad \bar{\pi}_i - \bar{\pi}_{i-1} \geq 0, \qquad i = 2,3,\ldots,m, \\ &\qquad \bar{\eta}_i \geq 0, \qquad i = 1,2,\ldots,m, \\ &\qquad \bar{\pi}_m \leq 0.\end{aligned}$$

The following result gives a bound on $V(\text{SCLP})$.

LEMMA 2.2. $V(\text{DP}) \leq V(\text{SCLP})$.

Proof. Suppose that $(\bar{x}_i, \bar{y}_i, \bar{z}_i)$, $i = 1,2,\ldots,m$, is an optimal solution to DP. Then by the duality theorem of linear programming there is some $(\bar{\pi}, \bar{\eta})$ that solves DP* with $V(\text{DP}^*) = V(\text{DP})$. It is easy to see that any feasible solution to DP* can be used to construct a piecewise constant feasible solution to SCLP*. Formally we set $\pi(T) = \eta(T) = 0$ and define $\pi(t) = \bar{\pi}_i$, $\eta(t) = \bar{\eta}_i$, for $t \in [t_{i-1}, t_i)$. Furthermore, since

$$a^T(t_0)\bar{\pi}_1 + \sum_{i=1}^m \bar{a}_i^T\bar{\pi}_i - \sum_{i=1}^m \bar{\eta}_i^T\bar{b}_i = -\int_0^T d\pi(t)^T a(t) - \int_0^T \eta(t)^T b(t)dt,$$

it follows that $V(\text{DP}^*) \leq V(\text{SCLP}^*)$, whence the result follows from Lemma 2.1. □

In what follows, we show how to define a partition so that the optimal solution $\bar{x}$ to DP with this partition has the same value as $V(\text{SCLP})$. This can be achieved by constructing from $\bar{x}$ a feasible solution x to SCLP, which is analytic in each interval of the partition and satisfies $\int_0^T x(t)dt = \sum_{i=1}^m \bar{x}_i$. Since c is constant this solution x will have the same value as $V(\text{DP})$, and is thus a piecewise analytic optimal solution for SCLP by virtue of Lemma 2.2.

For each interval of the partition, the solution $x(t)$ at time t is defined to be a convex combination of certain extreme points of a subset of $D(t) = \{\xi : H\xi \leq b(t), \xi \geq 0\}$, with the property that its integral over the interval matches $\bar{x}_i$. To ensure that this is possible we need to assume that $D(t)$ is bounded for each t. Furthermore, to show that the $x(\cdot)$ constructed is feasible for SCLP we must ensure that it generates variables $y(\cdot)$ that are nonnegative. This is achieved by working with a very large partition with the property that each component of $y(\cdot)$ generated by each of the possible extreme points is monotonic on each interval of the partition, thus guaranteeing that $y(\cdot)$ is nonnegative between the endpoints of each such interval.

To define the extreme-point solutions alluded to above, we make use of the matrix

$$K = \begin{bmatrix} G & I & -I & 0 \\ H & 0 & 0 & I \end{bmatrix},$$

obtained by differentiating (1) and writing these constraints with (2) in matrix form, where the columns correspond to variables $x(\cdot)$, $\dot{y}^+(\cdot)$, $\dot{y}^-(\cdot)$, $z(\cdot)$, and the right-hand side is $[\dot{a}(\cdot)\ b(\cdot)]^T$. Clearly K has full rank, and so any $n_2 + n_3$ linearly independent columns of K form a basis matrix K_B, say. Let

$$Q = \left\{ K_B^{-1} \begin{bmatrix} \dot{a}(\cdot) \\ b(\cdot) \end{bmatrix} : K_B \quad \text{is a basis matrix of} \quad K \right\},$$

where $\dot{a}$ denotes da/dt, and let

$$R = \{\rho_j(\cdot) : \quad \rho \in Q, \quad 1 \leq j \leq n_2 + n_3\}.$$

Since each component of $\dot{a}(\cdot)$ and $b(\cdot)$ is analytic on a neighbourhood of $[0, T]$, R consists of a finite set of analytic functions. By a standard result on analytic functions, each of these is either identically zero or has a finite number of zeros on $[0, T]$. Let $\{t_0, t_1, \ldots, t_m\}$ be the smallest partition of $[0, T]$ that contains all the zeros of each function in R that is not identically zero. We call this the *canonical partition* of $[0, T]$. It follows for any subinterval (t_{i-1}, t_i) of the canonical partition that for all $\rho_j(\cdot) \in R$, either $\rho_j(t) > 0$, $t \in (t_{i-1}, t_i)$, or $\rho_j(t) = 0$, $t \in (t_{i-1}, t_i)$, or $\rho_j(t) < 0$, $t \in (t_{i-1}, t_i)$.

Henceforth we concentrate our attention on a typical subinterval (t_{i-1}, t_i) of the canonical partition as defined above. For such an interval we make the following definitions. Given any vector $u^{(i)} \in \{-1, 0, 1\}^{n_2}$, we let $U^{(i)} = \text{diag}(u^{(i)})$, and define

$$K^{(i)} = \begin{bmatrix} G & U^{(i)} & 0 \\ H & 0 & I \end{bmatrix},$$

$$\bar{E}_i = \left\{ \bar{\psi} : K^{(i)}\bar{\psi} = \begin{bmatrix} \bar{a}_i \\ \bar{b}_i \end{bmatrix}, \quad \bar{\psi} \geq 0 \right\},$$

and for $t \in (t_{i-1}, t_i)$, let

$$E_i(t) = \left\{ \psi : K^{(i)}\psi = \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix}, \quad \psi \geq 0 \right\}.$$

Here $\bar{a}_i$ and $\bar{b}_i$ are defined as above for the canonical partition. Thus $\bar{E}_i$ consists of vectors $\bar{\psi} = (\bar{\xi}^T, \bar{\eta}^T, \bar{\zeta}^T)^T$ that, upon identifying $\bar{x}_i$ with $\bar{\xi}$ and $\bar{z}_i$ with $\bar{\zeta}$, are feasible for DP in the ith interval of the partition without the constraint $\bar{y}_i \geq 0$ but with the restriction that over the interval the jth component of the y variables must not decrease when $u_j^{(i)} = 1$, must not increase when $u_j^{(i)} = -1$, and must remain constant when $u_j^{(i)} = 0$. Similarly $E_i(t)$ consists of vectors that satisfy the differentiated constraints of SCLP in the ith interval of the partition, as well as having y components decreasing, held constant, or increasing, depending on the value of $u^{(i)}$. (To simplify the notation we choose to suppress the dependence of $\bar{E}_i$ and $E_i(t)$ on $u^{(i)}$.) Observe that both $\bar{E}_i$ and $E_i(t)$ are convex, polyhedral sets. The following lemmas show that each extreme point of $\bar{E}_i$ is the integral over (t_{i-1}, t_i) of the corresponding extreme point of $E_i(t)$.

LEMMA 2.3. *If $\bar{\psi}$ is a basic feasible solution for $\bar{E}_i$ with basis matrix K_B, then*

$$\bar{\psi} = \int_{t_{i-1}}^{t_i} K_B^{-1} \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix} dt.$$

Proof. The proof is immediate from the definition of $\bar{a}_i$ and $\bar{b}_i$. □

LEMMA 2.4. *If $\bar{\psi}$ is a basic feasible solution for $\bar{E}_i$ with basis matrix K_B, then for all $t \in (t_{i-1}, t_i)$,*

$$K_B^{-1} \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix} \geq 0.$$

Proof. Let

$$\psi(t) = K_B^{-1} \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix},$$

and suppose for some j that $e_j^T \psi(t) < 0$ for some $t \in (t_{i-1}, t_i)$. Since the columns of $K^{(i)}$ that appear in K_B also appear in K, K_B is a basis matrix of K. Thus $e_j^T \psi(t) \in R$ and by virtue of the definition of the canonical partition, $e_j^T \psi(t) < 0$ for every $t \in (t_{i-1}, t_i)$. By Lemma 2.3 it follows that

$$e_j^T K_B^{-1} \begin{bmatrix} \bar{a}_i \\ \bar{b}_i \end{bmatrix} < 0,$$

which contradicts the assumption that $\bar{\psi}$ is a basic feasible solution for $\bar{E}_i$. □

The following result is now immediate.

LEMMA 2.5. *For each extreme point $\bar{\psi}$ of $\bar{E}_i$ there exists a function $\psi(\cdot)$, analytic on a neighbourhood of $[0, T]$, with $\psi(t)$ an extreme point of $E_i(t)$, and $\int_{t_{i-1}}^{t_i} \psi(t)dt = \bar{\psi}$.*

We are now ready to prove the main result of the paper. This shows how a piecewise analytic optimal solution to SCLP may be constructed from an optimal solution to DP with the canonical partition. Since the cost coefficients are all constant, we need only ensure that the piecewise analytic solution is feasible for SCLP and has

$\int_{t_{i-1}}^{t_i} x(s)ds$ matching $\bar{x}_i$ on the partition. The feasibility of the piecewise analytic solution is guaranteed by constructing it from the extreme points of sets $E_i(t)$, which have been defined so that the y variables are monotonic within each interval of the partition.

THEOREM 2.6. *Suppose that in* SCLP *the components of* $a(\cdot)$ *and* $b(\cdot)$ *are analytic on a neighbourhood of* $[0,T]$ *and the components of* $c(\cdot)$ *are constant functions. If* $D(t) = \{\xi : H\xi \leq b(t), \xi \geq 0\}$ *is bounded for all* t *then* SCLP *has an optimal solution for which* $x(\cdot)$ *is piecewise analytic.*

Proof. Since $D(t)$ is bounded, SCLP has an optimal solution, say $x^*(\cdot)$. Let $P = \{t_0, t_1, \ldots, t_m\}$ be the canonical partition for SCLP, and for each i define $\bar{x}_i = \int_{t_{i-1}}^{t_i} x^*(s)ds$. It is clear that we may define $\bar{y}_i$ and $\bar{z}_i$ similarly so that $(\bar{x}_i, \bar{y}_i, \bar{z}_i)$, $i = 1, 2, \ldots, m$, is feasible for DP with the partition P. For each $i = 1, 2, \ldots, m$ we define $u^{(i)}$ by

$$e_j^T u^{(i)} = \begin{cases} -1, & e_j^T(\bar{a}_i - G\bar{x}_i) < 0, \\ 0, & e_j^T(\bar{a}_i - G\bar{x}_i) = 0, \\ 1, & e_j^T(\bar{a}_i - G\bar{x}_i) > 0, \end{cases}$$

and define $\bar{E}_i$ and $E_i(t)$ as above.

For the moment let us consider $\bar{E}_i$ for some i. Let $\{\bar{\psi}^{(k)} : k = 1, 2, \ldots, N\}$ be the set of extreme points of $\bar{E}_i$, and denote the vector $(\bar{x}_i^T, (\bar{a}_i - G\bar{x}_i)^T U^{(i)}, \bar{z}_i^T)^T$ by $\hat{x}_i$. It is clear from the definition of $u^{(i)}$, that $\hat{x}_i \in \bar{E}_i$, whence since $\bar{E}_i$ is bounded (because $D(t)$ is) $\hat{x}_i$ is a convex combination of the extreme points of $\bar{E}_i$, i.e., there exists $\theta^{(k)} \geq 0$, $k = 1, 2, \ldots, N$, with $\sum_{k=1}^N \theta^{(k)} = 1$, and $\hat{x}_i = \sum_{k=1}^N \theta^{(k)} \bar{\psi}^{(k)}$. By Lemma 2.5 we can find functions $\psi^{(k)}(\cdot)$, analytic on a neighbourhood of $[0,T]$, such that $\psi^{(k)}(t)$ is an extreme point for $E_i(t)$, $t \in (t_{i-1}, t_i)$, and $\int_{t_{i-1}}^{t_i} \psi^{(k)}(t) = \bar{\psi}^{(k)}$.

Let

$$\psi_i(t) = \sum_{k=1}^N \theta^{(k)} \psi^{(k)}(t), \qquad t \in [t_{i-1}, t_i).$$

It follows that

$$\int_{t_{i-1}}^{t_i} \psi_i(t)dt = \sum_{k=1}^N \theta^{(k)} \int_{t_{i-1}}^{t_i} \psi^{(k)}(t)dt = \sum_{k=1}^N \theta^{(k)} \bar{\psi}^{(k)} = \hat{x}_i, \tag{4}$$

Now, letting $\psi_i(\cdot) = (\xi_i(\cdot)^T, \eta_i(\cdot)^T, \zeta_i(\cdot)^T)^T$, define $\hat{x}(t) = (x(t)^T, y(t)^T, z(t)^T)^T$ for $t \in [t_{i-1}, t_i)$, $i = 1, 2, \ldots, m$ by

$$\begin{aligned} x(t) &= \xi_i(t), \\ y(t) &= a(t_0) + \sum_{j=1}^{i-1} \int_{t_{j-1}}^{t_j} U^{(j)} \eta_j(t_j) + \int_{t_{i-1}}^{t} U^{(i)} \eta_i(s)ds, \\ z(t) &= \zeta_i(t), \end{aligned}$$

and define $\hat{x}(T) = \lim_{t \uparrow T} \hat{x}(t)$. Then $\hat{x}(\cdot)$ is piecewise analytic, and by virtue of the definition of $E_i(t)$, it follows that $x(\cdot) \geq 0$ and $z(\cdot) \geq 0$.

To show that $y(\cdot) \geq 0$ observe that for each i

$$y(t_i) - y(t_{i-1}) = \int_{t_{i-1}}^{t_i} U^{(i)} \eta_i(t)dt = U^{(i)} \int_{t_{i-1}}^{t_i} \eta_i(t)dt = U^{(i)} U^{(i)} (\bar{a}_i - G\bar{x}_i),$$

by virtue of (4) and the definition of $\hat{x}_i$. It follows from the definition of $U^{(i)}$ that $y(t_i) - y(t_{i-1}) = \bar{a}_i - G\bar{x}_i$. Since $y(t_0) = a(t_0)$, we have for $i = 1, 2, \ldots, m$, that $y(t_i) = \bar{y}_i \geq 0$. Since for $t \in [t_{i-1}, t_i)$, $\dot{y}(t) = U^{(i)}\eta_i(t)$ and $\eta_i(t) \geq 0$, it is easy to see for each such interval that $y(t)$ lies between $y(t_{i-1}) = \bar{y}_{i-1}$ and $y(t_i) = \bar{y}_i$, and hence is nonnegative.

Finally it remains to show that $\hat{x}(\cdot)$ is optimal for SCLP, which we do by observing that

$$\int_0^T c^T x^*(t)dt = \sum_{i=1}^m c^T \bar{x}_i = \sum_{i=1}^m c^T \int_{t_{i-1}}^{t_i} x(t)dt = \int_0^T c^T x(t)dt,$$

from which the result follows. □

COROLLARY 2.7. *Under the conditions of Theorem* 2.6 *the optimal value of* SCLP *equals the optimal value of* DP *with the canonical partition.*

Proof. Given any feasible solution to DP with the canonical partition, the proof of Theorem 2.6 shows how to construct a feasible solution to SCLP with the same value. If DP is unbounded then we may apply the construction to show that SCLP is unbounded, which contradicts an assumption of the theorem. Thus DP has an optimal solution, which corresponds to a feasible solution to SCLP with the same value, yielding the result by virtue of Lemma 2.2. □

COROLLARY 2.8 (strong duality). *Under the conditions of Theorem* 2.6 *the optimal value of* SCLP *equals the optimal value of* SCLP* *with attainment in both primal and dual.*

Proof. With the canonical partition, $V(\text{SCLP}) = V(\text{DP}) = V(\text{DP}^*) = V(\text{SCLP}^*)$ and the solutions of DP and DP* can be used to construct optimal solutions to SCLP and SCLP*, respectively. □

3. Concluding remarks. Under suitable assumptions on the problem data we have shown above that SCLP has a piecewise analytic optimal solution. However, the piecewise analytic optimal solution we construct in Theorem 2.6 is not necessarily a basic feasible solution, and although our boundedness assumption ensures that an optimal basic feasible solution to SCLP must exist, there is, as Perold [13, p. 14] has observed, no guarantee that any such solution will be piecewise analytic.

Theorem 2.6 may be applied to a number of examples of SCLP with constant costs. In particular we may show that the continuous-time minimum-cost network flow problem CNP of [5] has a piecewise analytic optimal solution when the arc costs are constant with time, and the arc and node capacities are analytic. A further special case of this problem is the continuous-time maximum flow problem (MFP) dealt with in [4] and [14]. If the arc and node capacities in MFP are analytic then MFP has a piecewise analytic solution. Indeed we can also show in this case that the corresponding generalized cut for this problem has a finite number of switches, and thus corresponds to an optimal solution to the dual of MFP.

It is interesting to speculate on whether a stronger result than Theorem 2.6 is true. By taking canonical partitions between breakpoints of the functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ it is certainly possible to extend Theorem 2.6 to the case where $c(\cdot)$ is required to be piecewise constant and $a(\cdot)$ and $b(\cdot)$ are piecewise analytic, respectively. In view of Pullan's result for concave analytic $c(\cdot)$ cited above, it is natural to seek an extension of Theorem 2.6 to the case where $c(\cdot)$, as well as $a(\cdot)$ and $b(\cdot)$, is an analytic function. Unfortunately, although this result may well be true, it is certainly much harder to establish.

We might also conjecture that some result similar to Theorem 2.6 is true when $b(\cdot)$ is a (not necessarily analytic) continuous function. In the absence of integral constraints (1), SCLP with constant cost coefficients becomes a linear program with a time varying right-hand side vector. Böhm [8] has shown, using a continuous selection theorem, that such a parametric linear program has an optimal solution that varies continuously with time. However, it is not known whether SCLP has a continuous optimal solution when the integral constraints are included, even if $a(\cdot)$ is required to be constant.

REFERENCES

[1] E. J. ANDERSON, *A Continuous Model For Job-Shop Scheduling*, Unpublished PhD. thesis, University of Cambridge, 1978.

[2] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, J. Wiley and Sons, Chichester, 1987.

[3] E. J. ANDERSON, P. NASH, AND A. F. PEROLD, *Some properties of a class of continuous linear programs*, SIAM J. Control Optim., 21 (1983), pp. 758–765.

[4] E. J. ANDERSON, P. NASH, AND A. B. PHILPOTT, *A class of continuous network flow problems*, Math. Oper. Res., 7 (1982), pp. 501–514.

[5] E. J. ANDERSON AND A. B. PHILPOTT, *A continuous-time network simplex algorithm*, Networks, 19 (1989), pp. 395–425.

[6] K. M. ANSTREICHER, *Generation of Feasible Descent Directions In Continuous-Time Linear Programming*, Tech. Report SOL 83-18, Department of Operations Research, Stanford University, Stanford, CA, 1983.

[7] R. BELLMAN, *Bottleneck problems and dynamic programming*, Proc. Nat. Acad. Sci., 39 (1953), pp. 947–951.

[8] V. BÖHM, *On the continuity of the optimal policy set for linear programs*, SIAM J. Appl. Math., 28 (1975), pp. 303–306.

[9] R. N. BUIE AND J. ABRHAM, *Numerical solutions to continuous linear programming problems*, Z. Oper. Res., 17 (1973), pp. 107–117.

[10] W. P. DREWS, *A simplex-like algorithm for continuous-time linear optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 309–322.

[11] R. J. HARTBERGER, *Representation extended to continuous time*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 297–307.

[12] R. S. LEHMAN, *On the Continuous Simplex Method*, RM-1386, Rand Corporation, Santa Monica, CA, 1954.

[13] A. F. PEROLD, *Fundamentals of a Continuous Time Simplex Method*, Tech. Report SOL 78-26, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[14] A. B. PHILPOTT, *Continuous-time flows in networks*, Math. Oper. Res., 15 (1990), pp 640–661.

[15] M. PULLAN, *An Algorithm for a Class of Continuous Linear Programs*, SIAM J. Control Optim., 31 (1993), pp. 1558–1577.

[16] ———, *Separated Continuous Linear Programs: Theory and Algorithms*, Unpublished PhD. thesis, University of Cambridge, 1992.

[17] R. G. SEGERS, *A generalised function setting for dynamic optimal control problems*, in Optimization Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 279–296.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1297–1321, September 1994

006

THE RICCATI EQUATION FOR OPTIMAL CONTROL PROBLEMS WITH MIXED STATE-CONTROL CONSTRAINTS: NECESSITY AND SUFFICIENCY*

VERA ZEIDAN†

Abstract. The goal of this paper is to conduct a complete study of second-order conditions for the optimal control problem with mixed state-control constraints. The conjugate point theory is presented and a necessary condition in terms of the corresponding Riccati equation is obtained. Sufficiency criteria are developed in terms of strengthened necessary conditions, including the Riccati equation. The results generalize the known ones for pure control constraints as well as for the mixed state-control constraints.

Key words. optimal control, mixed state-control constraints, Riccati equation, conjugate points, strong normality, necessary conditions, sufficient conditions, weak and strong local minima

AMS subject classification. 49B10

1. Introduction.

Second-order conditions for the optimal control problem with equality or inequality constraints on the control, state, or on both variables have been the focus of several papers in the literature starting about two decades ago; see, for instance, [12], [6], [11], [20], [14], [15], and [13]. For the case of pure control equality and inequality constraints, the accessory problem is given in [20], where necessary conditions involving conjugate points and the Riccati equation also were developed for optimality of continuous controls. For the same problem, sufficiency criteria for weak and strong local minima were obtained in [11]. They consist of strengthened necessary conditions and in particular, the Riccati equation presented in [20]. For pure state constraints, sufficient conditions in terms of a certain Riccati inequality are given in [14]. Concerning the problem with mixed control-state equality and inequality constraints, we can find the accessory problem in [12] and [6]. For the equality constraints case, see also [15]. The accessory problem can be also obtained by applying results known for the abstract nonlinear programming setting [8] and [16] to the optimal control problem. A study of the case when the accessory problem, corresponding to the one with equality mixed constraints, is abnormal can be found in [3]. In a recent work [9], a set of sufficient conditions in terms of a Riccati-type equation were established by applying a generalization of [8] to the problem. However, between these conditions and the necessary conditions there is a gap larger than expected. Recently a sufficiency criterion for weak local minimality of a continuous control was given in [13].

In this paper two objectives are accomplished for the optimal control problem with mixed state-control constraints and L^∞-control functions. First, the notion of conjugate points is introduced and necessary conditions in terms of this notion, and then in terms of a Riccati-type equation, were established. Hence, the results of [20] are generalized to the case of L^∞-controls and mixed state-control constraints. As we will see, our results invoke the concept of strong normality, which also extends the one given in [20]. The second objective is to complete the study of problems with mixed state-control constraints by developing second-order sufficient conditions (Theorem 6.1) for weak and strong local optimality that are natural strengthening

* Received by the editors July 27, 1992; accepted for publication (in revised form) February, 19, 1993.

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

of the necessary conditions obtained in the first part of the paper. Thus, the gap between the two sets of conditions is as small as possible, and hence the results in [9] and [11] are included in Theorem 6.1. In Corollary 6.1 we show that if the ideas in [14] were adopted to our setting, the resulting sufficiency criterion would be a special case of Theorem 6.1. Finally we show that the recent sufficiency theorem obtained in [13] for continuous control candidates is included in this paper (see Remarks 6.4 and 6.5).

The paper is divided as follows. Section 2 contains the statement of the problem. Known results needed for the rest of the paper are given in §3. A thorough study of the normality and the controllability of the accessory problem is presented in §4. The second-order necessary and sufficient conditions are given, respectively, in §§5 and 6. An example illustrating the results is given in §7.

2. Statement of the problem. Consider the following optimal control problem:

$$\text{(C)} \qquad \text{minimize } J(x,u) := \ell(x(b)) + \int_a^b g(t,x(t),u(t))dt$$

$$\text{subject to } \dot{x}(t) = f(t,x(t),u(t)) \qquad \text{a.e.} \tag{2.1}$$

$$x(a) = A\,, \qquad \psi(x(b)) = 0 \tag{2.2}$$

and

$$G(t,x(t),u(t)) \leq 0 \qquad \text{a.e.,} \tag{2.3}$$

where $x(\cdot) : [a,b] \longrightarrow \mathbb{R}^n$ is absolutely continuous (AC), $u(\cdot) : [a,b] \longrightarrow \mathbb{R}^m$ is in $L^\infty[a,b]$, and

$$\begin{aligned} &g : [a,b] \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}, \quad && f : [a,b] \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n \\ &\psi : \mathbb{R}^n \longrightarrow \mathbb{R}^r, && G : [a,b] \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^k \\ &\ell : \mathbb{R}^n \longrightarrow \mathbb{R}\,, \end{aligned}$$

and $r \leq n$, $k \leq m$.

DEFINITION 2.1. *A pair* (x,u) *is* admissible *for* (C) *if* x *is* AC, u *is in* $L^\infty[a,b]$ *and the constraints* (2.1)–(2.3) *are satisfied by* (x,u).

DEFINITION 2.2. *An admissible pair* $(\hat{x},\hat{u})$ *is a* weak *local minimum for* (C) *if for some* $\epsilon > 0$, $(\hat{x},\hat{u})$ *minimizes* $J(x,u)$ *over all admissible pairs* (x,u) *satisfying*

$$\| x - \hat{x} \|_\infty < \epsilon \text{ and } \| u - \hat{u} \|_\infty < \epsilon.$$

The pair $(\hat{x},\hat{u})$ *is a* strong *local minimum if only the first inequality holds.*

3. Preliminary results. Given a pair $(\hat{x},\hat{u}) \in AC \times L^\infty[a,b]$. We define

$$T(\hat{x},\hat{u};\epsilon) := \{(t,x,u) \in [a,b] \times \mathbb{R}^n \times \mathbb{R}^m : |x - \hat{x}(t)| < \epsilon \text{ and } |u - \hat{u}(t)| < \epsilon\}\,,$$

$((x(\cdot),u(\cdot)) \in T(\hat{x},\hat{u};\epsilon)$ means that $(t,x(t),u(t)) \in T(\hat{x},\hat{u};\epsilon)$ almost everywhere),

$$I(t) := \{i \in \{1,\dots k\} : G^i(t,\hat{x}(t),\hat{u}(t)) = 0\}$$

(the elements of $I(t)$ are arranged in an increasing order), and

$$G^{I(t)} := \left\{G^i : i \in I(t)\right\}, \qquad G^\emptyset := 0.$$

For $\mathcal{F} := (f, g, G)$, the following regularity conditions will be recalled.

(R1) There exist $\epsilon > 0$ and $\alpha > 0$ such that for $t \in [a, b]$ almost everywhere, $\mathcal{F}(t, \cdot, \cdot)$ and its first derivatives in (x, u) are continuous on $\{(x, u) : (t, x, u) \in T(\hat{x}, \hat{u}; \epsilon)\}$ uniformly in t and essentially bounded at $(t, \hat{x}(t), \hat{u}(t)), \psi$ and ℓ are C^1 on $\{x : |x - \hat{x}(b)| < \epsilon\}$, $\nabla\hat{\psi}(b) := \nabla\psi(\hat{x}(b))$ is of full rank, and if $I(t) \neq \emptyset$

$$G_u^{I(t)}(t, \hat{x}(t), \hat{u}(t)) \left(G_u^{I(t)}(t, \hat{x}(t), \hat{u}(t))\right)^* \geq \alpha I_{\mathrm{Card} I(t)} \quad \text{a.e.},$$

where A^* is the transpose of the matrix A and $I_{\mathrm{Card} I(t)}$ is the identity matrix with a number of rows and columns equal to the number of elements in $I(t)$, i.e., Card $I(t)$.

(R2) There exists $\epsilon > 0$ such that $\mathcal{F}(t, \cdot, \cdot)$ and its derivatives in (x, u) up to second order are continuous on $\{(x, u) : (t, x, u) \in T(\hat{x}, \hat{u}; \epsilon)\}$ uniformly in t and essentially bounded at $(t, \hat{x}(t), \hat{u}(t)), \psi$ and ℓ are C^2 on $\{x : |x - \hat{x}(b)| < \epsilon\}$.

(R3) For every subset $A \subset \{1, \ldots, k\}$ and for every bounded set $M \subset [a, b] \times \mathbb{R}^n \times \mathbb{R}^m$, there exists $\epsilon > 0$ such that, either $M_\epsilon^A := \{(t, x, u) \in M : -\epsilon \leq G^i(t, x, u) \leq 0, \forall i \in A\}$ is empty or $G_u^A(t, x, u)(G_u^A(t, x, u))^* \geq \epsilon$ on M_ϵ^A.

(R4) The multivalued map $t \rightsquigarrow I(t)$ is piecewise constant.

The next result is a weak version of the minimum principle applied to (C) [10, VI. 3]. Any admissible pair $(\hat{x}, \hat{u})$ satisfying conditions (a)–(e) below is an *extremal*. An extremal $(\hat{x}, \hat{u})$ is *normal* if the corresponding l_0 is not zero.

THEOREM 3.1 (see [10]). *Let $(\hat{x}, \hat{u})$ be a weak local minimum for* (C). *Assume* (R1) *holds; then there exist $\hat{p} : [a, b] \to \mathbb{R}^n$ in AC, $\hat{q} : [a, b] \to \mathbb{R}^k$ in $L^\infty[a, b]$, $\nu \in \mathbb{R}^r$ and $\lambda_0 \in \mathbb{R}$ such that*

(a) $\lambda_0 \geq 0$, *for all i, $\hat{q}^i(t) \geq 0$ almost everywhere, and $\lambda_0 + |\nu| + \| \hat{q} \|_\infty \neq 0$;*

(b) $-\dot{\hat{p}}^*(t) = \hat{H}_x(t)$ *almost everywhere;*

(c) $\hat{H}_u(t) = 0$ *almost everywhere;*

(d) $\hat{q}^i(t)\hat{G}^i(t) = 0$ *almost everywhere, for all $i = 1, \ldots, k$;*

(e) $\hat{p}(b) = [\nabla\hat{\psi}(b)]^*\nu + \lambda_0[\nabla\ell(\hat{x}(b))]^*$,

where

$$H(t, x, u, p, q, \lambda_0) = \lambda_0 g(t, x, u) + p^* f(t, x, u) + q^* G(t, x, u). \tag{3.1}$$

$\hat{H}(t)$ is the evaluation of H at $(t, \hat{x}(t), \hat{u}(t), \hat{p}(t), \hat{q}(t), \lambda_0)$, and $\hat{G}(t)$ is $G(t, \hat{x}(t), \hat{u}(t))$.

Remark 3.1. If λ_0 and $\hat{p}$ are given, then (R1) and conditions (c) and (d) yield the uniqueness of $\hat{q}$ (see the proof of Proposition 4.1).

The following result is a necessary condition of second order. It is in terms of the accessory problem associated to (C). It is given with more generality in [12]. It can also be deduced by applying the second-order necessary conditions developed in [8] or [16] for the abstract optimization problem. For the case where only equality mixed state-control constraints are present, the accessory problem can be found in [15].

Set $\Lambda_0 := \{(\lambda_0, \nu, p, q)$ satisfying Theorem 3.1 with $\lambda_0 + |\nu| + \| q \|_\infty = 1\}$.

THEOREM 3.2 (see [12]). *Assume* (R1)–(R3) *and that $(\hat{x}, \hat{u})$ is a weak local minimum for* (C). *Then Λ_0 is nonempty. If Λ_0 is a singleton then*

$$J_2(\eta, v) := \frac{1}{2}\eta^*(b)\Gamma\eta(b) + \frac{1}{2}\int_a^b (\eta^*(t), v^*(t)) \nabla^2_{(x,u)} \hat{H}(t) \begin{pmatrix} \eta(t) \\ v(t) \end{pmatrix} dt \geq 0$$

for all $(\eta, v) \in AC \times L^\infty[a, b]$ such that

$$\dot{\eta}(t) = \hat{f}_x(t)\eta(t) + \hat{f}_u(t)v(t) \quad \text{a.e.}, \tag{3.2}$$

$$\eta(a) = 0, \ \nabla\hat{\psi}(b)\eta(b) = 0, \tag{3.3}$$

$$\hat{G}_x^{I(t)}(t)\eta(t) + \hat{G}_u^{I(t)}(t)v(t) = 0 \quad a.e., \tag{3.4}$$

where

$$\Gamma = \sum_{i=1}^{r} \nu_i \nabla^2 \psi_i(\hat{x}(b)) + \lambda_0 \nabla^2 \ell(\hat{x}(b)), \tag{3.5}$$

$\hat{f}(t)$ *is* $f(t, \hat{x}(t), \hat{u}(t))$ *and* $\hat{\psi}(b) = \psi(\hat{x}(b))$.

Remark 3.2. Theorem 3.2 gives rise to the accessory problem; that is,

(AP) minimize $J_2(\eta, v)$ over $(\eta, v) \in AC \times L^\infty[a, b]$ and satisfying (3.2)–(3.4).

Thus, Theorem 3.2 states that when Λ_0 contains a single element, a necessary condition for optimality is that the minimum value of (AP) is zero.

4. Normality and controllability. In this section we derive a condition namely, the strong normality, that insures the normality of both problem (C) and the accessory problem (AP) and the uniqueness of the multipliers, as required in Theorem 3.2. It will turn out that strong normality is equivalent to the M-controllability of the linearized system, where M is $\nabla\hat{\psi}(b)$.

DEFINITION 4.1. *A pair* $(\hat{x}, \hat{u})$ *satisfying Theorem* 3.1 *is* normal *if* $\lambda_0 \neq 0$.

DEFINITION 4.2. *An admissible pair* $(\hat{x}, \hat{u})$ *is* strongly *normal on an interval* $[c, b] \subseteq [a, b]$ *if* (R1) *is satisfied on* $[c, b]$ *and the only solution to the system*

$$-\dot{p}(t) = \hat{f}_x^*(t)p(t) + \hat{G}_x^*(t)q(t) \qquad t \in [c, b] \quad a.e., \tag{4.1}$$

$$\hat{f}_u^*(t)p(t) + \hat{G}_u^*(t)q(t) = 0 \qquad t \in [c, b] \quad a.e., \tag{4.2}$$

$$p(b) = [\nabla\hat{\psi}(b)]^*\nu, \quad \text{and} \quad q^i(t)\hat{G}^i(t) = 0 \qquad \forall i = 1, \ldots, k, t \in [c, b] \quad a.e., \tag{4.3}$$

is $p \equiv 0$, *where* $q(\cdot)$ *in* $L^\infty[c, b]$ *and* $\nu \in \mathbb{R}^r$.

Remark 4.1. If $p \equiv 0$ solves the above system, then from (R1) it results that also $q \equiv 0$ on $[c, b]$ and $\nu = 0$.

Remark 4.2. Assume that the extremal $(\hat{x}, \hat{u})$ is strongly normal on $[a, b]$. Then we can easily show that any solution to (AP) is normal, $(\hat{x}, \hat{u})$ is normal for (C), and when $\lambda_0 = 1$, the multipliers $\hat{p}$, $\hat{q}$, and ν in Theorem 3.1 are unique.

Note that when no final state constraint is present; that is, when $\psi : \mathbb{R}^n \longrightarrow \{0\}$, then any admissible pair is automatically strongly normal on any subinterval $[c, b]$ of $[a, b]$.

Now we define the tangent subspace to the active constraints

$$T^{I(t)}(t) := \left\{ u \in \mathbb{R}^m : \hat{G}_u^{I(t)}(t)u = 0 \right\}.$$

Let $Y^I(\cdot)$ be uniformly bounded on $[a, b]$ with $Y^{I(t)}(t)$ a matrix whose columns form an orthonormal basis for $T^{I(t)}(t)$, and $Y^{I(t)}(t) = 0$ if $T^{I(t)}(t) = \{0\}$. Thus

$$\hat{G}_u^{I(t)}(t)Y^{I(t)}(t) = 0 \qquad \text{a.e.} \tag{4.4}$$

Set

$$\beta^{I(t)}(t) := \begin{cases} \left[\hat{G}_u^{I(t)}(t)\right]^* \left[\hat{G}_u^{I(t)}(t)\left(\hat{G}_u^{I(t)}(t)\right)^*\right]^{-1} & \text{if } I(t) \neq \emptyset \\ 0 & \text{if } I(t) = \emptyset. \end{cases} \tag{4.5}$$

Condition (R1) implies that the multivalued map $t \hookrightarrow I(t)$ from $[a,b]$ to the subsets of $\{1,\dots,k\}$ is measurable. Hence $\beta^I(\cdot)\,\hat{G}_x^I(\cdot)$ is in $L^\infty[a,b]$ and $Y^I(\cdot)$ can be chosen in $L^\infty[a,b]$.

The following result rephrases the strong normality definition in terms of a system that does not contain q.

PROPOSITION 4.1. *Assume* (R1) *on* $[c,b] \subseteq [a,b]$. *Then* $(\hat{x},\hat{u})$ *is strongly normal on* $[c,b]$ *if and only if the only solution on* $[c,b]$ *of the system*

$$-\dot{p}(t) = \left[\hat{f}_x^*(t) - \left(\hat{G}_x^{I(t)}(t)\right)^* \left(\beta^{I(t)}(t)\right)^* \hat{f}_u^*(t)\right] p(t), \qquad a.e., \tag{4.6}$$

$$\left(Y^{I(t)}(t)\right)^* (t)\hat{f}_u^*(t)p(t) = 0 \qquad a.e., \tag{4.7}$$

$$p(b) = [\nabla\hat{\psi}(b)]^*\nu \tag{4.8}$$

is $p \equiv 0$ *and* $\nu = 0$.

Proof of Proposition 4.1. Assume $(\hat{x},\hat{u})$ is strongly normal on $[c,b]$. If the system (4.6)–(4.8) has a nonzero solution (p,ν) then, by (R1), $p\not\equiv 0$. Define

$$q = \begin{pmatrix} q^1 \\ \vdots \\ q^k \end{pmatrix}$$

as follows:

$$q^i(t) := \begin{cases} -e_{\alpha_i}^* \left(\beta^{I(t)}(t)\right)^* \hat{f}_u^*(t)p(t) & \text{if } i \in I(t) \\ 0 & \text{if } i \notin I(t), \end{cases}$$

where α_i is the position of i in $I(t)$, and $e_{\alpha_i}^* = (0,0,\dots,0,1,0,\dots 0)$ has "1" in the α_ith position. First, it is clear that for all i, $q^i(t)\hat{G}^i(t) = 0$, for $t \in [c,b]$ almost everywhere, and $q(\cdot)$ is in $L^\infty[c,b]$. Thus, (4.8) yields that (4.3) satisfied. Now using q in (4.6) and (4.7) we obtain

$$-\dot{p}(t) = \hat{f}_x^*(t)p(t) + \left(\hat{G}_x^{I(t)}(t)\right)^* q^{I(t)}(t) \tag{4.9}$$

$$\hat{f}_u^*(t)p(t) + \left(\hat{G}_u^{I(t)}(t)\right)^* q^{I(t)}(t) = 0, \tag{4.10}$$

where $q^{I(t)} = \left(q^i\right)_{i\in I(t)}$.

This is equivalent to saying that p and q solve (4.1) and (4.2). Therefore, (p,q,ν) solve (4.1)–(4.3) with $p\not\equiv 0$. Thus, we obtain a contradiction and whence (4.6)–(4.8) has zero as the only solution.

Conversely, assume that zero is the only solution to (4.6)–(4.8). Let us show that $(\hat{x}, \hat{u})$ is strongly normal on $[c, b]$. If not, there exist $p \not\equiv 0$, $\nu \in \mathbb{R}^r$, and q in $L^\infty[c, b]$ solving (4.1)–(4.3). This is equivalent to (4.9), (4.10), and (4.8) being satisfied by $p \not\equiv 0$, ν, and q. Using (R1) we solve (4.10) for $q^{I(t)}$ to obtain

$$q^{I(t)}(t) = -\left(\beta^{I(t)}(t)\right)^* \hat{f}_u^*(t)p(t)\,,$$

which, if used in (4.9), produces (4.6). Equations (4.4) and (4.10) yield (4.7). From the hypothesis we must have $p \equiv 0$ and $\nu = 0$, which is a contradiction. Therefore $(\hat{x}, \hat{u})$ must be strongly normal. □

Now consider the linearized system of (2.1) and (2.3); that is,

$$\dot{\eta}(t) = \hat{f}_x(t)\eta(t) + \hat{f}_u(t)v(t) \qquad \text{a.e.}, \tag{4.11}$$

$$\hat{G}_x^{I(t)}\eta(t) + \hat{G}_u^{I(t)}v(t) = 0 \qquad \text{a.e.}, \tag{4.12}$$

where η is absolutely continuous and v is in $L^\infty[c, b]$. This system coincides with (3.2) and (3.4). If (R1) holds on $[c, b]$ then (4.12) yields that

$$v(t) = Y^{I(t)}(t)\alpha(t) - \beta^{I(t)}(t)\hat{G}_x^{I(t)}(t)\eta(t) \qquad \text{a.e.}, \tag{4.13}$$

where $\alpha(\cdot)$ is in $L^\infty[c, b]$.

Set $\Phi(t, c)$ as the fundamental matrix of the linear system

$$\dot{\eta}(t) = \left(\hat{f}_x(t) - \hat{f}_u(t)\beta^{I(t)}(t)\hat{G}_x^{I(t)}(t)\right)\eta(t). \tag{4.14}$$

Then the *reachable* set at b from $\eta(c) = 0$ of the system (4.11) and (4.12) is

$$\mathcal{R}_c(b) = \left\{ \int_c^b \Phi(b, c)\Phi^{-1}(s, c)\hat{f}_u(s)Y^{I(t)}(s)\alpha(s)ds : \alpha(\cdot) \in L^\infty[c, b] \right\}. \tag{4.15}$$

DEFINITION 4.3. *Let M be an $r \times n$-matrix of full rank $(r \leq n)$. We say that the system* (4.11), (4.12) *is* M-controllable *on* $[c, b]$ *if*

$$M\mathcal{R}_c(b) = \mathbb{R}^r.$$

When $M = \bigtriangledown\psi(\hat{x}(b))$, the M-controllability is shown below to be equivalent to the strong normality.

PROPOSITION 4.2. *Assume* (R1), *then $(\hat{x}, \hat{u})$ is strongly normal on $[c, b]$ if and only if the system* (4.11) *and* (4.12) *is $\bigtriangledown\hat{\psi}(b)$-controllable.*

Proof. It suffices to show that the nonstrong normality of $(\hat{x}, \hat{u})$ is equivalent to

$$0 \notin \operatorname{int} \bigtriangledown \hat{\psi}(b)\mathcal{R}_c(b).$$

By the separation theorem, this latter condition is equivalent to the existence of $\nu \in \mathbb{R}^r \backslash \{0\}$ with $\nu^* \bigtriangledown \hat{\psi}(b)\mathcal{R}_c(b) \geq 0$. This is equivalent to

$$\left(Y^{I(t)}(t)\right)^*(t)\hat{f}_u^*(t)p(t) = 0 \qquad \text{a.e.},$$

where

$$p(t) = [\Phi^{-1}]^*(t, c)\Phi^*(b, c) \bigtriangledown \hat{\psi}(b)\nu\,.$$

Thus p satisfies (4.6)–(4.8) with $p(b) \neq 0$. This is equivalent, by Proposition 4.1, to $(\hat{x}, \hat{u})$ is not strongly normal on $[c, b]$. □

5. Necessary conditions: Conjugate points–Riccati equation. In this section we introduce the notion of a conjugate point at an extremal $(\hat{x}, \hat{u})$. This definition is shown to be expressed in terms of a linear system of differential equations. The nonexistence of conjugate points to b in (a, b) turns out to be necessary for the optimality of an extremal that is strongly normal on each subinterval $[c, b]$ of $[a, b]$. Finally, another necessary condition in terms of a certain Riccati-type equation is derived.

Throughout this section (R1)–(R4) are assumed to hold on $[a, b]$ at a given extremal $(\hat{x}, \hat{u})$ that is strongly normal on $[a, b]$. By Remark 4.2, we can take λ_0 in Theorem 3.1 to be 1 and in this case the corresponding multipliers $\hat{p}$, $\hat{q}$ and ν are unique. The Hamiltonian of the problem is then

$$H(t, x, u, p, q) = g(t, x, u) + p^* f(t, x, u) + q^* G(t, x, u)$$

and $\hat{H}(t) = H(t, \hat{x}(t), \hat{u}(t), \hat{p}(t), \hat{q}(t))$.

DEFINITION 5.1. *A point $c \in [a, b)$ is* conjugate *to b along $(\hat{x}, \hat{u})$ if there exists a nonzero $(\eta, \lambda, \mu, v) \in AC \times AC \times L^\infty[a, b] \times L^\infty[a, b]$ with $\eta \not\equiv 0$ on $[a, c]$ and satisfying*

$$\dot{\eta}(t) = \hat{f}_x(t)\eta(t) + \hat{f}_u(t)v(t) \quad a.e., \tag{5.1}$$

$$-\dot{\lambda}(t) = \hat{f}_x^*(t)\lambda(t) + \hat{H}_{xx}(t)\eta(t) + \hat{H}_{xu}(t)v(t) + \hat{G}_x^*(t)\mu(t) \quad a.e., \tag{5.2}$$

$$\hat{f}_u^*(t)\lambda(t) + \hat{H}_{ux}(t)\eta(t) + \hat{H}_{uu}(t)v(t) + \hat{G}_u^*(t)\mu(t) = 0 \quad a.e., \tag{5.3}$$

$$\hat{G}_x^{I(t)}(t)\eta(t) + \hat{G}_u^{I(t)}(t)v(t) = 0 \quad a.e., \qquad \mu^i(t) = 0 \text{ for } i \notin I(t), \tag{5.4}$$

$$\lambda(b) = (\nabla\hat{\psi}(b))^*\delta + \Gamma\eta(b) \text{ for some } \delta \in \mathbb{R}^r, \tag{5.5}$$

$$\nabla\hat{\psi}(b)\eta(b) = 0, \quad \text{and} \quad \eta(c) = 0, \tag{5.6}$$

where Γ is defined by (3.5) *with $\lambda_0 = 1$ and $\mathbb{R}^0 = \{0\}$.*

Remark 5.1. Since $(\hat{x}, \hat{u})$ is strongly normal, the accessory problem (AP) is normal at any extremal (η, v). Thus, if (η, v) is an extremal for (AP), (5.1)–(5.6) hold for $c = a$ and for some (λ, μ, δ).

In the following proposition the conjugate point definition is rephrased in terms of a linear system in η and λ.

DEFINITION 5.2. *We say that the* strengthened Legendre–Clebsch *condition is satisfied at $(\hat{x}, \hat{u})$ if for some $\bar{\alpha} > 0$ we have*

$$\left(Y^{I(t)}(t)\right)^* \hat{H}_{uu}(t)Y^{I(t)}(t) \geq \bar{\alpha} I_{m-\mathrm{Card}I(t)} \quad t \in [a, b] \quad a.e. \tag{5.7}$$

whenever $Y^{I(t)}(t) \neq 0$, where $Y^{I(t)}(t)$ is defined in (4.4).

This condition is a strengthening of the Legendre–Clebsch condition

$$\left(Y^{I(t)}(t)\right)^* \hat{H}_{uu}(t)Y^{I(t)}(t) \geq 0 \text{ a.e.}$$

PROPOSITION 5.1. *Assume that* (5.7) *holds at $(\hat{x}, \hat{u})$. Then, $c \in [a, b)$ is conjugate to b if and only if there exists a nonzero $(\eta, \lambda) \in AC \times AC$ with $\eta \not\equiv 0$ on $[a, c]$, satisfying the following linear system:*

$$\dot{\eta} = \left[\hat{f}_x + \hat{f}_u(Z^I\hat{H}_{uu} - I_m)\beta^I\hat{G}_x^I - \hat{f}_u Z^I \hat{H}_{ux}\right]\eta - \hat{f}_u Z^I \hat{f}_u^*\lambda \quad a.e., \tag{5.8}$$

$$-\dot{\lambda} = \left[\hat{f}_x^* + \left(\hat{G}_x^I\right)^* (\beta^I)^* \left(\hat{H}_{uu} Z^I - I_m\right) \hat{f}_u^* - \hat{H}_{xu} Z^I \hat{f}_u^*\right] \lambda$$

$$+ \left[\hat{H}_{xx} - \left(\hat{H}_{uu}\beta^I \hat{G}_x^I - \hat{H}_{ux}\right)^* Z^I \left(\hat{H}_{uu}\beta^I \hat{G}_x^I - \hat{H}_{ux}\right) - \hat{H}_{xu}\beta^I \hat{G}_x^I \right. \tag{5.9}$$

$$\left. - \left(\hat{H}_{xu}\beta^I \hat{G}_x^I\right)^* + \left(\beta^I \hat{G}_x^I\right)^* \hat{H}_{uu}\beta^I \hat{G}_x^I\right] \eta \quad a.e.,$$

$$\lambda(b) = [\nabla\hat{\psi}(b)]^*\delta + \Gamma\eta(b) \quad \textit{for some } \delta \in \mathbb{R}^r \tag{5.10}$$

$$\nabla\hat{\psi}(b)\eta(b) = 0, \quad \textit{and} \quad \eta(c) = 0, \tag{5.11}$$

where the functions in (5.8) *and* (5.9) *are evaluated at* t, *the functions* Y^I *and* β^I *are given by* (4.4) *and* (4.5), *and*

(5.12)

$$Z^{I(t)}(t) = \begin{cases} Y^{I(t)}(t)\left[\left(Y^{I(t)}(t)\right)^* \hat{H}_{uu}(t) Y^{I(t)}(t)\right]^{-1} \left(Y^{I(t)}(t)\right)^*(t) & \textit{if } Y^{I(t)}(t) \neq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. First let us show that the existence of $(\eta, \lambda, \mu, v) \neq 0$ that solves (5.1)–(5.4) is equivalent to saying that $(\eta, \lambda) \neq 0$ and (η, λ, μ, v) satisfies (5.1)–(5.4). In fact, if $(\eta, \lambda, \mu, v) \neq 0$ and solves (5.1)–(5.4) with $(\eta, \lambda) \equiv 0$, it follows that

$$\hat{G}_u^I v = 0 \quad \text{a.e.}, \qquad \mu^i(t) = 0 \ \text{ for } \ i \notin I(t)$$

and

$$\hat{H}_{uu} v + \left(\hat{G}_u\right)^* \mu = 0 \quad \text{a.e.}$$

If $Y^{I(t)}(t) = 0$ for almost all t in an interval $\mathcal{J}_i \subset [a, b]$, then $\hat{G}_u^I$ is invertible almost everywhere there, and thus the above implies that $(\mu, v) \equiv 0$ on $\mathcal{J}_i$. If $Y^{I(t)}(t) \neq 0$ on an interval $\mathcal{J}_i$, then for $\mu^{I(t)}(t) = \left(\mu^i(t)\right)_{i \in I(t)}$ the above equations yield that for some $\alpha(\cdot)$,

$$v(t) = Y^{I(t)}(t)\alpha(t) \qquad \text{a.e.} \in \mathcal{J}_i,$$

and

$$\hat{H}_{uu}(t) Y^{I(t)}(t)\alpha(t) + \left(\hat{G}_u^{I(t)}(t)\right)^* \mu^{I(t)}(t) = 0 \quad \text{a.e.} \in \mathcal{J}_i.$$

Now, using (5.7) and (4.4), the last equation yields $\alpha(t) \equiv 0$ and $\mu^{I(t)}(t) \equiv 0$. Whence $(\mu, v) \equiv 0$, yielding a contradiction. Therefore, $(\eta, \lambda) \neq 0$.

To complete the proof we show that (η, λ, μ, v) solve (5.1)–(5.4) is equivalent to (η, λ) solve (5.8) and (5.9). Note that the first of (5.4) is equivalent to (4.13); that is,

$$v(t) = Y^{I(t)}(t)\alpha(t) - \beta^{I(t)}(t)\hat{G}_x^{I(t)}(t)\eta(t) \quad \text{a.e.} \tag{5.13}$$

for some $\alpha(\cdot) \in L^\infty[a, b]$, and (5.3) is equivalent to

$$(Y^I)^* \left[\hat{f}_u^* \lambda + \hat{H}_{ux}\eta + \hat{H}_{uu} v\right] = 0 \quad \text{a.e.} \tag{5.14}$$

with

$$\mu^I = \left(\beta^I\right)^* \left[-\hat{f}_u^* \lambda - \hat{H}_{ux}\eta - \hat{H}_{uu}v\right] \quad \text{a.e.} \tag{5.15}$$

Using (5.13) in (5.14), (5.15), (5.1), and (5.2), it results that (5.1)–(5.4) are equivalent to

$$v = \left[\left(Z^I \hat{H}_{uu} - I_m\right)\beta^I \hat{G}_x^I - Z^I \hat{H}_{ux}\right]\eta - Z^I f_u^* \lambda \quad \text{a.e.} \tag{5.16}$$

$$\mu^I = \left(\beta^I\right)^* \left(\hat{H}_{uu} Z^I - I_m\right)\left[\hat{f}_u^* \lambda + \left(\hat{H}_{ux} - \hat{H}_{uu}\beta^I \hat{G}_x^I\right)\eta\right] \quad \text{a.e.} \tag{5.17}$$

and (5.8) and (5.9) hold. □

Now we are ready to state the necessary condition involving the conjugate point theory.

THEOREM 5.1. *Let $(\hat{x}, \hat{u})$ be a weak local minimum for* (C). *Assume that $(\hat{x}, \hat{u})$ is strongly normal on each interval $[c, b] \subset [a, b]$, then there exists no point in (a, b) conjugate to b.*

Proof. We will argue by contradiction. If there exists $c \in (a, b)$ conjugate to b, by Proposition 5.1 we have a nonzero $(\eta, \lambda) \in AC \times AC$ with $\eta \not\equiv 0$ on $[a, c]$ satisfying (5.8)–(5.11).

Define v through (5.16) and set

$$(\bar{\eta}, \bar{\lambda}, \bar{v})(t) = (\eta, \lambda, v)(t)\chi_{[c,b]}(t),$$

where $\chi_{[c,b]}(\cdot)$ is the characteristic function of $[c, b]$. Thus, using Proposition 5.1 and the fact that (5.8), (5.9), (5.16), and (5.17) are equivalent to (5.1)–(5.4), it follows that $(\bar{\eta}, \bar{v})$ is admissible for the accessory problem (AP) and $J_2(\bar{\eta}, \bar{v}) = 0$. Hence, $(\bar{\eta}, \bar{v})$ solves (AP). Then, applying Theorem 3.1 to (AP), taking into account Remark 5.1, we obtain $(\tilde{\lambda}, \tilde{\mu}, \tilde{\delta})$ satisfying, with $(\bar{\eta}, \bar{v})$, (5.1)–(5.6) with $c = a$. The strong normality on $[c, b]$ yields that $\lambda \equiv \bar{\lambda} \equiv \tilde{\lambda}$ on $[c, b]$, and $\delta = \tilde{\delta}$. Since (η, λ) and $(\bar{\eta}, \tilde{\lambda})$ satisfy (5.8) and (5.9) and coincide on $[c, b]$, there they are equal on $[a, b]$. Thus, $\eta \equiv 0$ on $[a, c]$, which leads a contradiction. Therefore no $c \in (a, b)$ is conjugate to b. □

Now we define what would have been the extension to the mixed state-control constraints of the classical notion of conjugate points.

DEFINITION 5.3. *A point $c \in (a, b)$ is* classically conjugate *to b along $(\hat{x}, \hat{u})$ if there exists a nonzero $(\eta, \lambda, \mu, v) \in AC \times AC \times L^\infty[a, b] \times L^\infty[a, b]$ satisfying* (5.1)–(5.6).

Remark 5.2. By Proposition 5.1, the above definition is equivalent to saying that there exists $(\eta, \lambda) \neq 0$ satisfying (5.8)–(5.11). Thus, when G depends only on u, the notion of classically conjugate point reduces to one given in [20, Def. 6.1]. As we will soon see, whenever $(\hat{x}, \hat{u})$ is strongly normal on any interval of the form $[a, c]$, Definitions 5.1 and 5.3 are equivalent. However, in the literature (see, e.g., [20]) the nonexistence of conjugate points is proven necessary for optimality of a piecewise continuous control $\hat{u}$, under the two-sided strong normality at $(\hat{x}, \hat{u})$. Therefore Theorem 5.1 is a generalization of those results, not only to the case where $\hat{u}$ in $L^\infty[a, b]$ and the constraints are in terms of both the control and the state variables, but also to the case when only strong normality on intervals of the form $[c, b]$ holds, that is, only one-sided strong normality is required.

DEFINITION 5.4. *A pair $(\hat{x}, \hat{u})$ is* strongly normal *on $[a, c] \subseteq [a, b]$ if on $[a, c]$ the system* (4.1), (4.2) *and*

$$q^i(t)\hat{G}^i(t) = 0 \quad \forall i = 1, \ldots, k$$

has only $p \equiv 0$ as a solution.

Remark 5.3. Since (R1) holds, Proposition 4.1 yields that Definition 5.4 is equivalent to saying that on $[a, c]$ the system (4.6), (4.7) has only $(p, \nu) = 0$ as a solution.

PROPOSITION 5.2. *Assume that $(\hat{x}, \hat{u})$ is strongly normal on any interval $[a, c] \subseteq [a, b]$. Then, Definitions* 5.1 *and* 5.3 *are equivalent.*

Proof. One direction of the equivalence is trivial. Now, assume c is classically conjugate to b but not conjugate to b. Then, by Proposition 5.1, there exists $(\eta, \lambda) \neq 0$ satisfying (5.8)–(5.11) with $\eta \equiv 0$ on $[a, c]$. Using this last property in (5.8) and (5.9) we obtain that λ satisfies on $[a, c]$ the system (4.6), (4.7). Using the strong normality on $[a, c]$, it results that $\lambda \equiv 0$ on $[a, c]$. Thus, (η, λ) must be zero on $[a, b]$. This is a contradiction. Therefore, the two definitions are equivalent. □

Consider the matrix system associated with the system (5.8)–(5.11):

$$\dot{X} = AX - \hat{f}_u Z^I \hat{f}_u^* \Lambda, \tag{5.18}$$

$$-\dot{\Lambda} = A^*\Lambda + \Big[\hat{H}_{xx} - \Big(\hat{H}_{uu}\beta^I \hat{G}_x^I - \hat{H}_{ux}\Big)^* Z^I \Big(\hat{H}_{uu}\beta^I \hat{G}_x^I - \hat{H}_{ux}\Big) - \hat{H}_{xu}\beta^I \hat{G}_x^I - \Big(\hat{H}_{xu}\beta^I \hat{G}_x^I\Big)^* + \Big(\beta^I \hat{G}_x^I\Big)^* \hat{H}_{uu}\beta^I \hat{G}_x^I\Big] X, \tag{5.19}$$

$$NX(b) = 0, \quad (I_n - N)(\Gamma X(b) - \Lambda(b)) = 0, \tag{5.20}$$

where

$$A = \hat{f}_x + \hat{f}_u(Z^I \hat{H}_{uu} - I_m)\beta^I \hat{G}_x^I - \hat{f}_u Z^I \hat{H}_{ux},$$

and

$$N = (\nabla\hat{\psi}(b))^*[\nabla\hat{\psi}(b)(\nabla\hat{\psi}(b))^*]^{-1} \nabla\hat{\psi}(b) \tag{5.21}$$

is a projection.

Theorem 5.1 and Propositions 5.1 and 5.2 lead to the following corollary.

COROLLARY 5.1. *Let $(\hat{x}, \hat{u})$ satisfy the conditions of Theorem* 5.1. *Assume in addition that $(\hat{x}, \hat{u})$ is strongly normal on each interval of the form $[a, c] \subseteq [a, b]$. Then there exists (X, Λ) solving* (5.18)–(5.20) *with*

$$X^*\Lambda = \Lambda^* X \quad \textit{and} \quad \det X(t) \neq 0 \textit{ on } (a, b).$$

Proof. Let (X, Λ) be the solution of (5.18), (5.19) with the following boundary conditions:

$$X(b) = I_n - N \quad \text{and} \quad \Lambda(b) = \Gamma(I_n - N) - N; \tag{5.22}$$

then (X, Λ) satisfies (5.18)–(5.20) and $d/dt[X^*\Lambda - \Lambda^* X] = 0$. Using (5.22) we deduce that (X, Λ) satisfies $X^*\Lambda = \Lambda^* X$.

Now, if for some $c \in (a, b)$ and $\alpha (\neq 0) \in \mathbb{R}^n$ we have $X(c)\alpha = 0$, then $(\eta(t), \lambda(t)) := (X(t)\alpha, \Lambda(t)\alpha)$ satisfies (5.8)–(5.11) with $\delta = -[\nabla\hat{\psi}(b)(\nabla\hat{\psi}(b))^*]^{-1} \nabla\hat{\psi}(b)\alpha$. Moreover, $(\eta, \lambda) \neq 0$ since $(\eta(b), \lambda(b)) \neq 0$. Thus, from Proposition 5.1 and Definition 5.3, c is classically conjugate to b, and hence c is conjugate to b (see Proposition 5.2). However, this contradicts Theorem 5.1. Therefore the result is true. □

The next result is the final one in this section. It basically states that the existence of a solution to a certain Riccati equation is necessary for optimality. A natural strengthening of this condition will be shown in this paper to be sufficient.

COROLLARY 5.2. *Assume the conditions of Corollary* 5.1. *Then there exists a Lipschitz continuous symmetric matrix function* $W(\cdot)$ *satisfying on* (a,b) *the equation*

$$
\begin{aligned}
L^I(W) := {} & \dot{W} + \hat{f}_x^* W + W\hat{f}_x \\
& - \left(\hat{f}_u^* W + \hat{H}_{ux} - \hat{H}_{uu}\beta^I \hat{G}_x^I\right)^* Z^I \left(\hat{f}_u^* W + \hat{H}_{ux} - \hat{H}_{uu}\beta^I \hat{G}_x^I\right) \\
& - \left(\beta^I \hat{G}_x^I\right)^* \left(\hat{f}_u^* W + \hat{H}_{ux}\right) - \left(W\hat{f}_u + \hat{H}_{xu}\right)\beta^I \hat{G}_x^I \\
& + \left(\beta^I \hat{G}_x^I\right)^* \hat{H}_{uu}\beta^I \hat{G}_x^I + \hat{H}_{xx} = 0
\end{aligned} \tag{5.23}
$$

with

$$
\lim_{t\to b}(I_n - N)(\Gamma - W(t)X(t))(I_n - N) = 0\,, \tag{5.24}
$$

where $X(\cdot)$ *is continuous and* $X(b) = I_n - N$.

Proof. Define on (a,b)

$$
W := \Lambda X^{-1},
$$

where (X,Λ) is the pair in Corollary 5.1. We can easily check that W is Lipschitz symmetric and satisfies (5.23). Moreover

$$
W(t)X(t) = \Lambda(t) \text{ on } (a,b)\,,
$$

$X(\cdot)$ is continuous on $[a,b]$ with $X(b) = I_n - N$, and

$$
\Lambda(b) = \Gamma(I_n - N) - N\,.
$$

Therefore, (5.24) follows. □

Remark 5.3. If $W(\cdot)$ is continuous at b then (5.24) becomes

$$
\Gamma - W(b) = 0 \quad \text{on } \{y : \nabla\hat{\psi}(b)y = 0\}\,.
$$

6. Sufficient conditions: Riccati equation. Consider the problem (C) of §2. The goal of this section is to provide a sufficiency criterion for weak and strong local optimality in (C) of a pair $(\hat{x},\hat{u}) \in AC \times L^\infty[a,b]$. The conditions involved here are basically strengthening of the necessary conditions of §5 and of the Pontryagin minimum principle in §3.

Let $(\hat{x},\hat{u})$ be an extremal for (C) with corresponding multipliers $\lambda_0 = 1$, $\hat{q}(\cdot)$, $\hat{p}(\cdot)$ and ν. Define

$$
J(t) := \{i \in \{1,\dots,k\} : \hat{q}^i(t) > 0\}
$$

and, for $\gamma > 0$, (6.1)

$$J_\gamma(t) := \{i \in \{1, \dots, k\} : \hat{q}^i(t) > \gamma\}.$$

In this section we adopt the notation used in §§3 and 4 with $I(t)$ replaced by $J(t)$ or $J_\gamma(t)$, e.g., G^J, T^J, β^J, Z^J, etc, or $G^{J_\gamma}, T^{J_\gamma}, \beta^{J_\gamma}, Z^{J_\gamma}$, etc.

The following theorem consists of sufficiency results for weak and strong local minimality in (C).

THEOREM 6.1. *Let $(\hat{x}, \hat{u})$ be an admissible pair. Assume that there exist $(\hat{p}, \hat{q}) \in AC \times L^\infty[a,b]$, and $\nu \in \mathbb{R}^r$ such that $\lambda_0 = 1$, $(\hat{p}, \hat{q})$, and ν satisfy Theorem* 3.1, *and that* (R1), (R2) *hold, where $J(\cdot)$, defined in* (6.1), *replaces $I(\cdot)$. Suppose in addition that*

(1) *there exists $\gamma > 0$ such that $J(t) = J_\gamma(t)$ almost everywhere;*

(2) $\left(Y^{J(t)}(t)\right)^* (t)\hat{H}_{uu}(t)Y^{J(t)}(t) \geq \bar{\alpha} I_{m-\mathrm{Card}J(t)}$ $t \in [a,b]$ *almost everywhere, whenever $Y^{J(t)}(t) \neq 0$, where Y^J is defined by* (4.4) *with $J(t)$ instead of $I(t)$;*

(3) *there exist a Lipschitz symmetric matrix function W on $[a,b]$ and $\bar{\delta} > 0$ satisfying*

$$L^J(W) \geq \bar{\delta} I_n \qquad t \in [a,b] \quad a.e. \tag{6.2}$$

and

$$\Gamma - W(b) > 0 \quad \text{on } \{y : \nabla\hat{\psi}(b)y = 0\}$$

where $L^J(W)$ is defined by (5.23), *in which $J(t)$ replaces $I(t)$, and Γ and N are given by* (3.5) *and* (5.21), *respectively.*

Then $(\hat{x}, \hat{u})$ is a weak local minimum for (C).

Moreover, if condition (c) *of Theorem* 3.1 *is replaced by the following:*

$$\begin{aligned} &\text{There exists } \delta > 0 \text{ and a mapping } u : T(\hat{x}, \hat{p}; \delta) \to \mathbb{R}^m \text{ such that} \\ &u(t, \cdot, \cdot) \text{ is continuous uniformly in } t,\ u(t, \hat{x}(t), \hat{p}(t)) = \hat{u}(t) \text{ a.e., and} \\ &u(t,x,p) \in \arg\min_u \{p^* f(t,x,u) + g(t,x,u) : G(t,x,u) \leq 0\}, \end{aligned} \tag{6.3}$$

then $(\hat{x}, \hat{u})$ is a strong local minimum for (C).

Remark 6.1. When the Lagrange multipliers associated with the active constraints are all nonzero (no degenerate inequalities) $J(t)$ coincides with $I(t)$. Thus, all the conditions of Theorem 6.1 are natural strengthenings of the necessary conditions presented in §§3 and 5. Hence, the above theorem provides sufficient conditions as close as possible to the necessary ones. An attempt was made in [9] to obtain similar results for this problem using the abstract nonlinear programming approach. However, the conditions imposed there are strong and hence too far from the necessary ones. Furthermore, Theorem 6.1 generalizes Theorems 2.1 and 2.2 in [11] to the case where $\hat{u}$ is in $L^\infty[a,b]$ and mixed state-control constraints are present.

A second-order sufficiency criterion was given in [14, Thm. 4.2] for the problem (C) where G depends only on x. The technique used there is the one employed earlier in [17] for the case where $u \in U$. However, condition (d) of Theorem 4.2 in [14] does not take *fully* into account the presence of the inequality state constraints. In fact, if we were to extend Sorger's condition (d) to our case, that is, where G depends on (x,u), we obtain condition (b) of the next corollary, which is clearly stronger than conditions (2) and (3) of Theorem 6.1. As explained at the end of this section, in Remarks 6.5 and 6.6, the results of this section include those developed recently in [13] for the case when $\hat{u}$ and the data as a function of t are continuous on $[a,b]$, the

end points of $\hat{x}$ are fixed and condition (R1) is strengthened. In that case condition 1 of Theorem 6.1 is not needed.

COROLLARY 6.1. *In Theorem* 6.1 *conditions* 2 *and* 3 *can be replaced by the following stronger conditions*:

(a) $\hat{H}_{uu}(t) \geq \bar{\alpha} I_m$ *a.e.t.*,

(b) *there exists a Lipschitz symmetric matrix function* W_0 *satisfying on* $[a, b]$

$$D(W) := \dot{W} + \hat{f}_x^* W + W\hat{f}_x + \hat{H}_{xx} - [W\hat{f}_u + \hat{H}_{xu}]\hat{H}_{uu}^{-1}[\hat{f}_u^* W + \hat{H}_{ux}] \geq 0 \text{ a.e.},$$

and

$$\Gamma - W(b) \geq 0 \quad \text{on } \{y : \nabla\hat{\psi}(b)y = 0\}.$$

Proof. Condition (a) yields that $\hat{H}_{uu}^{-1}$ has essentially bounded entries. Hence, (R1) and (R2) imply the conditions of the embedding theorem for ordinary differential equations (see, e.g., Theorem 4.1 in the Appendix of [4]). This theorem leads to the existence of $\lambda > 0$ and a Lipschitz symmetric matrix function W on $[a, b]$ satisfying

$$D(W) \geq \lambda I_n \quad \text{a.e.}$$

and

$$\Gamma - W(b) \geq \lambda I_n \quad \text{on } \{y : \nabla\hat{\psi}(b)y = 0\}.$$

Since $\hat{H}_{uu}(t) \geq Z^J(t)$ almost everywhere, it follows from the inequalities above that $L^J(W) \geq D(W) \geq \lambda I_n$. □

Remark 6.2. Conditions (a) and (b) above are too strong, since they imply that condition (c) of Proposition 6.2, given below, holds for all $(x, u) \in \mathbb{R}^{n+m}$.

The following result sheds light on how to obtain a function W satisfying condition 3 of Theorem 6.1 without imposing strong conditions, but reasonable and verifiable ones.

COROLLARY 6.2. *Condition* 3 *of Theorem* 6.1 *can be replaced by the following*: *there exists on* $[a, b]$ *a Lipschitz symmetric solution* W_0 *of*

$$L^J(W) \geq 0 \quad \text{a.e.}$$

and

$$\Gamma - W(b) \geq 0 \quad \text{on } \{y : \nabla\hat{\psi}(b)y = 0\}.$$

Proof. We have $Y^J(\cdot)$ and $\beta^J(\cdot)\hat{G}_x^J(\cdot)$ in $L^\infty[a, b]$. Hence, condition 2 of Theorem 6.1 implies that $Z^J(\cdot)$ is also in $L^\infty[a, b]$ and hence the embedding theorem for ordinary differential equations [4] yields the result. □

The proof of Theorem 6.1 is based on the following result involving the Hamilton–Jacobi inequality. This technique was used earlier in [17], [11], [18] and recently in [13].

PROPOSITION 6.1. *Suppose there exists a function* $V : [a, b] \times \mathbb{R}^n \to \mathbb{R}$ *such that for almost all* t, $V(\cdot, \cdot)$ *is differentiable on* $T(\hat{x}; \epsilon_0)$, *and* $V(\cdot, x(\cdot))$ *is absolutely continuous whenever* $x(\cdot)$ *is. Assume that, for* $\bar{F}(t, x, u) := V_t(t, x) + V_x(t, x)f(t, x, u) + g(t, x, u)$

(i) $\min_{(x,u)} \{\bar{F}(t, x, u) : (t, x, u) \in T(\hat{x}, \hat{u}; \epsilon_0)$ *and* $G(t, x, u) \leq 0\} = \bar{F}(t, \hat{x}(t), \hat{u}(t))$ *a.e.*,

and

(ii) $\min_x \{\ell(x)_- V(b,x) : \psi(x) = 0 \text{ and } |x - \hat{x}(b)| < \epsilon_0\} = \ell(\hat{x}(b))_- V(b, \hat{x}(b))$.

Then $(\hat{x}, \hat{u})$ *is a weak local minimum for* (P).

If condition (i) is true for $(t,x,u) : (t,x) \in T(\hat{x}; \epsilon_0)$ and $G(t,x,u) \leq 0$, then $(\hat{x}, \hat{u})$ is a strong local minimum.

Proof. If (x,u) is admissible for (C) and in $T(\hat{x}, \hat{u}; \epsilon_0)$ then by (i) and (ii)

$$J(x,u) - J(\hat{x}, \hat{u}) = \int_a^b \{\bar{F}(t, x(t), u(t)) - \hat{\bar{F}}(t)\} dt + \ell(x(b)) - \ell(\hat{x}(b))$$

$$+ \int_a^b \frac{d}{dt}[V(t, \hat{x}(t)) - V(t, x(t))] dt$$

$$\geq 0.$$

If (i) holds for $(t,x) \in T(\hat{x}; \epsilon_0)$ and $G(t,x,u) \leq 0$, take (x,u) admissible with $x(\cdot) \in T(\hat{x}; \epsilon)$. The same arguments as before yield $J(x,u) \geq J(\hat{x}, \hat{u})$. □

The next result represents a sufficiency criterion for condition (i) of Proposition 6.1 to hold. This is a generalization of [1, Thm. 3.4.3] to the case where a parameter t is present. The following condition will be used:

(R5) $$\exists \alpha > 0 : \nabla_{(x,u)} \hat{G}^{J(t)}(t)(\nabla_{(x,u)} \hat{G}^{J(t)}(t))^* \geq \alpha I_{\mathrm{Card} J(t)} \quad \text{a.e.}$$

PROPOSITION 6.2. *Let* $F(t,x,u)$ *be a real valued map such that* (F,G) *satisfies for some* $\epsilon > 0$ *condition* (R2). *Assume* (R5) *holds and that there exists* $\hat{q} : [a,b] \to \mathbb{R}^k$ *in* $L^\infty[a,b]$ *such that*

(a) *for all* i *and for* $t \in [a,b]$ *almost everywhere,* $\hat{q}^i(t) \geq 0$ *and* $(\hat{q}^i(t))^* \hat{G}^i(t) = 0$;

$$\nabla_{(x,u)} \hat{F}(t) + \hat{q}^*(t) \nabla_{(x,u)} \hat{G}(t) = 0 \quad a.e.,$$

(b) *condition* (1) *of Theorem* 6.1 *holds for some* γ,

(c) *for almost all* t, $\nabla^2_{(x,u)} \hat{F}(t) + \sum_{i=1}^k \hat{q}^i(t) \nabla^2_{(x,u)} \hat{G}^i(t) \geq 2\bar{\gamma} I_{n+m}$ *on*

$$\mathcal{L}^J(t) := \left\{ (x,u) \in \mathbb{R}^{n+m} : \hat{q}^i(t) \nabla_{(x,u)} \hat{G}^i(t) \begin{pmatrix} x \\ u \end{pmatrix} = 0 \ \forall i \right\}. \tag{6.4}$$

Then there exists $0 < \epsilon_0 \leq \epsilon$ *such that condition* (i) *of Proposition* 6.1 *is satisfied for* $\bar{F} = F$.

Remark 6.3. Proposition 6.2 is a key result for proving Theorem 6.1. Consider the special case when $\hat{u}$ and (F,G) and its derivatives in (x,u) up to second order are continuous in t. Then $\hat{q}$ is also continuous and condition (b) of Proposition 6.2 is not needed. In this case the proof is much simpler. It is analogous to the proof of the corresponding result in [11], where the constraints $d = G$ and $h = 0$ depend only on u. However, if we wish to assume that $F(t,\cdot,\cdot)$ and $G(t,\cdot,\cdot)$ are only C^{1+} (have Lipschitz gradients [19]), then the Hessian matrices in condition (c) are replaced by Clarke's generalized Jacobians $\partial_{(x,u)} \nabla_{(x,u)} F$ and $\partial_{(x,u)} \nabla_{(x,u)} G$ (see [2]), which are assumed to be upper semicontinuous in (t,x,u), where

$$\partial \nabla S(s) := \text{Convex hull} \left\{ M = \lim_{i \to \infty} \nabla^2 S(s_i) : s_i \underset{i \to \infty}{\longrightarrow} s \text{ and } \nabla^2 S(s_i) \text{exists} \right\}.$$

Thus, by using [5, Lem. 1], the proof also follows from that in [11].

Since the dependence on t is only $L^\infty[a,b]$, the proof of Proposition 6.2 will require the next result, which is a generalization of Hoffman's Lemma [1, §3.3.4] to our setting, where the cone $\mathcal{K}$, the elements $x_i^*(t) := \nabla \hat{G}^i(t)$ and the index set J depend on a parameter t.

LEMMA 6.1. *Let $G(t,\cdot,\cdot)$ be differentiable uniformly in t on $T(\hat{x},\hat{u};\epsilon)$, $(\hat{x},\hat{u})$ and $\nabla_{(x,u)}\hat{G}(t)$ be in $L^\infty[a,b]$, and* (R4) *hold. Define, for $t \in [a,b]$ almost everwhere,*

$$\mathcal{K}^J(t) := \left\{ h \in \mathbb{R}^{n+m} : \nabla_{(x,u)}\hat{G}^i(t)h \le 0 \text{ for } i \in J(t) \right\}$$

and

$$\rho(h, \mathcal{K}^J(t)) := \inf_{k \in \mathcal{K}(t)} |h-k|.$$

Then there exists a constant C independent of h and t such that

$$\rho(h, \mathcal{K}^J(t)) \le C \left\{ \sum_{i \in J(t)} \left[\nabla_{(x,u)}\hat{G}^i(t)h \right]_+ \right\},$$

where

$$\left[\nabla_{(x,u)}\hat{G}^i(t)h \right]_+ = \begin{cases} 0 & \text{if} \quad \nabla_{(x,u)}\hat{G}^i(t)h < 0 \\ \nabla_{(x,u)}\hat{G}^i(t)h & \text{if} \quad \nabla_{(x,u)}\hat{G}^i(t)h \ge 0. \end{cases}$$

Hence, for

$$\mathcal{L}^J(t) := \left\{ h \in \mathbb{R}^{n+m} : \nabla_{(x,u)}\hat{G}^i(t)h = 0 \, \forall i \in J(t) \right\},$$

$$\rho(h, \mathcal{L}^J(t)) \le C \sum_{i \in J(t)} |\nabla_{(x,u)} \hat{G}^i(t)h|.$$

Proof of Lemma 6.1. The proof follows the steps of Hoffman's Lemma in [1] with some modifications. As in (6) of that proof,

$$\rho(h, \mathcal{K}^J(t)) = \sup \left\{ \sum_{i \in J(t)} \lambda_i \nabla_{(x,u)} \hat{G}^i(t)h : \left| \sum_{i \in J(t)} \lambda_i \nabla_{(x,u)} \hat{G}^i(t) \right| \le 1, \ \lambda_i \ge 0 \ \forall i \right\}.$$

Define the finite-dimensional space

$$L_t := \left\{ z = \sum_{i \in J(t)} \lambda_i \nabla_{(x,u)} \hat{G}^i(t) : \lambda_i \in \mathbb{R} \quad \forall i \right\}$$

and the linear continuous operator

$$\Lambda_t : \mathbb{R}^{\operatorname{Card} J(t)} \longrightarrow L_t$$

$$\lambda = \left(\lambda_1, \ldots \lambda_{\operatorname{Card} J(t)}\right)^* \longrightarrow \Lambda_t(\lambda) = \sum_{i \in J(t)} \lambda_i \nabla_{(x,u)} \hat{G}^i(t).$$

Thus, from (R4) it results that Λ_t is a bijection and

$$\Lambda_t(\lambda) = X := \left(\lambda^* \hat{G}_x^{J(t)}(t),\, \lambda^* \hat{G}_u^{J(t)}(t)\right)$$

implies that

$$\lambda^* = X \left(\nabla_{(x,u)} \hat{G}^{J(t)}(t)\right)^* \left[\nabla_{(x,u)} \hat{G}^{J(t)}(t) \left(\nabla_{(x,u)} \hat{G}^{J(t)}(t)\right)^*\right]^{-1},$$

and hence, for some constant C independent of h and t,

$$|\lambda| \leq C|X| \leq C \left| \sum_{i \in J(t)} \lambda_i \nabla_{(x,u)} \hat{G}^i(t) \right|.$$

Therefore, the above equation for ρ yields the result. □

Proof of Proposition 6.2. If $\hat{q} \equiv 0$ almost everywhere, then (c) and the continuity of $\nabla^2_{(x,u)} F(t,\cdot,\cdot)$ uniformly in t yield that for some $\epsilon_0 > 0 (\epsilon_0 \leq \epsilon)$, $F(t,\cdot,\cdot)$ is strictly convex and hence, by (a) the result follows. If $\hat{q} \not\equiv 0$, define

$$L(t,x,u,q) := F(t,x,u) + q^* G(t,x,u)$$

and $\mathcal{K}^J(t)$ and $\mathcal{L}^J(t)$ as in Lemma 6.1. By this same lemma, there exists a positive constant C_1 (independent of t) such that

$$\forall h \in \mathbb{R}^{n+m}, h = h_1 + h_2 \text{ with } h_1 \in \mathcal{K}^J(t) \text{ and}$$

$$|h_2| \leq C_1 \sum_{i \in J(t)} [\nabla \hat{G}^i(t) h]_+,$$

where

$$[\nabla \hat{G}^i(t) h]_+ = \begin{cases} 0 & \text{if } \nabla \hat{G}^i(t) h < 0 \\ \nabla \hat{G}^i(t) h & \text{if } \nabla \hat{G}^i(t) h \geq 0, \end{cases} \tag{6.5}$$

and ∇G stands for $\nabla_{(x,u)} G$. Furthermore, we can represent h_1 as

$$\begin{aligned} &h_1 = h_1' + h_1'', \text{ with } h_1' \in \mathcal{L}^J(t) \text{ and} \\ &|h_1''| \leq C_2 \sum_{i \in J(t)} |\nabla \hat{G}^i(t) h_1| = C_2 \left\{ - \sum_{i \in J(t)} \nabla \hat{G}^i(t) h_1'' \right\}, \end{aligned} \tag{6.6}$$

where $C_2 > 0$.

Choose $A > 0$ such that

$$A C_2^{-1} \gamma - C_1 \max_{i=1}^{i=k} \| \hat{q}^i \|_\infty \| \nabla \hat{G}^i \|_\infty - 1 > 0, \tag{6.7}$$

where γ is the constant in condition (b). Let $\delta_1 \in (0,1]$ be such that $\delta := (C_1 + A)\delta_1$ satisfy $\delta < 1$, and

$$\bar{\gamma}(1-\delta)^2 - 2 \| B \|_\infty \delta(1+\delta) - \| B \|_\infty \| \delta^2 - \frac{\bar{\gamma}}{2} \geq 0 \tag{6.8}$$

where $B(t) := \frac{1}{2} \nabla^2_{(x,u)} \hat{L}(t)$, and $\bar{\gamma}$ the constant in condition (c).

The regularity hypotheses on F, G, and $\hat{q}$ yield that $L(t, \cdot, \cdot, \hat{q}(t))$ satisfies the same regularity conditions as F does. Thus, for any $(x(\cdot), u(\cdot)) : G(t, x(t), u(t)) \leq 0$ almost everywhere, and for

$$h(t) := \begin{pmatrix} x(t) - \hat{x}(t) \\ u(t) - \hat{u}(t) \end{pmatrix},$$

Taylor's expansion and condition (a) yield

(6.9)

$$\begin{aligned} F(t, x(t), u(t)) &= L(t, x(t), u(t), \hat{q}(t)) - \hat{q}^*(t) G(t, x(t), u(t)) \\ &= \hat{L}(t) - \hat{q}^*(t) \hat{G}(t) + \left[\nabla \hat{L}(t) - \hat{q}^*(t) \nabla \hat{G}(t) \right] h(t) + r_1(h(t)) \\ &= \hat{F}(t) - \hat{q}^*(t) \nabla \hat{G}(t) h(t) + r_1(h(t)) \qquad \text{a.e.}, \end{aligned}$$

where as

$$|h(t)| \to 0, \frac{r_1(h(t))}{|h(t)|} \longrightarrow 0$$

uniformly in t. Moreover,

$$F(t, x(t), u(t)) \geq L(t, x(t), u(t), \hat{q}(t)) = \hat{L}(t) + h^*(t) B(t) h(t) + r_2(h(t)) \quad \text{a.e.} \tag{6.10}$$

where as

$$|h(t)| \to 0, \frac{r_2(h(t))}{|h(t)|^2} \longrightarrow 0$$

uniformly in t, and for all $i = 1, \ldots, k$,

$$0 \geq G^i(t, x(t), u(t)) = \hat{G}^i(t) + \nabla \hat{G}^i(t) h(t) + \rho_i(h(t)) \tag{6.11}$$

where as

$$|h(t)| \to 0, \frac{\rho_i(h(t))}{|h(t)|} \longrightarrow 0$$

uniformly in t.

Choose $\epsilon_0 \in (0, \|B\|_\infty^{-1})$ such that for all $h(\cdot)$ with $|h(t)| < \epsilon_0$ almost everywhere, we have

$$\sum_{i=1}^{k} |\rho_i(h(t))| + |r_1(h(t))| \leq \delta_1 |h(t)| \quad \text{a.e.} \tag{6.12}$$

and

$$|r_2(h(t))| \leq \frac{\bar{\gamma}}{2} |h(t)|^2 \quad \text{a.e.} \tag{6.13}$$

Now, let $(x(\cdot), u(\cdot))$ such that $G(t, x(t), u(t)) \leq 0$ almost everwhere, with

$$h(t) := \begin{pmatrix} x(t) - \hat{x}(t) \\ u(t) - \hat{u}(t) \end{pmatrix}$$

satisfying $|h(t)| < \epsilon_0$ almost everywhere. Let us show that $F(t, x(t), u(t)) \leq \hat{F}(t)$ almost everywhere.

From (6.5), (6.6), (6.11), and (6.12) we have

$$h(t) = h_1(t) + h_2(t) \text{ with } h_1(t) \in \mathcal{K}^J(t) \quad \text{a.e.},$$

$$|h_2(t)| \leq C_1 \sum_{i \in J(t)} |\rho_i(h(t))| \leq C_1 \delta_1 |h(t)| \quad \text{a.e.}, \tag{6.14}$$

$$h_1(t) = h_1'(t) + h_1''(t) \text{ with } h_1'(t) \in \mathcal{L}^J(t) \quad \text{a.e.},$$

and

$$|h_1''(t)| \leq C_2 \left[\sum_{i \in J(t)} - \nabla \hat{G}^i(t) h_1''(t) \right] \quad \text{a.e.} \tag{6.15}$$

Case 1. $|h_1''(t)| > A\delta_1 |h(t)|$ for $t \in M$ almost everwhere, where $M \subseteq [a, b]$ has a positive measure. Then, (6.15) yields that, on M,

$$A\delta_1 |h(t)| < |h_1''(t)| \leq C_2 \left\{ \sum_{i \in J(t)} - \nabla \hat{G}^i(t) h_1'' \right\}. \tag{6.16}$$

Thus, from (6.9), (6.14), (6.15), and the above inequality we have, for $t \in M$,

$$F(t, x(t), u(t)) - \hat{F}(t) = - \sum_{i \in J(t)} \langle \hat{q}_i(t) \nabla \hat{G}^i(t), h_2(t) + h_1''(t) \rangle + r_1(h(t)).$$

By (b), (6.16), (6.14), and (6.12),

$$F(t, x(t), u(t)) - \hat{F}(t) \geq \delta_1 |h(t)| \left\{ AC_2^{-1} \gamma - C_1 \max_{i=1}^{k} \| q_i \|_\infty \| \nabla \hat{G}_i \|_\infty - 1 \right\}.$$

Thus, using (6.7),

$$F(t, x(t), u(t)) - \hat{F}(t) \geq 0 \quad \text{for } t \in M \quad \text{a.e.}$$

Case 2. $|h_1''(t) \leq A\delta_1 |h(t)|$ on a subset $S \subseteq [a, b]$ of positive measure. Then, from (6.14) we have

$$|h_1''(t) + h_2(t)| \leq \delta |h(t)| \quad \text{a.e. on } S,$$

and hence, $h(t) = h_1'(t) + h_2'(t)$ with $h_1'(t) \in \mathcal{L}^J(t)$, $h_2'(t) := h_1''(t) + h_2(t)$. The above inequality implies

$$(1 - \delta)|h(t)| \leq |h_1'(t)| \leq (1 + \delta)|h(t)|.$$

Now, use these inequalities with (6.10), conditions (a) and (c), and (6.12) we obtain

$$\begin{aligned} F(t, x(t), u(t)) - \hat{F}(t) &\geq (h_1'(t) + h_2'(t))^* B(t)(h_1'(t) + h_2'(t)) + r_2(h(t)) \\ &\geq \bar{\gamma} |h_1'(t)|^2 - 2 \| B \|_\infty |h_1'(t)| |h_2'(t)| \\ &\quad - \| B \|_\infty |h_2'(t)|^2 - \frac{\bar{\gamma}}{2} |h(t)|^2 \\ &\geq \left[\bar{\gamma}(1 - \delta)^2 - 2 \| B \|_\infty (1 + \delta)\delta - \| B \|_\infty \delta^2 - \frac{\bar{\gamma}}{2} \right] |h(t)|^2 \\ &\geq 0 \qquad \text{for } t \in S \text{ a.e.} \end{aligned}$$

from (6.8). Therefore the result is proved. □

Remark 6.4. If condition (1) of Theorem 6.1 is not satisfied, Proposition 6.2 remains valid whenever there exists $\gamma > 0$ such that J in (R5) and $\mathcal{L}^J$ is replaced by J_γ, where

$$\mathcal{L}^{J_\gamma}(t) := \left\{ (x,u) \in \mathbb{R}^{n+m} : \nabla_{(x,u)} \hat{G}^i(t) \begin{pmatrix} x \\ u \end{pmatrix} = 0 \ \ \forall\, i \in J_\gamma(t) \right\}.$$

In fact, in this case Lemma 6.1 and the proof of Proposition 6.2 remain valid when we replace J by J_γ, the constant γ in (6.7) by the above γ, and (6.9) by

$$\begin{aligned} F(t,x(t),u(t)) &\geq L(t,x(t),u(t),\hat{q}(t)) - \sum_{i \in J_\gamma(t)} \hat{q}^i(t) G^i(t,x(t),u(t)) \\ &= \hat{L}(t) - \sum_{i \in J_\gamma(t)} \hat{q}^i(t) \hat{G}^i(t) + [\nabla \hat{L}(t) - \sum_{i \in J_\gamma(t)} \hat{q}^i(t) \nabla \hat{G}^i(t)] h(t) \\ &\quad + r_1(h(t)) \\ &= \hat{F}(t) - \sum_{i \in J_\gamma(t)} \hat{q}^i(t) \nabla \hat{G}^i(t) h(t) + r_1(h(t)) \qquad \text{a.e.} \end{aligned}$$

Proof of Theorem 6.1. The idea of the proof is to construct in terms of the given functions $\hat{p}, \hat{q}, W$, and $(\hat{x}, \hat{u})$ a function V that satisfies the conditions of Proposition 6.1.

Define

$$V(t,x) := \langle \hat{p}(t), x - \hat{x}(t) \rangle + \tfrac{1}{2} \langle x - \hat{x}(t),\, W(t)(x - \hat{x}(t)) \rangle \tag{6.17}$$

then the function $\bar{F}(t,x,u)$ in Proposition 6.1 becomes

$$\begin{aligned} \bar{F}(t,x,u) \quad &= \langle \dot{\hat{p}}(t), x - \hat{x}(t) \rangle - \langle \hat{p}(t), \dot{\hat{x}}(t) \rangle + \tfrac{1}{2} \langle x - \hat{x}(t),\, \dot{W}(t)(x - \hat{x}(t)) \rangle \\ &\quad + \langle x - \hat{x}(t),\, W(t) \dot{\hat{x}}(t) \rangle + \langle \hat{p}(t) + W(t)(x - \hat{x}(t)), f(t,x,u) \rangle + g(t,x,u). \end{aligned} \tag{6.18}$$

To prove condition (i) of Proposition 6.1, we shall use Proposition 6.2. From (6.18), it is clear that $\bar{F}$ satisfies the regularity conditions required there. From Theorem 3.1, it follows that

$$\nabla \hat{\bar{F}}(t) + \hat{q}^*(t) \nabla \hat{G}(t) = \left(\dot{\hat{p}}^*(t) + \hat{H}_x(t), \hat{H}_u(t) \right) = 0 \quad \text{a.e.},$$

and hence condition (a) of Proposition 6.2 holds. Moreover

$$\nabla^2 \hat{\bar{F}}(t) + \sum_{i=1}^{k} \hat{q}^i(t) \nabla^2 \hat{G}^i(t) = \begin{bmatrix} \dot{W} + \hat{f}_x^* W + W \hat{f}_x + \hat{H}_{xx} & W \hat{f}_u + \hat{H}_{xu} \\ \hat{f}_u^* W + \hat{H}_{ux} & \hat{H}_{uu} \end{bmatrix}(t). \tag{6.19}$$

Note that (6.4) and (6.1) yield that

$$\mathcal{L}^J(t) = \left\{ (x,u) \in \mathbb{R}^n \times \mathbb{R}^m : u = Y^J(t) d - \beta^{J(t)}(t) \hat{G}_x^{J(t)}(t) x \quad \text{for some } d \right\},$$

where $d \in \mathbb{R}^{m - \mathrm{Card} J(t)}$ and β^J and Y^J are defined in (4.4), (4.5) with $I(t)$ replaced by $J(t)$. Thus, condition (c) of Proposition 6.2 is equivalent to the following:

$$\begin{aligned} C(t) := &\begin{bmatrix} I_n & -\left(\beta^J \hat{G}_x^J\right)^* \\ O^* & \left(Y^J\right)^* \end{bmatrix} \begin{bmatrix} \dot{W} + \hat{f}_x^* W + W \hat{f}_x + \hat{H}_{xx} & W \hat{f}_u + \hat{H}_{xu} \\ \hat{f}_u^* W + \hat{H}_{ux} & \hat{H}_{uu} \end{bmatrix} \\ &\cdot \begin{bmatrix} I_n & 0 \\ -\beta^J \hat{G}_x^J & Y^J \end{bmatrix}(t) \geq 2\bar{\gamma} I_t \text{ a.e.}, \end{aligned}$$

where I_t is the $(n+m-\mathrm{Card}J(t)) \times (n+m-\mathrm{Card}J(t))$-identity matrix, and $0 = 0_{n\times(m-\mathrm{Card}J(t))}$. After straightforward calculation we can show that this condition is

$$C(t) = N^*(t) \begin{bmatrix} L^J(W) & 0 \\ 0^* & \left(Y^J\right)^* \hat{H}_{uu} Y^J \end{bmatrix} (t) N(t) \geq 2\bar{\gamma} I_t \quad \text{a.e.},$$

where

$$N(t) := \begin{bmatrix} I_n & 0 \\ E(t) & I_{m-\mathrm{Card}J(t)} \end{bmatrix},$$

$$E := \left(\left(Y^J\right)^* \hat{H}_{uu} Y^J\right)^{-1} \left(Y^J\right)^* [\hat{f}_u^* W + \hat{H}_{ux} - \hat{H}_{uu} \beta^J \hat{G}_x^J].$$

Condition (2) of Theorem 6.1 and condition (R1) yield that $\left(\left(Y^J\right)^* \hat{H}_{uu} Y^J\right)^{-1}$ and $\left[\hat{G}_u^J \left(\hat{G}_u^J\right)^*\right]^{-1}$ are bounded for almost all t in $[a,b]$. Thus, (R1) and (R2) imply that E is bounded for almost all t in $[a,b]$. Since

$$N^{-1}(t) = \begin{bmatrix} I_n & 0 \\ -E(t) & I_{m-\mathrm{Card}J(t)} \end{bmatrix}$$

then, for some $\gamma_0 > 0$,

$$\left|N^{-1}(t)\right| \leq \gamma_0 \text{ for } t \in [a,b] \quad \text{a.e.}$$

Whence, for almost all t and for any

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}^{m-\mathrm{Card}J(t)},$$

$$|v|^2 = \left|N^{-1}(t)N(t)v\right|^2 \leq \left|N^{-1}(t)\right|^2 |N(t)v|^2 \leq \gamma_0^2 |N(t)v|^2 .$$

Therefore, using conditions (2) and (3)(a) of Theorem 6.1 we get for $t \in [a,b]$ almost everhwere and for $\bar{\lambda} := \min\{\bar{\delta}, \bar{\alpha}\}$

$$\begin{aligned} v^* C(t) v &= v_1^* L^J(W)(t) v_1 + (v_1^* E^*(t) + v_2^*) \left(\left(Y^J\right)^* \hat{H}_{uu} Y^J\right)(t)(E(t)v_1 + v_2) \\ &\geq \bar{\delta}|v_1|^2 + \bar{\alpha}|E(t)v_1 + v_2|^2 \\ &\geq \bar{\lambda}|N(t)v|^2 \\ &\geq 2\bar{\gamma}|v|^2 , \end{aligned}$$

where $\bar{\gamma} = \bar{\lambda}/2\gamma_0^2$, proving that condition (c) of Proposition 6.2 holds. Thus, by the latter, there exists $\epsilon_1 > 0 (\epsilon_1 \leq \epsilon)$ such that

$$\bar{F}(t,x,u) \leq \hat{\bar{F}}(t) \quad \text{a.e.}$$

and for $(t,x,u) \in T(\hat{x},\hat{u};\epsilon_0)$ with $G(t,x,u) \leq 0$, that is, condition (i) of Proposition 6.1 is satisfied. On the other hand, consider the problem

$$\begin{array}{ll} \text{minimize} & \mathcal{F}(x) := \ell(x) - V(b,x) \\ \text{over} & \psi(x) = 0 \end{array}$$

where V is defined by (6.17).

From condition (e) of Theorem 3.1 we have

$$\nabla \mathcal{F}(\hat{x}(b)) + \nu^* \nabla \psi(\hat{x}(b)) = 0,$$

and from (3.5) and Condition (3)(b) of the Theorem, it follows

$$\nabla^2 \mathcal{F}(\hat{x}(b)) + \sum_{i=1}^{r} \nu_i \nabla^2 \psi_i(\hat{x}(b)) > 0$$

on $\{y : \nabla\psi(\hat{x}(b))y = 0\}$. Thus, from the finite-dimensional version of Proposition 6.2 (e.g., see [7]), it results that for some $\epsilon_0 > 0 (\epsilon_0 \leq \epsilon_1)$ condition (ii) of Proposition 6.1 holds. Therefore, by this proposition the first part of Theorem 6.1 is proved.

Now, to prove the rest of Theorem 6.1 we will use the second part of Proposition 6.1. Let δ be the constant in (6.3). Define

$$p(t,x) = \hat{p}(t) + W(t)(x - \hat{x}(t)).$$

Then, the continuity uniformly in t of $p(t,\cdot)$ and of $u(t,\cdot,\cdot)$ in (6.3) implies that there exists $\tilde{\epsilon} > 0$ $(\tilde{\epsilon} \leq \epsilon_0)$ such that, for almost all t and for $x : |x - \hat{x}(t)| < \tilde{\epsilon}$, we have

$$|u(t,x,p(t,x)) - \hat{u}(t)| < \epsilon_0. \tag{6.20}$$

Thus, for $(t,x) \in T(\hat{x};\tilde{\epsilon})$ and all $u : G(t,x,u) \leq 0$,

$$\bar{F}(t,x,u) \leq \bar{F}(t,x,u(t,x,p(t,x))).$$

Since $G(t,x,u(t,x,p(t,x))) \leq 0$, from (6.20) and the proof of the first part of Theorem 6.1 we obtain

$$\bar{F}(t,x,u) \leq \hat{\bar{F}}(t) \quad \text{for } (t,x) \in T(\hat{x};\tilde{\epsilon}) \text{ and } G(t,x,u) \leq 0.$$

From Proposition 6.1 we obtain that $(\hat{x},\hat{u})$ is a strong local minimum. □

Remark 6.5. If $\hat{G}_u^{J(t)}(t)$ is not of full rank uniformly in t, i.e., does not satisfy (R1), but instead (R5) is satisfied, then the above proof with Proposition 6.2 show that condition 2 and inequality (6.2) in Theorem 6.1 can be replaced by a more primitive condition involving (6.19), namely, for $t \in [a,b]$ almost everywhere,

$$\begin{bmatrix} \dot{W} + \hat{f}_x^* W + W\hat{f}_x + \hat{H}_{xx} & W\hat{f}_u + \hat{H}_{xu} \\ \hat{f}_u^* W + \hat{H}_{ux} & \hat{H}_{uu} \end{bmatrix}(t) \geq \bar{\gamma} I_{n+m} \tag{6.21}$$

on $\mathcal{L}^J(t)$, where $\mathcal{L}^J(t)$ is defined by (6.4) and $\bar{\gamma} > 0$. Hence, in the special setting where $\hat{u}$ and the t-dependence of the data are continuous, W is C^1 and both end points of x are fixed, the first part of Theorem 6.1 was recently proved in [13]. There, the condition used is (6.21) instead of inequality (6.2) and condition 2. If in this special setting the data g and f were only C^{1+} in (x,u) (as opposed to C^2), Remark 6.3 and the proof of Theorem 6.1 imply that Theorem 6.1 remains valid when inequality (6.2) and condition 2 are replaced by the following condition: for all $t \in [a,b]$,

$$\begin{bmatrix} \dot{W} + \hat{f}_x^* W + W\hat{f}_x + \sum_{i=1}^{k} \hat{q}^i \hat{G}^i_{xx} + \alpha & W\hat{f}_u + \sum_{i=1}^{k} \hat{q}^i \hat{G}^i_{xu} + \beta \\ \hat{f}_u^* W + \sum_{i=1}^{k} \hat{q}^i \hat{G}^i_{ux} + \gamma & \sum_{i=1}^{k} \hat{q}^i \hat{G}^i_{uu} + \delta \end{bmatrix}(t) \geq \bar{\gamma} I_{n+m}$$

on $\mathcal{L}^J(t)$, for all

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \in \partial_{(x,u)} \nabla_{(x,u)} \hat{\mathcal{H}}(t),$$

where $\mathcal{H}(t,x,u) = g(t,x,u) + \langle \hat{p}(t), f(t,x,u) \rangle$. Thus, the results in [13] for the problem (C) are included in this section.

Remark 6.6. All the results of this paper can be extended to the multidimensional control problem, that is, when $t \in \Omega \subseteq \mathbb{R}^s$. Therefore, from the previous remark it follows that this section generalizes [13, Thm. 2] to the case where the control $\hat{u}$ at the t-dependence of the data are merely essentially bounded, the function W is only Lipschitz and one end point of x is varying. Moreover, we provide strong as well as weak local minimality criteria.

As indicated in Remark 6.1, the lack of the t-continuity of the data is compensated for in Theorem 6.1 with condition 1. However, as we shall see below, if this condition is violated, we can still obtain a sufficiency criterion when conditions 2 and 3 of Theorem 6.1 are strengthened by using J_γ instead of J, where γ is some positive number.

THEOREM 6.2. *Suppose all the conditions, except condition* 1, *in Theorem* 6.1 *are satisfied where* J *is replaced by* J_γ, *for some* $\gamma > 0$. *Then the results of Theorem* 6.1 *remain valid.*

Proof. The proof is identical to that of Theorem 6.1, where J is replaced by J_γ and Remark 6.4 is used instead of Proposition 6.2. □

Remark 6.7. Corollaries 6.1 and 6.2, where J is replaced by γ, hold for Theorem 6.2.

7. Numerical example. For

$$\tilde{x}(t) := \begin{cases} t^2 |\sin \frac{\pi}{t}| & \text{for } t \neq 0 \\ 0 & \text{for } t = 0 \end{cases}$$

and

$$\tilde{u}(t) := 2t \left| \sin \frac{\pi}{t} \right| - \pi \cos \frac{\pi}{t} \operatorname{sgn}\left(\sin \frac{\pi}{t} \right) \quad \text{a.e.},$$

define the optimal control problem

$$(\tilde{\mathrm{C}}) \qquad \text{minimize} \int_0^1 \left[(u_1 - \tilde{u})^3 - \frac{1}{2} u_2^2 - \frac{1}{8}(x - \tilde{x} + 64)(u_1 - \tilde{u}) \right] dt$$

$$\text{subject to } \dot{x} = u_1 - \frac{1}{8}(x - \tilde{x}) \quad \text{a.e.},$$

$$x(0) = x(1) = 0,$$

$$(u_1 - \tilde{u})^2 + u_2^2 + 2x - 2\tilde{x} \leq 4 \quad \text{a.e.}$$

Set $\hat{x} = \tilde{x}, \hat{u} = (\tilde{u}, 2), \hat{p} \equiv 8$ and $\hat{q} \equiv \frac{1}{2}$. We will soon show the $(\hat{x}, \hat{u}, \hat{p}, \hat{q})$ satisfies the conditions of Theorem 6.1 and hence $(\hat{x}, \hat{u})$ is a *strong local minimum* for $(\tilde{\mathrm{C}})$.

Here we have $u \in \mathbb{R}^2, x \in \mathbb{R}$,

$$g(t,x,u) = (u_1 - \tilde{u}(t))^3 - \tfrac{1}{2} u_2^2 - \tfrac{1}{8}(x - \tilde{x}(t) + 64)(u_1 - \tilde{u}(t)),$$

$$f(t,x,u) = u_1 - \tfrac{1}{8}(x - \tilde{x}(t)),$$

$$G(t,x,u) = (u_1 - \tilde{u}(t))^2 + u_2^2 + 2x - 2\tilde{x}(t),$$

and

$$H(t,x,u,p,q) = pf(t,x,u) + g(t,x,u) + qG(t,x,u).$$

We can easily check that $l_0 = 1, \hat{x}, \hat{u}, \hat{p}$ and $\hat{q}$ satisfy the conditions of Theorem 3.1. Moreover, since $\tilde{x}$ is Lipschitz and $\tilde{u}$ is in $L^\infty[0,1]$, the data of the problem satisfies the regularity assumptions in (R1) and (R2). Furthermore, $J(t) \equiv 1$, and

$$\hat{G}_u(t)\hat{G}_u^*(t) = 16 \quad \text{a.e.}$$

Take

$$Y(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Then $Y^*(t)\hat{H}_{uu}(t)Y(t) = 1$. For this problem we have

$$\beta^1 = \begin{bmatrix} 0 \\ \frac{1}{4} \end{bmatrix},$$

$$Z^1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

and hence

$$L^1(W) = \dot{W} - \tfrac{1}{4}W - (W - \tfrac{1}{8})^2.$$

Take $W_0(t) = \frac{t}{2}$. It results that

$$L^1(W_0(t)) = \frac{-t^2}{4} + \frac{31}{64} \geq \frac{15}{64} \quad \text{a.e.}$$

Therefore, by the first part of Theorem 6.1, it follows that $(\hat{x}, \hat{u})$ is a weak local minimum for $(\tilde{C})$.

Note that the strong local normality on any subinterval of $[0,1]$ of the form $[0,c]$ or $[c,1]$ holds true at $(\hat{x}, \hat{u})$ and hence, $W_0(t) = \frac{t}{2}$ satisfies the necessary condition of Corollary 5.2.

Now we shall show that condition (6.3) of Theorem 6.1 holds.

Take $q \equiv \frac{1}{2}$. Then $H_{u_2} \equiv 0$. Consider the equations

$$\begin{cases} H_{u_1}(t,x,u,p,\frac{1}{2}) = 0 \\ G(t,x,u) = 0. \end{cases}$$

Since

$$\nabla_u(H_{u_1}, G) = \begin{bmatrix} 1 + 6(u_1 - \tilde{u}) & 0 \\ 2(u_1 - \tilde{u}) & 2u_2 \end{bmatrix},$$

by the implicit function theorem on Banach spaces we obtain that for some $\bar{\delta} > 0$ there exists a mapping $u : T(\hat{x}, \hat{p}; \bar{\delta}) \to \mathbb{R}^2$ such that $u(t, \hat{x}(t), \hat{p}(t)) = \hat{u}(t) = (\tilde{u}(t), 2)$ and $u(t, ., .)$ is continuous uniformly in t. Hence $u_2(t, x, p) > 0$ and

$$\left\{(u_1, u_2) : \nabla_u G(t, x, u(t, x, p)) \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = 0\right\}$$

is of the form $c\bar{u}$, where c is in $\mathbb{R}$ and

$$\bar{u} := \begin{pmatrix} 1 \\ \frac{\tilde{u}(t) - u_1(t,x,p)}{u_2(t,x,p)} \end{pmatrix}.$$

Moreover

$$\bar{u}^* \nabla_u^2 H\left(t, x, u, p, \frac{1}{2}\right) \bar{u} = 1 + 6(u_1 - \tilde{u}(t)) > \frac{1}{2}$$

for $\|u_1 - \tilde{u}\|_\infty < \frac{1}{12}$. Thus, for some $\delta < \bar{\delta}$, $u(t, x, p)$ provides a minimum for H over u whenever $(t, x, p) \in T(\hat{x}, \hat{p}; \delta)$. Therefore, from Theorem 6.1 it results that $(\hat{x}, \hat{u})$ is a strong local minimum.

Acknowledgment. The author thanks Professor H. Maurer for bringing to her attention the recent work [13].

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov and S. V. Fomin, *Optimal Control*, Consultants Bureau, New York, London, 1987.

[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[3] A. V. Dmitruk, *Jacobi-type conditions for the problem of Bolza with inequalities*, Matematicheskie Zametki, 35 (1984), pp. 813–827.

[4] M. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[5] D. Klatte and K. Tammer, *On second-order sufficient optimality conditions for $C^{1,1}$-optimization problems*, Optimization, 19 (1988), pp. 169–179.

[6] E. S. Levitin, A. A. Milyutin, and N. P. Osmolovkii, *Conditions of high order for a local minimum in problems with constraints*, Uspekhi Mat. Nauk, 33 (1978), pp. 85–148.

[7] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading, MA, 1984.

[8] H. Maurer and J. Zowe, *First- and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[9] H. Maurer, *The Two-Norm Approach for Second-Order Sufficient Conditions in Mathematical Programming and Optimal Control*, preprint, Department of Mathematics, Univeristät Münster, 1992.

[10] L. W. Neustadt, *Optimization. A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.

[11] D. Orrell and V. Zeidan, *Another Jacobi sufficiency criterion for optimal control with smooth constraints*, J. Optim. Theory Appl., 58 (1988), pp. 283–300.

[12] N. P. Osmolovskii, *Second-order conditions for a weak local minimum in an optimal control problem (necessity, sufficiency)*, Soviet Math. Dokl., 16 (1975), pp. 1480–1484.

[13] S. Pickenhain, *Sufficiency Conditions for Weak Local Minima in Multidimensional Optimal Control Problems with Mixed Control-State Restrictions*, preprint, Leipzig, Germany.

[14] G. Sorger, *Sufficient optimality conditions for nonconvex control problems with state constraints*, J. Optim. Theory Appl., 62 (1989), pp. 289–310.

[15] G. Stefani and P. Zezza, *Optimal control problems with mixed state-control constraints*, unpublished manuscript, to appear.

[16] J. WARGA, *Second-order necessary conditions in optimization*, SIAM J. Control. Optim., 22 (1984), pp. 524–528.

[17] V. ZEIDAN, *First- and second-order sufficient conditions for optimal control and the calculus of variations*, Appl. Math. Optim., 11 (1984), pp. 209–226.

[18] ———, *Sufficient conditions with minimal regularity assumptions*, Appl. Math. Optim., 20 (1989), pp. 19–31.

[19] ———, *Sufficient conditions for the generalized problem of Bolza*, Trans. Amer. Math. Society, 275 (1983), pp. 561–586.

[20] V. ZEIDAN AND P. ZEZZA, *The conjugate point condition for smooth control sets*, J. Math. Anal. Appl., 132 (1988), pp. 572–589.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1322–1331, September 1994

007

A MULTISTATE, MULTICONTROL PROBLEM WITH UNBOUNDED CONTROLS *

J. R. DORROH† AND GUILLERMO FERREYRA†

Abstract. This paper gives the optimal synthesis for a two-dimensional singular control problem of the Vidale–Wolfe type. The controls take values in an unbounded set. Moreover, the optimal feedback control turns out to be impulsive on certain regions, and the order in which these impulses occur is important. A parameterization of time is introduced into the problem. This helps to elucidate the optimal synthesis, to prove its optimality by the verification method, and to design suboptimal physical approximations to the optimal impulsive control.

Key words. unbounded multidimensional control, optimal advertising, dynamic programming

AMS subject classifications. 93C10, 93C75, 90A05

1. Introduction.

In this paper, we consider the singular control problem ($\mathcal{P}$): Maximize

$$J = J(x_0, y_0, u, v, \tau) = \int_0^\infty [Ax\tau + By\tau - u - v]e^{-\rho t}\, ds \tag{1.1}$$

over the set $\mathcal{U}$ of ordered triples (u, v, τ) of nonnegative locally integrable functions on $[0, \infty)$ subject to

$$\begin{aligned} \dot{x} &= \alpha(1-x)u - \beta x\tau, \quad x(0) = x_0, \qquad 0 < x_0 < 1, \\ \dot{y} &= \gamma(1 - y/x)v - \eta y\tau, \quad y(0) = y_0, \qquad 0 < y_0 < x_0, \\ \dot{t} &= \tau, \quad t(0) = 0, \quad t(\infty) = \infty, \end{aligned} \tag{1.2}$$

where $\dot{} = d/ds$, $\beta, \eta \geq 0$, and $A, B, \rho, \alpha, \gamma > 0$. In §2, we discuss necessary conditions for the problem ($\mathcal{P}$) in this generality. In §3, we give the optimal synthesis in the special case $\beta = 0$, $\eta = 0$. These restrictions will be imposed at that point.

The problem ($\mathcal{P}$) is motivated by the singular control problem ($\mathcal{P}_0$): Maximize

$$J_0 = J_0(x_0, y_0, u, v) = \int_0^\infty [Ax + By - u - v]e^{-\rho t}\, dt$$

over the set $\mathcal{U}_0$ of ordered pairs (u, v) of nonnegative locally integrable functions on $[0, \infty)$ subject to

$$\begin{aligned} x' &= \alpha(1-x)u - \beta x, \quad x(0) = x_0, \qquad 0 < x_0 < 1, \\ y' &= \gamma(1 - y/x)v - \eta y, \quad y(0) = y_0, \qquad 0 < y_0 < x_0, \end{aligned} \tag{1.3}$$

where $' = d/dt$.

As stated, ($\mathcal{P}_0$) has no solution, for it leads to impulsive controls. There is a vast literature on impulsive control problems; see, for example, [1], [2]. However,

*Received by the editors April 2, 1992; accepted for publication (in revised form) March 2, 1993.
†Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana 70803.

the methods described in these references require that the cost of a jump be given in the model. This would be highly artificial in our situation; moreover it is not clear how these costs would be arrived at without introducing the parameter s. See the discussion in [6, p. 365]. The problem $(\mathcal{P})$ is obtained from $(\mathcal{P}_0)$ by introducing the parameter s and treating t as a state variable. It is well known [3], [11] that since the vector fields $(1-x,0)$ and $(0,1-y/x)$ do not commute, serious difficulties occur in the attempt to define, by means of the parameterization introduced in (1.2), an extension of the notion of the solution of (1.3) to the case where u and v have simultaneous impulses. However, in an optimal control problem such as ours, there is a functional to be optimized, and in fact, an optimal parameterization of t can be determined along with optimal u and v. Obviously, because of the parameterization, the optimal synthesis will not be unique. However, given an initial state, the image in (x,y,t)-space of the optimal trajectory is unique, as are the optimal profit and the feedback controls. In the optimal synthesis, the control τ is equal to zero on some nondegenerate s-intervals. On these intervals, the time t is constant, the control u or v is nonzero, and the corresponding state variable x or y increases to a well-defined value. This would correspond to an impulsive control and a jump in x or y for the problem $(\mathcal{P}_0)$. Even though x and y are continuous functions of the parameter s in the problem $(\mathcal{P})$ that we treat, we refer to increases in x or y that occur on an s-interval on which t is constant as *jumps*. Remarkably, in some of these s-intervals, both x and y increase, but not simultaneously, and the order is important. To our knowledge, this is the first example that exhibits this phenomenon. The complex jumps have interesting implications for continuous approximations of the optimal synthesis. This is discussed in §4.

The problem $(\mathcal{P}_0)$ is the Vidale–Wolfe advertising model for two products of a company in which the saturation level for the rate of sales $x(t)$ of the first product is taken to be 1, and the saturation level for the rate of sales $y(t)$ of the second product is equal to the rate of sales of the first product. The controls u and v are the rates of investment in advertising each product. The quantities $1-x$ and $1-y/x$ are the portions of the potential sales rates of the two products upon which advertising has an effect. The terms $-\beta x$ and $-\eta y$ represent the "forgetting effect." That is, if no advertising of a product is done ($u=0$ or $v=0$), then its rate of sales decays exponentially. Note that it follows from the state equations that $0<x(t)<1$ and that $0<y(t)<x(t)$. The functional describing the profit contains two positive terms due to the rates of sales $x(t)$ and $y(t)$ and two negative terms due to the expenditure on advertising. The discount factor $e^{-\rho t}$ reflects the time value of money. Several variations of the Vidale–Wolfe model for a single product [12] have been considered in [4]–[6], [9], and [10].

We assume throughout the paper that

$$A\alpha > \rho+\beta, \qquad B\gamma > \rho+\eta. \tag{1.4}$$

From an economic viewpoint, these assumptions are quite natural for a profitable enterprise, and mathematically they lead to the richest case in which all features occur.

2. Necessary conditions. The Hamiltonian for the system (1.1), (1.2) is given by

$$\begin{aligned} H(x,y,t,u,v,\tau,p_1,p_2,p_3,p_4) =& [p_3-\beta x p_1-\eta y p_2-(Ax+By)e^{-\rho t}p_4]\tau \\ &+[\alpha(1-x)p_1+e^{-\rho t}p_4]u+[\gamma(1-y/x)p_2+e^{-\rho t}p_4]v, \end{aligned} \tag{2.1}$$

where p_1, p_2, p_3, p_4 satisfy the adjoint equations

$$\begin{aligned} \dot{p}_1 &= \alpha u p_1 + \beta\tau p_1 - \gamma\frac{y}{x^2}vp_2 + A\tau e^{-\rho t}p_4, \\ \dot{p}_2 &= \gamma\frac{v}{x}p_2 + \eta\tau p_2 + B\tau e^{-\rho t}p_4, \\ \dot{p}_3 &= \rho(-Ax\tau - By\tau + u + v)e^{-\rho t}p_4, \\ \dot{p}_4 &= 0. \end{aligned} \tag{2.2}$$

Assuming $p_4 = -1$, Pontryagin's maximum principle implies the following necessary conditions.

If $\tau = 0$ and $v > 0$, then $p_2 = e^{-\rho t}/\gamma(1 - y/x)$. Differentiating this expression with respect to s, substituting for $\dot{x}$ and $\dot{y}$ from (1.2), and substituting into the adjoint equation yields $(y/x^2)\alpha(1 - x)u = 0$, or $u = 0$. This implies that jumps are either horizontal or vertical, since $\tau = 0$, $u > 0$, $v > 0$ is not optimal. In the next section, we will show that if both x and y undergo jumps while t remains fixed, then it is optimal to jump first in x and then in y.

If $\tau > 0$, $u = 0$, $v > 0$, then $p_2 = e^{-\rho t}/\gamma(1 - y/x)$. Proceeding as before, we deduce the necessary condition

$$B\gamma(1 - y/x)^2 - (\rho + \beta)(1 - y/x) - (\eta - \beta) = 0.$$

Let $z = \tilde{z}$ be the smaller root of

$$B\gamma(1 - z)^2 - (\rho + \beta)(1 - z) - (\eta - \beta) = 0. \tag{2.3}$$

Then $0 < \tilde{z} < 1$.

If $\tau > 0$, $u > 0$, $v = 0$, then $p_1 = e^{-\rho t}/\alpha(1 - x)$. As before, we deduce the condition

$$A\alpha(1 - x)^2 - \rho(1 - x) - \beta = 0.$$

If $\tau > 0$, $u > 0$, $v > 0$, then $p_1 = e^{-\rho t}/\alpha(1 - x)$, and $p_2 = e^{-\rho t}/\gamma(1 - y/x)$. These lead to

$$A\alpha(1 - x)^2 - \rho(1 - x) - \beta = -\frac{y}{x^2}\frac{v}{\tau}\frac{\alpha(1 - x)^2}{1 - y/x},$$

and

$$B\gamma(1 - y/x)^2 - (\rho + \beta)(1 - y/x) - (\eta - \beta) = \alpha(1 - x)\frac{u}{\tau}.$$

These two equations, together with (1.2), define a spiral around the point $(\tilde{x}, \tilde{z}\tilde{x})$, where $\tilde{x}$ is the smaller root of

$$\left[A + \frac{\eta\tilde{z}^2}{\gamma(1 - \tilde{z})^2}\right]\alpha(1 - x)^2 - \rho(1 - x) - \beta = 0$$

(thus, $0 < \tilde{x} < 1$). These spirals are trajectories of our system only on the region $\{(x, y) : x \geq \tilde{x}, \quad y/x \leq \tilde{z}\}$ since we must have $u, v \geq 0$.

Moreover, higher-order conditions [8] imply that no portion of any one of these spirals is optimal. In fact, considering the system (1.2) with the added state variables s and w defined by

$$\begin{aligned} \dot{s} &= 1, \qquad s(0) = 0, \\ \dot{w} &= (-Ax\tau - By\tau + u + v)e^{-\rho t}, \qquad w(0) = 0, \end{aligned}$$

we have, for $\bar{x} = (s, x, y, t, w)$,

$$\dot{\bar{x}} = a_0(\bar{x}) + a_1(\bar{x})u + a_2(\bar{x})v + a_3(\bar{x})\tau,$$

where

$$a_0 = (1,0,0,0,0), \qquad a_1 = (0, \alpha(1-x), 0, 0, e^{-\rho t}),$$
$$a_2 = (0,0,\gamma(1-y/x), 0, e^{-\rho t}), \qquad a_3 = (0, -\beta x, -\eta y, 1, -(Ax+By)e^{-\rho t}).$$

Let

$$\bar{H}(\bar{x}, \bar{u}, \bar{p}) = H(x, y, t, u, v, \tau, p_1, p_2, p_3, p_4) + p_0,$$

where $\dot{p}_0 = 0$, $\bar{p} = (p_0, p_1, p_2, p_3, p_4)$, and $\bar{u} = (u, v, \tau)$. Following the notation of [8, p. 286], the controls are of degree h, with $1 \le h \le \infty$, since

$$\frac{\partial}{\partial \bar{u}_i} \frac{d}{ds} \frac{\partial \bar{H}}{\partial \bar{u}_i} = 0.$$

Then [8, Thm. 6.2, p. 286] implies that for $\tau, u, v, > 0$, we must have

$$\frac{\partial}{\partial \bar{u}_i} \frac{d}{ds} \frac{\partial}{\partial \bar{u}_j} \bar{H}(\bar{x}, \bar{u}, \bar{p}) = 0, \tag{2.4}$$

for $i, j = 1, 2, 3$. Now (2.4) implies

$$\frac{\partial}{\partial v} \frac{d}{ds} \frac{\partial H}{\partial u} = -\gamma \frac{y}{x^2} \alpha(1-x) p_2 = 0, \tag{2.5}$$

$$\frac{\partial}{\partial \tau} \frac{d}{ds} \frac{\partial H}{\partial u} = \beta \alpha p_1 + (A\alpha(1-x) - \rho) e^{-\rho t} p_4 = 0. \tag{2.6}$$

Thus $p_2 = 0$ by (2.5), and $p_4 = 0$ by (2.2), so that (2.6) implies that $p_1 = 0$. Since $\tau > 0$, its coefficient in (2.1) must be zero, and thus $p_3 = 0$. This contradicts the nontriviality of the adjoint variable. Therefore, no piece of a spiral is optimal.

3. The optimal synthesis. For the remainder of this paper we assume

$$\beta = \eta = 0.$$

Thus $\tilde{z}$ and $\tilde{x}$ are given by

$$B\gamma(1-\tilde{z}) - \rho = 0, \qquad A\alpha(1-\tilde{x}) - \rho = 0.$$

Since we are assuming that $B\gamma, A\alpha > \rho$, this implies $0 < \tilde{x}, \tilde{z} < 1$. We also define $\tilde{y} = \tilde{x}\tilde{z}$. We need one more technical assumption; it is

$$\left(B - \frac{\rho}{\gamma}\right)(B\gamma - \rho) \le A(A\alpha - \rho). \tag{3.1}$$

This assumption is made so that one of the switching curves will have geometry that enables us to obtain the optimal synthesis. The assumption that $\eta = 0$ eliminates a switching curve whose presence was very perplexing.

We resolve the nonuniqueness due to parameterization by taking τ to be 1 whenever $\tau \neq 0$, and by taking $u = 1$ whenever $u \neq 0$ and $\tau = 0$, and likewise for v. We have

already demonstrated that u and v simultaneously nonzero is not optimal. In fact, when $\beta = \eta = 0$, it is not optimal for any two controls to be simultaneously nonzero. To obtain our last switching curve, consider a trajectory starting at a state (x_0, y_0, t_0), moving horizontally to (x_1, y_0, t_0), then vertically to $(x_1, \tilde{z}x_1, t_0)$, and finally $x = x_1$, $y = \tilde{z}x_1$ for $t \geq t_0$. This is achieved by taking controls $u = 1$, $v = 0$, $\tau = 0$ first, then $u = 0$, $v = 1$, $\tau = 0$, and finally $u = 0$, $v = 0$, $\tau = 1$. Solving the adjoint equations with $p_i(\infty) = 0$ for $i = 1, 2, 3$ leads to $E(x_1, y_0) = 0$, where E is the function defined in Lemma 3.1.

LEMMA 3.1. *Let the function E be defined on the set*

$$\mathcal{E} = \{(x, y) : \tilde{x} < x < 1,\ 0 \leq y < x\tilde{z}\}$$

by

$$E(x, y) = \frac{1}{\gamma} \log\left(\frac{1 - \tilde{z}}{1 - y/x}\right) - \frac{y/x}{\gamma(1 - y/x)} + \frac{\tilde{z}}{\gamma(1 - \tilde{z})} - \frac{1}{\alpha(1 - x)} + \frac{A}{\rho}.$$

Then there is a unique function φ on the interval $[0, \tilde{y}]$ such that $x = \varphi(y)$ satisfies the equation $E(x, y) = 0$ for $0 \leq y \leq \tilde{y}$. Furthermore, $\varphi(\tilde{y}) = \tilde{x}$, φ is strictly decreasing, $\tilde{x} < \varphi(0) < 1$, and $E(x, y) < 0$ for $(x, y) \in \mathcal{E}$ and $y \geq \tilde{y}$ or $0 \leq y \leq \tilde{y}$, $x > \varphi(y)$.

Proof. It is easy to see that $E(\tilde{x}, \tilde{y}) = 0$, and that $E(x, y) < 0$ for $\tilde{x} < x < 1$, $\tilde{y} \leq y \leq x\tilde{z}$. We also see that

$$E_y(x, y) = -\frac{z}{\gamma x(1 - z)^2}$$

and that

$$E_x(x, y) = \frac{z^2}{\gamma x(1 - z)^2} - \frac{1}{\alpha(1 - x)^2},$$

where $z = y/x$. Thus $E_y(x, y) \leq 0$ for all $(x, y) \in \mathcal{E}$, and $E_y(x, y) < 0$ for $y > 0$. It is also clear that E_x has a strong maximum at $(\tilde{x}, \tilde{y})$. But the assumption (3.1) is exactly the assumption that $\tilde{x}E_x(\tilde{x}, \tilde{y}) \leq 0$. We also see that $E(\tilde{x}, y) > 0$ and $E(1^-, y) = -\infty$ for $0 \leq y < \tilde{y}$. The lemma now follows from the implicit function theorem and the intermediate value property for continuous functions.

DEFINITION 3.1. *We denote by w the function on $[0, \tilde{y}]$ defined implicitly by the requirement that $x = w(y)$ be a solution of $E(x, y) = 0$, where E is the function defined in Lemma* 3.1. *That is, w denotes the function φ defined in Lemma* 3.1.

The Hamilton–Jacobi–Bellman equation for the value function V is

$$0 = \max\{Pu + Qv + R\tau : u, v, \tau \geq 0\},$$

where

$$P = \alpha(1 - x)V_x - e^{-\rho t},$$
$$Q = \gamma(1 - y/x)V_y - e^{-\rho t},$$
$$R = V_t + (Ax + By)e^{-\rho t}.$$

We will give controls and prove that they are optimal. We will do this by computing the corresponding payoff function and proving that this function satisfies the Hamilton–Jacobi–Bellman equation and is thus the value function. Let the regions Ω_1, Ω_2, Ω_3, and Ω_4 be defined by

$$\Omega_1 = \{(x, y) : \tilde{x} < x < 1,\ x\tilde{z} < y < x\},$$
$$\Omega_2 = \{(x, y) : 0 < y \leq \tilde{y},\ w(y) < x < 1\} \cup \{(x, y) : \tilde{x} < x < 1,\ \tilde{y} \leq y < x\tilde{z}\},$$
$$\Omega_3 = \{(x, y) : 0 < y < \tilde{y},\ y < x < w(y)\},$$
$$\Omega_4 = \{(x, y) : \tilde{y} < y < \tilde{x},\ y < x < \tilde{x}\}.$$

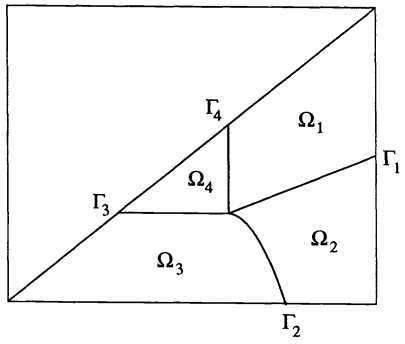


FIG. 1

Let the curves Γ_1, Γ_2, Γ_3, and Γ_4 be defined by

$$\begin{aligned}\Gamma_1 &= \{(x,y) : \tilde{x} < x < 1,\ y = x\tilde{z}\},\\ \Gamma_2 &= \{(x,y) : 0 < y < \tilde{y},\ x = w(y)\},\\ \Gamma_3 &= \{(x,y) : \tilde{y} < x < \tilde{x},\ y = \tilde{y}\},\\ \Gamma_4 &= \{(x,y) : x = \tilde{x},\ \tilde{y} < y < \tilde{x}\}.\end{aligned}$$

The regions Ω_k and the curves Γ_k are illustrated in Fig. 1. Let $\tilde{\Omega}_k = \Omega_k \times [0,\infty)$ and $\tilde{\Gamma}_k = \Gamma_k \times [0,\infty)$ for $k = 1,2,3,4$. The surfaces $\tilde{\Gamma}_1$, $\tilde{\Gamma}_2$, and $\tilde{\Gamma}_4$ are switching surfaces, but $\tilde{\Gamma}_3$ is not.

THEOREM 3.1. *Under the assumption* (3.1) (*and* $\beta = \eta = 0$, $A\alpha, B\gamma > \rho$), *optimal controls for the problem* ($\mathcal{P}$) *are given by*

$$u = \begin{cases} 0 & \text{on } \tilde{\Omega}_1 \cup \tilde{\Omega}_2 \cup \tilde{\Gamma}_1 \cup \tilde{\Gamma}_2 \cup \tilde{\Gamma}_4 \cup \{(\tilde{x},\tilde{y})\} \times [0,\infty),\\ 1 & \text{on } \tilde{\Omega}_3 \cup \tilde{\Omega}_4 \cup \tilde{\Gamma}_3,\end{cases}$$

$$v = \begin{cases} 0 & \text{on } \tilde{\Omega}_1 \cup \tilde{\Omega}_3 \cup \tilde{\Omega}_4 \cup \tilde{\Gamma}_1 \cup \tilde{\Gamma}_3 \cup \tilde{\Gamma}_4 \cup \{(\tilde{x},\tilde{y})\} \times [0,\infty),\\ 1 & \text{on } \tilde{\Omega}_2 \cup \tilde{\Gamma}_2,\end{cases}$$

$$\tau = \begin{cases} 0 & \text{on } \tilde{\Omega}_2 \cup \tilde{\Omega}_3 \cup \tilde{\Omega}_4 \cup \tilde{\Gamma}_2 \cup \tilde{\Gamma}_3,\\ 1 & \text{on } \tilde{\Omega}_1 \cup \tilde{\Gamma}_1 \cup \tilde{\Gamma}_4 \cup \{(\tilde{x},\tilde{y})\} \times [0,\infty).\end{cases}$$

Proof. The payoff function V is given below by giving its restrictions to the regions $\tilde{\Omega}_1$ through $\tilde{\Omega}_4$. V^k is the restriction of V to $\tilde{\Omega}_k$. Each of the functions V^k extends continuously to the boundary of $\tilde{\Omega}_k$, and the extensions agree on the common boundaries of the regions $\tilde{\Omega}_k$. The payoff function is obtained by solving (1.2) (with $t(0) = t$, $x(0) = x$, $y(0) = y$) with the given controls and evaluating (1.1). Let

$$V^1 = \frac{Ax}{\rho}e^{-\rho t} + \frac{By}{\rho}e^{-\rho t},$$

$$V^2 = \frac{x}{\gamma}\log\left(\frac{1-\tilde{z}}{1-y/x}\right)e^{-\rho t} + \frac{1}{\rho}[Ax + Bx\tilde{z}]e^{-\rho t},$$

$$V^3 = \frac{1}{\alpha}\log\left(\frac{1-w(y)}{1-x}\right)e^{-\rho t} + V^2(t, w(y), y),$$

$$V^4 = \frac{1}{\alpha}\log\left(\frac{1-\tilde{x}}{1-x}\right)e^{-\rho t} + \frac{A\tilde{x}}{\rho}e^{-\rho t} + \frac{By}{\rho}e^{-\rho t}.$$

The payoff function is $C^{(1)}$ (but not $C^{(2)}$). The classical verification theorem [7, Thm. 4.4, p. 87] can be easily modified to apply to this infinite horizon problem. Two observations are relevant for this. The payoff function V approaches zero along any trajectory as s approaches ∞. Also, if u, v, τ are admissible controls, (x, y, t) is the corresponding trajectory, and $S > 0$, then

$$\int_S^\infty (Ax\tau + By\tau - u - v)\, e^{-\rho t} ds \leq \int_S^\infty (A+B)\,\tau e^{-\rho t} ds = \frac{A+B}{\rho} e^{-\rho t(S)}.$$

All we need is to verify the Hamilton–Jacobi–Bellman equation on $\tilde{\Omega}_1 \cup \tilde{\Omega}_2 \cup \tilde{\Omega}_3 \cup \tilde{\Omega}_4$, which means that we need to show that

$$P = 0 \quad \text{on } \tilde{\Omega}_3 \cup \tilde{\Omega}_4, \tag{3.2}$$

$$P \leq 0 \quad \text{on } \tilde{\Omega}_1 \cup \tilde{\Omega}_2, \tag{3.3}$$

$$Q = 0 \quad \text{on } \tilde{\Omega}_2, \tag{3.4}$$

$$Q \leq 0 \quad \text{on } \tilde{\Omega}_1 \cup \tilde{\Omega}_3 \cup \tilde{\Omega}_4, \tag{3.5}$$

$$R = 0 \quad \text{on } \tilde{\Omega}_1, \tag{3.6}$$

$$R \leq 0 \quad \text{on } \tilde{\Omega}_2 \cup \tilde{\Omega}_3 \cup \tilde{\Omega}_4. \tag{3.7}$$

We give P, Q, and R by giving their restrictions to each set $\tilde{\Omega}_k$; with P^k denoting the restriction of P to $\tilde{\Omega}_k$, etc.,

$$e^{\rho t}P^1 = \alpha(1-x)\frac{A}{\rho} - 1,$$

$$e^{\rho t}P^2 = \alpha(1-x)\left[\frac{1}{\gamma}\log\left(\frac{1-\tilde{z}}{1-y/x}\right) - \frac{y/x}{\gamma(1-y/x)} + \frac{1}{\rho}[A + B\tilde{z}]\right] - 1,$$

$$e^{\rho t}P^3 = 0,$$

$$e^{\rho t}P^4 = 0,$$

$$e^{\rho t}Q^1 = \frac{B\gamma}{\rho}(1-y/x) - 1,$$

$$e^{\rho t}Q^2 = 0,$$

$$e^{\rho t}Q^3 = \frac{1-y/x}{1-y/w(y)} - 1,$$

$$e^{\rho t}Q^4 = \frac{B\gamma}{\rho}(1-y/x) - 1,$$

$$e^{\rho t}R^1 = 0,$$

$$e^{\rho t}R^2 = -\frac{\rho x}{\gamma}\log\left(\frac{1-\tilde{z}}{1-y/x}\right) + B(y - x\tilde{z}),$$

$$e^{\rho t}R^3 = e^{\rho t}R^2(t, w(y), y) + \frac{\rho}{\alpha}\log\left(\frac{1-x}{1-w(y)}\right) + A(x - w(y)),$$

$$e^{\rho t}R^4 = \frac{\rho}{\alpha}\log\left(\frac{1-x}{1-\tilde{x}}\right) + A(x - \tilde{x}).$$

The derivation of most of these formulas is straightforward. The formula for Q^3 uses the fact that $e^{\rho t}V_x^2(t, w(y), y) = 1/\alpha(1 - w(y))$, which follows from the definition of w and the definition of $\tilde{z}$.

Conditions (3.2), (3.4), and (3.6) are immediate from the above formulas, and (3.5) is easily seen, as well as the fact that $P < 0$ on $\tilde{\Omega}_1$. We can use the definition of $\tilde{z}$ to see that $e^{\rho t}P^2 = \alpha(1 - x)E(x, y)$, and thus the fact that $P < 0$ on $\tilde{\Omega}_2$ follows from Lemma 3.1. To establish (3.7), consider the functions G and H defined on (0,1) by

$$H(z) = \frac{\rho}{\gamma}\log(1 - z) + Bz,$$
$$G(x) = \frac{\rho}{\alpha}\log(1 - x) - A(1 - x).$$

It is easy to see that G is increasing on $(0, \tilde{x}]$ and decreasing on $[\tilde{x}, 1)$ and that H is increasing on $(0, \tilde{z}]$ and decreasing on $[\tilde{z}, 1)$. Since

$$e^{\rho t}R^2(t, x, y) = x[H(y/x) - H(\tilde{z})],$$

it is clear that $R < 0$ on $\tilde{\Omega}_2$. Since

$$e^{\rho t}R^3(t, x, y) = e^{\rho t}R^2(t, w(y), y) + G(x) - G(w(y)),$$

it is clear that $R < 0$ on $\tilde{\Omega}_3$. Since

$$e^{\rho t}R^4(t, x, y) = G(x) - G(\tilde{x}),$$

it is clear that $R < 0$ on $\tilde{\Omega}_4$. This establishes (3.7).

4. Suboptimal controls for $(\mathcal{P}_0)$. It is clear from the solution of $(\mathcal{P})$ that $(\mathcal{P}_0)$ has no solution in the class of locally integrable control functions, unless $(x_0, y_0) \in \Omega_1 \cup \Gamma_1 \cup \Gamma_4$. In other words, J_0 has no maximum for u, v nonnegative and locally integrable unless $(x_0, y_0) \in \Omega_1 \cup \Gamma_1 \cup \Gamma_4$. What we show in this section is that the supremum of J_0 over the class of nonnegative locally integrable controls is equal to the maximum of J. We do this for an initial state in Ω_3, which is the most interesting case. Since the suboptimal controls are for the problem $(\mathcal{P}_0)$, they do not involve the parameter s. However, the parameterization $s \to t(s)$ introduced in the problem $(\mathcal{P})$ permits a visualization of the optimal synthesis that is crucial for the design of suboptimal locally integrable controls u_ε and v_ε for the problem $(\mathcal{P}_0)$. To design suboptimal controls for the problem $(\mathcal{P}_0)$, we approximate first the jump from (x_0, y_0) to $(w(y_0), y_0)$ and then the jump from $(w(y_0), y_0)$ to $(w(y_0), \tilde{z}w(y_0))$.

Let $(x_0, y_0) \in \Omega_3$, and let

$$u_\varepsilon^1(t) = \begin{cases} \dfrac{1}{\varepsilon\alpha}\log\left(\dfrac{1 - x_0}{1 - w(y_0)}\right) & \text{for } 0 < t < \varepsilon, \\ 0 & \text{for } \varepsilon < t, \end{cases}$$

$$v_\varepsilon^1(t) = \begin{cases} 0 & \text{for } 0 < t < \varepsilon, \\ \dfrac{w(y_0)}{\varepsilon\gamma}\log\left(\dfrac{1 - y_0/w(y_0)}{1 - \tilde{z}}\right) & \text{for } \varepsilon < t < 2\varepsilon, \\ 0 & \text{for } 2\varepsilon < t. \end{cases}$$

Then solving

$$\frac{d}{dt}x_\varepsilon^1 = \alpha(1-x_\varepsilon^1)u_\varepsilon^1, \qquad x_\varepsilon^1(0) = x_0$$
$$\frac{d}{dt}y_\varepsilon^1 = \gamma(1-y_\varepsilon^1/x_\varepsilon^1)v_\varepsilon^1, \qquad y_\varepsilon^1(0) = y_0,$$

we find that

$$x_\varepsilon^1(t) = \begin{cases} 1-(1-x_0)^{1-t/\varepsilon}(1-w(y_0))^{t/\varepsilon} & \text{for } 0 \le t \le \varepsilon, \\ w(y_0) & \text{for } \varepsilon \le t, \end{cases}$$

$$y_\varepsilon^1(t) = \begin{cases} y_0 & \text{for } 0 \le t \le \varepsilon, \\ w(y_0) - w(y_0)(1-y_0/w(y_0))^{(2\varepsilon-t)/\varepsilon}(1-\tilde{z})^{(t-\varepsilon)/\varepsilon} & \text{for } \varepsilon \le t \le 2\varepsilon, \\ w(y_0)\tilde{z} & \text{for } 2\varepsilon \le t. \end{cases}$$

Letting $x^1(t) = \lim_{\varepsilon\to 0} x_\varepsilon^1(t)$, $y^1(t) = \lim_{\varepsilon\to 0} y_\varepsilon^1(t)$, it is clear that

$$x^1(t) = \begin{cases} x_0 & \text{for } t = 0, \\ w(y_0) & \text{for } t > 0, \end{cases}$$

$$y^1(t) = \begin{cases} y_0 & \text{for } t = 0, \\ w(y_0)\tilde{z} & \text{for } t > 0. \end{cases}$$

The Lebesgue dominated convergence theorem implies that

$$\lim_{\varepsilon\to 0}\int_0^\infty [Ax_\varepsilon^1(t) + By_\varepsilon^1(t)]e^{-\rho t}dt = [Aw(y_0) + Bw(y_0)\tilde{z}]/\rho,$$

and direct calculation implies that

$$\lim_{\varepsilon\to 0}\int_0^\infty [-u_\varepsilon^1(t) - v_\varepsilon^1(t)]e^{-\rho t}dt = \frac{1}{\alpha}\log\left(\frac{1-w(y_0)}{1-x_0}\right) + \frac{w(y_0)}{\gamma}\log\left(\frac{1-\tilde{z}}{1-y_0/w(y_0)}\right).$$

Thus

$$J_\varepsilon^1 = \int_0^\infty [Ax_\varepsilon^1(t) + By_\varepsilon^1(t) - u_\varepsilon^1(t) - v_\varepsilon^1(t)]e^{-\rho t}dt$$

converges to $V(0, x_0, y_0)$ as $\varepsilon \to 0$.

On the other hand, the controls given below, which look equivalent if one only considers the problem $(\mathcal{P}_0)$, are not suboptimal. Let

$$u_\varepsilon^2(t) = \begin{cases} 0 & \text{for } 0 < t < \varepsilon, \\ \dfrac{1}{\varepsilon\alpha}\log\left(\dfrac{1-x_0}{1-w(y_0)}\right) & \text{for } \varepsilon < t < 2\varepsilon, \\ 0 & \text{for } 2\varepsilon < t, \end{cases}$$

$$v_\varepsilon^2(t) = \begin{cases} \dfrac{x_0}{\varepsilon\gamma}\log\left(\dfrac{1-y_0/x_0}{1-\tilde{z}}\right) & \text{for } 0 < t < \varepsilon, \\ 0 & \text{for } \varepsilon < t. \end{cases}$$

These controls also produce trajectories $x_\varepsilon^2(t)$, $y_\varepsilon^2(t)$ that converge to $x^1(t)$, $y^1(t)$, respectively, as $\varepsilon \to 0$. However,

$$\lim_{\varepsilon\to 0}\int_0^\infty [-u_\varepsilon^2(t) - v_\varepsilon^2(t)]e^{-\rho t}dt = \frac{1}{\alpha}\log\left(\frac{1-w(y_0)}{1-x_0}\right) + \frac{x_0}{\gamma}\log\left(\frac{1-\tilde{z}}{1-y_0/x_0}\right).$$

Therefore,

$$J_\varepsilon^2 = \int_0^\infty [Ax_\varepsilon^2(t) + By_\varepsilon^2(t) - u_\varepsilon^2(t) - v_\varepsilon^2(t)]e^{-\rho t}dt$$

does not converge to $V(0, x_0, y_0)$ as $\varepsilon \to 0$.

Remark. There is an economic explanation why, for an initial state in Ω_3, we should increase x before y to approach the supremum of J_0. The reason is that the advertising of the second product is more effective when the market share of the first product is greater. This is intuitively correct, and it is reflected in the model.

REFERENCES

[1] A. BENSOUSSAN AND J.-L. LIONS, *Impulsive Control and Quasivariational Inequalities*, Bordas, Paris, 1984.

[2] A. BLAQUIÈRE, *Impulsive optimal control with finite or infinite time horizon*, J. Optim. Theory Appl., 46 (1985), pp. 431–439.

[3] A. BRESSAN AND F. RAMPAZZO, *On differential systems with vector-valued impulsive controls*, Boll. Un. Mat. Ital., 3 (1988), pp. 641–656.

[4] J. R. DORROH AND G. FERREYRA, *Optimal advertising in exponentially decaying markets*, J. Optim. Theory Appl., 79 (1993), pp. 219–236.

[5] ———, *Optimal advertising in growing-stabilizing markets*, Optim. Control Appl. Meth., 14 (1993), pp. 221–228.

[6] G. FERREYRA, *The optimal control problem for the Vidale–Wolfe advertising model revisited*, Optim. Control Appl. Meth., 11 (1990), pp. 363–368.

[7] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[8] A. J. KRENER, *The high order maximal principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 250–293.

[9] S. P. SETHI, *Optimal control of the Vidale–Wolfe model*, Oper. Res., 21 (1973), pp. 998–1013.

[10] ———, *Dynamic optimal control models in advertising: A survey*, SIAM Review, 19 (1977), pp. 685–725.

[11] H. J. SUSSMANN, *Semigroup representations, bilinear approximation of input-output maps, and generalized inputs*, in Mathematical Systems Theory, Lecture Notes in Econ. and Math Sys., Proc. Int. Symp., Udine, 131 (1976), pp. 172–191.

[12] M. L. VIDALE AND H. B. WOLFE, *An operations research study of sales response to advertising*, Oper. Res., 5 (1957), pp. 370–381.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1332–1363, September 1994

008

NUMERICAL APPROXIMATIONS FOR HEREDITARY SYSTEMS WITH INPUT AND OUTPUT DELAYS: CONVERGENCE RESULTS AND CONVERGENCE RATES*

A. MANITIUS† AND H. T. TRAN‡

Abstract. In this paper, the averaging approximation scheme for linear retarded functional differential equations with delays in control and observation is considered in the context of the state space theory developed by Pritchard and Salamon [*SIAM J. Control Optim.*, 25 (1987), pp. 121–144]. Using known results from linear semigroup theory, convergence and estimate of convergence rate of the approximating semigroups are established. These extend results due to Banks and Burns [*SIAM J. Control Optim.*, 16 (1978), pp. 169–208] and Lasiecka and Manitius [*SIAM J. Numer. Anal.*, 25 (1988), pp. 883–907] on hereditary systems with delays in state, to the case when delays in control and observation are included. The main difference from the case when delays in input and output are excluded is that unbounded input and output operators must be dealt with in the abstract formulation. Moreover, in the presence of the unboundedness of the input and output operators, new convergence results of the state solutions and the output are also obtained.

Key words. functional differential equations, unbounded input and output operators, averaging approximation, convergence, convergence rate

AMS subject classifications. 34K30, 34K35, 65J10

1. Introduction.

The object of this paper is to extend previous results on the averaging approximation scheme for linear retarded functional differential equations (RFDE) to the case when general delays in control and observation are included.

The averaging approximation scheme has been invented (in a different context) by Soviet authors Repin [R1] and Krasovskii [K2] in the early sixties. A detailed historical review can be found in Banks and Burns [B1]. In this paper by Banks and Burns, precise statements of convergence results and applications to open-loop control problems on finite time intervals were given for the first time. Later, Gibson [G1] showed that for the finite time interval, convergence result of the feedback control laws can be obtained using the averaging scheme. In the case of infinite time horizon, it is important to know whether the approximation scheme preserves, uniformly with respect to the discretization mesh, the asymptotic stability of the original systems. This stability preservation property of the averaging scheme was later proved by Salamon [S1]. Later, Lasiecka and Manitius [L1] established a stronger version of Salamon's result. They showed that in the case of RFDE and averaging approximations one obtains uniform differentiability of the approximating semigroups. As is shown in [L1], this fact has far reaching consequences. First, when the original semigroup is

*Received by the editors February 7, 1989; accepted for publication (in revised form) March 19, 1993.

†Department of Electrical & Computer Engineering, George Mason University, Fairfax, Virginia, 22030 (amanitiu@bass.gmu.edu). This research was supported in part by the National Science Foundation under grant MCS-8201719 and the Air Force Office of Scientific Research under contract ISSA-8400052.

‡Center for Research in Scientific Computation, Box 8205, North Carolina State University, Raleigh, North Carolina, 27695-8205 (tran@control.math.ncsu.edu). This research was supported in part by the National Science Foundation under grant MCS-8504316 and the Air Force Office of Scientific Research under contract F49620-86-C-0111. Part of this research was carried out while this author was at the Center for Control Sciences, Division of Applied Mathematics, Brown University, Providence, Rhode Island.

stable it implies almost immediately that the approximating semigroups are uniformly exponentially stable with a decay rate that can be made arbitrarily close (as $N \to \infty$) to the decay rate of the original equations. Second, and more important, this makes it possible to obtain convergence rates of the homogeneous solution. In [L1], precise statements of the optimal convergence estimates and their dependence on the initial data and system parameters were presented for the first time. Moreover, they showed that convergence of the approximating semigroups in the uniform operator topology can be obtained for sufficiently large time. The strong convergence of approximating semigroups has been proved via the Trotter–Kato approximation theorem in [B1].

A problem that has not yet been considered is whether analogous results can be developed for averaging approximation of RFDE with general delays in control and observation. In this paper we extend the results in Banks and Burns [B1] and Lasiecka and Manitius [L1] to retarded systems with unbounded input and output operators considered in the state space framework developed by Pritchard and Salamon [P2]. Moreover, convergence properties of the approximating output operators and the approximating state solution corresponding to the nonhomogeneous problem were also analyzed. The convergence results of the approximating output operators were obtained based on detailed analyses of the convergence and bounds of the composition of the unbounded input and output operators with the well-known operators characterizing the resolvents. Finally, we noted that the development presented in this paper gives a basis for numerical investigations of control problems, in particular using the operator Riccati equations. In particular, recently Ito and Tran [I3] have developed a general approximation framework for the numerical treatment of Riccati operators for a class of linear infinite-dimensional systems with unbounded input and output operators. As a simple application, the convergence theory was applied to the linear quadratic control problem for linear retarded systems with point delays in the controls. Later, Tran [T1] provided numerical evidence demonstrating the feasibility of the general approach in the context of feedback control of retarded systems with delay in the control using the averaging approximation scheme.

The organization of the paper is as follows. We first review in §2 a general setting for our system approximation problem in some appropriate Hilbert space. Much of the material in this section comes from a recent article by Pritchard and Salamon [P2] in which a unified theory of control systems with delays in state, control and observation based on C_0-semigroups is provided. Within the framework in §2, the averaging approximating scheme for linear RFDE with general delays in control and observation is developed in §3.1. Section 3.2 contains results on uniform differentiability of the approximating semigroups and uniform exponential stability of these semigroups when the original semigroup is exponentially stable. The main part of the paper is §4, where we give precise statements of convergence results as well as estimates of convergence rates with their proofs. Finally, §5 gives our concluding remarks.

2. Linear retarded systems with delays in input and output. In this section we define the type of hereditary systems to be considered in this paper and review some well-known results on the state space description of linear RFDE with general delays in input and output in terms of semigroups and evolution equations. Much of the material in this section comes from Pritchard and Salamon [P2] but must be presented to give a clear discussion of the developments in the subsequent sections.

We consider the linear RFDE of the form

$$\dot{x}(t) = Lx_t + \hat{B}u_t, \qquad t \geq 0, \tag{1a}$$

$$y(t) = \hat{C}x_t, \tag{1b}$$

where $x(t) \in R^n$, $u(t) \in R^m$, $y(t) \in R^p$ and x_t, u_t are defined by $x_t(\theta) = x(t+\theta)$, $u_t(\theta) = u(t+\theta)$ for $\theta \in [-h, 0]$, $0 < h < \infty$. The bounded linear operators $L, \hat{B}$, and $\hat{C}$ are mappings from spaces of real vector-valued continuous functions to real finite-dimensional spaces and are given by

$$\begin{aligned} L\phi &= \int_{-h}^{0} d\eta(\theta)\phi(\theta) + A_0\phi(0) \\ &= \sum_{i=0}^{q} A_i\phi(-h_i) + \int_{-h}^{0} A_{01}(\theta)\phi(\theta)d\theta, \\ \hat{C}\phi &= \int_{-h}^{0} d\gamma(\theta)\phi(\theta) + C_0\phi(0) \\ &= \sum_{i=0}^{q} C_i\phi(-h_i) + \int_{-h}^{0} C_{01}(\theta)\phi(\theta)d\theta, \qquad \phi \in C(-h, 0; R^n), \\ \hat{B}\xi &= \int_{-h}^{0} d\beta(\theta)\xi(\theta) + B_0\xi(0) \\ &= \sum_{i=0}^{q} B_i\xi(-h_i) + \int_{-h}^{0} B_{01}(\theta)\xi(\theta)d\theta, \qquad \xi \in C(-h, 0; R^m), \end{aligned}$$

where $0 = h_0 < h_1 < \cdots < h_q = h$, $A_i \in R^{n\times n}$, $B_i \in R^{n\times m}$, $C_i \in R^{p\times n}$, $i = 0, \ldots, q$, and $A_{01}(\cdot) \in L^2(-h, 0; R^{n\times n})$, $B_{01}(\cdot) \in L^2(-h, 0; R^{n\times m})$, $C_{01}(\cdot) \in L^2(-h, 0; R^{p\times n})$. It is clear that the matrix function $\eta : [-h, 0] \to R^{n\times n}$ of bounded variation is of the form

$$\eta(\theta) = -\sum_{i=1}^{q} A_i\chi_{(-\infty,-h_i]}(\theta) - \int_{\theta}^{0} A_{01}(s)ds, \qquad \theta \in [-h, 0],$$

where χ_I denotes the characteristic function of the interval I. Similarly, we can define the matrix functions β and γ. Let

$$X = R^n \times L^2(-h, 0; R^n) \times L^2(-h, 0; R^m)$$

Consider the solution of (1) with the initial data

$$x(0) = \phi^0, \quad x(\theta) = \phi^1(\theta), \quad u(\theta) = \phi^2(\theta), \quad \theta \in [-h, 0),$$

where $\phi = (\phi^0, \phi^1, \phi^2) \in X$, and $u(\cdot) \in L^2_{\text{loc}}(-h, \infty; R^m)$. Routine extensions of standard results (see, e.g., [D1], [P2], [S2]) guarantee existence, uniqueness, and continuous dependence of the solution on the initial data $\phi \in X$. This motivates the "natural" definition of the state of system (1) at time $t \geq 0$ to be the triple $z(t) = \big(x(t), x_t, u_t\big) \in X$, and thus justifies the choice of X as the state space for (1).

In problems involving Laplace and Fourier transforms and eigenvalue analysis of (1), we will need to use the natural complex extension of X, i.e., the space

$$\tilde{X} = \mathbf{C}^n \times L^2(-h, 0; \mathbf{C}^n) \times L^2(-h, 0; \mathbf{C}^m).$$

This being understood, we will use notation X for both the state space and its complex extension without an explicit mention each time.

To accomodate the effect of delays in the observation, we will also use the space

$$X^T = R^n \times L^2(-h, 0; R^n) \times L^2(-h, 0; R^p),$$

or its complex extension (the integer p is the dimension of the output vector).

The evolution of the state $z(t)$ in time is governed by the variation-of-constants formula [P2]

$$z(t) = S(t)\phi + \int_0^t S(t-\theta)Bu(\theta)d\theta, \qquad t \geq 0 \tag{2}$$

with the corresponding output $y(t)$ given by

$$y(t) = Cz(t), \qquad t \geq 0.$$

To allow for possible unboundedness of the input and output operators, B and C, we assume that $B \in L(R^m, V)$, $C \in L(W, R^p)$ where W and V are Hilbert spaces such that

$$W \subset X \subset V$$

with continuous dense injections $i : W \to X$ and $j : X \to V$. Moreover, the input operator $B : R^m \to V$ is given by [I1]

$$Bu = (B_0 u, 0, T_{\delta_0}(u)), \qquad u \in R^m,$$

where the generalized function T_{δ_0} is the Dirac distribution concentrated at the point $0 \in R^m$, and the output operator $C : W \to R^p$ is given by

$$C\phi = \hat{C}\phi^1, \qquad \phi \in W.$$

The C_0-semigroup $S(\cdot)$ corresponding to the free motion of the system, $u(t) = 0$ for $t \geq 0$, i.e., the bounded linear operator $S(t) : X \to X$ is defined by

$$S(t)\phi = (x(t), x_t, u_t), \qquad \phi \in X.$$

The infinitesimal generator A of $S(\cdot)$ is given by

$$\begin{aligned} \mathrm{dom}A = \Big\{ & \phi \in X / \phi^1 \in W^{1,2}(-h, 0; R^n), \\ & \phi^2 \in W^{1,2}(-h, 0; R^m), \quad \phi^1(0) = \phi^0, \quad \phi^2(0) = 0 \Big\} \\ A\phi = (L\phi^1 + \hat{B}\phi^2 - B_0\phi^2(0), \dot{\phi}^1, \dot{\phi}^2), & \qquad \phi \in \mathrm{dom}A, \end{aligned} \tag{3}$$

[P2]. To make (2) precise and to allow for trajectories in all three spaces W, X, and V, we assume that $S(\cdot)$ is also a strongly continuous semigroup on W and V and that the following hypotheses are satisfied.

Let $k = ji$ be the continuous dense injection from W into V.

(H1) For any $u(\cdot) \in L^2(0, T; R^m)$,

$$\int_0^T S(T-\theta)Bu(\theta)d\theta \in kW$$

and there exists a positive constant b such that

$$\left\| k^{-1} \int_0^T S(T-\theta)Bu(\theta)d\theta \right\|_W \leq b\|u\|_{L^2(0,T;R^m)}.$$

(H2) There exists a positive constant c such that

$$\left\| CS(\cdot)\phi \right\|_{L^2(0,T;R^p)} \leq c\|k\phi\|_V, \qquad \phi \in W.$$

The dual statements of (H1) and (H2) are as follows.

(H1*) For every $x \in V^*$,

$$\left\| B^*S^*(T-\cdot)x \right\|_{L^2(0,T;R^m)} \leq b\|k^*x\|_{W^*}.$$

(H2*) For every $y(\cdot) \in L^2(0,T;R^p)$

$$\int_0^T S^*(T-\theta)C^*y(\theta)d\theta \in k^*V^*$$

and

$$\left\| (k^*)^{-1} \int_0^T S^*(T-\theta)C^*y(\theta)d\theta \right\|_{V^*} \leq c\|y\|_{L^2(0,T;R^p)}.$$

The function $z(t)$ as defined in (2) is a mild solution of the abstract Cauchy problem

$$(\Sigma) \qquad \begin{aligned} \frac{d}{dt}z(t) &= Az(t) + Bu(t), \\ y(t) &= Cz(t) \end{aligned}$$

in the Hilbert spaces W, respectively, V [P2]. We note here that, if we consider the Cauchy problem (Σ) in the state space W, then the ouput operator C will be bounded but the input operator B may be unbounded. Nevertheless, the solution of (Σ) in W is well defined, since B satisfies hypothesis (H1). Conversely, if we consider the Cauchy problem (Σ) in the bigger state space V, then B is bounded but C may be unbounded. Nevertheless, the output is well defined, since C satisfies hypothesis (H2).

Before closing this section, we will describe the structural operator F associated with (1) and the resolvent operator $R(\lambda, A)$ of the generator A, both being generalizations of the operators F and $R(\lambda, A)$ described in [D2].

Let $F : X \to X^{T^*}$ be defined by

$$[F\phi]^0 = \phi^0, \qquad [F\phi]^1(\theta) = \int_{-h}^{\theta} d\eta(s)\phi^1(s-\theta) + \int_{-h}^{\theta} d\beta(s)\phi^2(s-\theta),$$
$$[F\phi]^2(\theta) = \int_{-h}^{\theta} d\gamma(s)\phi^1(s-\theta), \qquad \theta \in [-h, 0].$$

This definition corresponds to the results of [P2]. We note that in the work of Delfour and Karrakchou [D1] the definition of F does not include the term with $\beta(\cdot)$, which instead is added as an exogenous component of the system state. Both choices of the structural operator are valid; ours is more suitable to describe the duality between the original system (1) and the "transposed" system (with input $v(t)$ and output $q(t)$)

$$(4) \qquad \begin{aligned} \frac{d}{dt}w(t) &= L^T w_t + \hat{C}^T v_t, \\ q(t) &= \hat{B}^T w_t, \end{aligned}$$

and the resulting duality by transposition between the finite-dimensional approximations of (1) and (4). The structural operator F has the same basic properties as the one in [D2]. In particular, its dual operator $F^* : X^T \to X^*$ is

$$
\begin{aligned}
[F^*\psi]^0 = \psi^0, \qquad [F^*\psi]^1(\theta) &= \int_{-h}^{\theta} d\eta^T(s)\psi^1(s-\theta) + \int_{-h}^{\theta} d\gamma^T(s)\psi^2(s-\theta) \\
[F^*\psi]^2(\theta) &= \int_{-h}^{\theta} d\beta^T(s)\psi^1(s-\theta), \qquad \theta \in [-h, 0].
\end{aligned} \tag{5}
$$

Remark 2.1. The structural operator is extremely useful in numerical approximations for two reasons: (a) it plays a crucial role in the proof of convergence rates; see the proof of Theorem 4.3, estimate of $\|T_3\|$, and (b) it often enables us to make a dramatic reduction in the dimensionality of approximating systems (see [T1]).

In the remainder of this section, X and X^T will be their complex extensions. For $\lambda \in \mathbf{C}$, let the "exponential map" $E_\lambda : \mathbf{C}^n \to X$ be defined by

$$
[E_\lambda x]^0 = x, \quad [E_\lambda x]^1(\theta) = e^{\lambda\theta}x, \quad [E_\lambda x]^2(\theta) = 0, \qquad x \in \mathbf{C}^n.
$$

An analogous definition holds for $E_\lambda^T : \mathbf{C}^n \to X^T$. The dual operators $E_\lambda^* : X^* \to \mathbf{C}^n$ and $E_\lambda^{T^*} : X^{T^*} \to \mathbf{C}^n$ are given by

$$
\begin{aligned}
E_\lambda^*\phi &= \phi^0 + \int_{-h}^{0} e^{\bar\lambda\theta}\phi^1(\theta)d\theta, \\
E_\lambda^{T^*}\psi &= \psi^0 + \int_{-h}^{0} e^{\bar\lambda\theta}\psi^1(\theta)d\theta.
\end{aligned}
$$

Note that the composition of maps $E_\lambda^{T^*}F : X \to \mathbf{C}^n$ is well defined and does not depend on $\gamma(\cdot)$. That is,

$$
E_\lambda^{T^*}F\phi = \phi^0 + \int_{-h}^{0} e^{\bar\lambda\theta}(F\phi)^1(\theta)d\theta.
$$

The linear operator $T_\lambda : X \to X$ is defined by

$$
\begin{aligned}
[T_\lambda\phi]^0 = 0, \quad [T_\lambda\phi]^1(\theta) &= \int_{\theta}^{0} e^{\lambda(\theta-s)}\phi^1(s)ds, \\
[T_\lambda\phi]^2(\theta) &= \int_{\theta}^{0} e^{\lambda(\theta-s)}\phi^2(s)ds.
\end{aligned}
$$

Define $\Delta(\lambda) = \lambda I - L(e^{\lambda\cdot})$ and the resolvent set $\rho(A) = \{\lambda/\det\,\Delta(\lambda) \neq 0\}$. It is easy to see that the spectrum of A is $\sigma(A) = \{\lambda/\det\Delta(\lambda) = 0\}$. Moreover, by using the same arguments as in [D2] we can also show that the resolvent $R(\lambda, A) = (\lambda I - A)^{-1}$ is given by

$$
R(\lambda, A) = E_\lambda\Delta^{-1}(\lambda)E_{\bar\lambda}^{T^*}F + T_\lambda, \quad \lambda \in \rho(A).
$$

3. Finite-dimensional approximations for retarded systems with delays in input and output. The object of this section is the approximation of solutions of the RFDE (2.1) via use of approximate solutions of the abstract Cauchy problem (Σ). To approximate solutions of the abstract evolution equation in $X = R^n \times L^2 \times L^2$, we

will use subspaces $Z^N \subseteq L^2$ being sets of piecewise constant functions on the delay interval $[-h, 0]$ and thereby obtain the semidiscrete finite-difference scheme widely known in the literature as the averaging approximation scheme. Based on some careful analysis of the structure of the approximating systems, it will be shown in §3.2 that the approximating semigroups are all uniformly differentiable with respect to the index N determining the mesh size. These are extended results of Lasiecka and Manitius [L1] to the case when delays are included in the control and observation. A necessary condition for an approximation scheme to have the uniform differentiability characteristic is that the location of the spectrum must be contained in some logarithmic sector. Such a property is by no means obvious and not all approximation schemes have this property. For example, it is shown by numerical calculations [L1] that the spectra of approximating generators corresponding to linear splines contain eigenvalues of large modulus located arbitrarily close, as $N \to \infty$, to the imaginary axis. Thus we cannot hope to prove the same property for the spline approximation scheme in [B2].

3.1. Averaging approximation. For every positive integer N, we define the finite-dimensional linear subspace X^N of X by

$$X^N = \Big\{ \phi \in X / \phi^0 = z_0 \in R^n; \phi^1 = \sum_{j=1}^{N} z_j \chi_j, \quad z_j \in R^n; \\ \phi^2 = \sum_{j=1}^{N} v_j \chi_j, \quad v_j \in R^m \Big\},$$

where χ_j denote the characteristic function of $[t_j, t_{j-1})$ for $j = 1, 2, \ldots, N$ and $t_j = -jh/N, j = 0, 1, \ldots, N$. This subspace can be made isometrically isomorphic to the Euclidean space $R^{n(N+1)+mN}$ by means of the embedding $\wr^N : R^{n(N+1)+mN} \to X^N$, which associates with every $z = \operatorname{col}(z_0, \ldots, z_N, v_1, \ldots, v_N) \in R^{n(N+1)+mN}$, where $z_i \in R^n$ and $v_i \in R^m$, the triple

$$\begin{aligned} &[\wr^N z]^0 = z_0, \\ &[\wr^N z]^1(\theta) = z_i, \quad \theta \in [t_i, t_{i-1}), \qquad i = 1, \ldots, N, \\ &[\wr^N z]^2(\theta) = v_i, \quad \theta \in [t_i, t_{i-1}), \qquad i = 1, \ldots, N. \end{aligned}$$

On $R^{n(N+1)+mN}$, we define the induced inner product

$$\langle z, y \rangle_N = z^T Q^N y, \qquad z, y \in R^{n(N+1)+mN},$$

where $Q^N = \operatorname{diag}(Q_{11}^N, Q_{22}^N)$ is an $n(N+1)+mN \times n(N+1)+mN$ matrix and $Q_{11}^N = \operatorname{diag}(I_n, \frac{h}{N} I^n, \ldots, \frac{h}{N} I_n)$, $Q_{22}^N = \operatorname{diag}(\frac{h}{N} I_m, \ldots, \frac{h}{N} I_m)$, are $n(N+1) \times n(N+1)$ and $mN \times mN$ matrices, respectively. The corresponding vector and matrix norms will be denoted by $\| \cdot \|_N$. It can be shown that the map $\pi^N : X \to R^{n(N+1)+mN}$, an extension of the dual map of $\wr^N$, is given by

$$\pi^N \phi = \operatorname{col}(z_0, \ldots, z_N, v_1, \ldots, v_N),$$

where

$$z_0 = \phi^0, \quad z_j = \frac{N}{h} \int_{t_j}^{t_{j-1}} \phi^1(\theta) d\theta, \quad v_j = \frac{N}{h} \int_{t_j}^{t_{j-1}} \phi^2(\theta) d\theta, \qquad j = 1, \ldots, N.$$

It is clear that $\iota^N\pi^N = P^N$ is an orthogonal projection of X onto X^N and $\pi^N\iota^N = I$.

We now define the approximating formulas of the generator A, the input operator B, and the output operator C. First we introduce the following matrices:

$$A_j^N = \lim_{\theta \uparrow t_j} \left[\eta\left(\theta + \frac{h}{N}\right) - \eta(\theta)\right], \qquad j = 1, 2, \ldots, N,$$

$$B_j^N = \lim_{\theta \uparrow t_j} \left[\beta\left(\theta + \frac{h}{N}\right) - \beta(\theta)\right], \qquad j = 1, 2, \ldots, N,$$

$$C_j^N = \lim_{\theta \uparrow t_j} \left[\gamma\left(\theta + \frac{h}{N}\right) - \gamma(\theta)\right], \qquad j = 1, 2, \ldots, N.$$

For $\phi \in X$, let $\pi^N\phi = z = \operatorname{col}(z_0, \ldots, z_N, v_1, \ldots, v_N) \in R^{n(N+1)+mN}$ and define the linear maps

$$L_1^N\pi^N\phi = A_0 z_0 + \sum_{j=1}^{N} A_j^N z_j + \sum_{j=1}^{N} B_j^N v_j,$$

$$L_2^N\pi^N\phi = C_0 z_0 + \sum_{j=1}^{N} C_j^N z_j,$$

$$\nabla_1 P^N\phi = \sum_{j=1}^{N} \frac{N}{h}(z_{j-1} - z_j)\chi_j,$$

$$\nabla_2 P^N\phi = \sum_{j=1}^{N} \frac{N}{h}(v_{j-1} - v_j)\chi_j,$$

where we define $v_0 = 0$. The approximating operators $A^N : X \to X^N$ are then defined by

$$A^N\phi = (L_1^N\pi^N\phi, \nabla_1 P^N\phi, \nabla_2 P^N\phi).$$

The approximating input operators $B^N : R^m \to X^N$ are defined by

$$B^N u = (B_0 u, 0, \tfrac{N}{h}\chi_1 u).$$

Finally, the approximating output operators $C^N : X \to R^p$ are defined by

$$C^N\phi = L_2^N\pi^N\phi.$$

Let $S^N(t)$ denote the semigroups generated by A^N on X. The following theorem concerns the important question of stability of the averaging approximations for (1).

THEOREM 3.1. *There exists constants M and ω independent of N such that*

$$\|e^{\pi^N A^N \iota^N t}\|_N \leq M e^{\omega t}.$$

Proof. The proof is based on the well-known Gronwall's inequality and the use of weighting functions in L_2-type norms as suggested in [B1].

Let N be sufficiently large so that the points $-h_i$ (corresponding to the delays), $i = 1, \ldots, q$, lie in distinct intervals in the partition of $[-h, 0]$ by $\{t_j\}$. Let $J^N = \{j_1, j_2, \ldots, j_q\}$ be the subset of indices in $\{1, 2, \ldots, N\}$ such that $-h_i \in [t_{j_i}, t_{j_i - 1}), i = 1, \ldots, q$. Following the notation of Banks and Burns [B1], we define the piecewise constant function τ_N, for each fixed N, on $[-h, 0]$ by

$$\tau_N(\theta) = a_j^N, \quad \theta \in [t_j, t_{j-1}), \qquad j = 1, 2, \ldots, N,$$

where the a_j^N's are defined recursively by

$$a_{N+1}^N = 1,$$
$$a_j^N = \begin{cases} a_{j+1}^N + 1 & \text{if } j \in J^N \\ a_{j+1}^N & \text{if } j \notin J^N \end{cases}, \qquad j = N, N-1, \ldots, 1.$$

Let $\langle .,. \rangle_{\tau^N}$ and $\| \cdot \|_{\tau^N}$ denote the induced inner product and Euclidean norm on $R^{n(N+1)+mN}$ using the weighting function $\tau_N(\theta)$, i.e.,

$$\langle z, y \rangle_{\tau^N} = z^T Q_\tau^N y, \qquad z, y \in R^{n(N+1)+mN},$$

where $Q_\tau^N = \text{diag}\,(Q_{\tau_{11}}^N, Q_{\tau_{22}}^N)$ of dimension $n(N+1)+mN \times n(N+1)+mN$ and $Q_{\tau_{11}}^N = \text{diag}\,(I_n, \frac{h}{N} a_1^N I_n, \ldots, \frac{h}{N} a_N^N I_n), Q_{\tau_{22}}^N = \text{diag}\,(\frac{h}{N} a_1^N I_m, \ldots, \frac{h}{N} a_N^N I_m)$ of dimension $n(N+1) \times n(N+1)$ and $mN \times mN$, respectively.

For $z \in R^{n(N+1)+mN}$, let

$$e^{\pi^N A^N \wr^N t} z = z(t) = \text{col}\,(z_0(t), \ldots, z_N(t), v_1(t), \ldots, v_N(t)).$$

Then

$$\begin{aligned} \frac{d}{dt} \|z(t)\|_{\tau^N}^2 &= \frac{d}{dt}[z^T(t) Q_\tau^N z(t)] \\ &= 2 z^T(t) Q_\tau^N \pi^N A^N \wr^N z(t) \\ &= 2 z_0^T(t)[A_0 z_0(t) + \sum_{j=1}^N A_j^N z_j(t) + \sum_{j=1}^N B_j^N v_j(t)] \\ &\quad + 2 \sum_{j=1}^N z_j^T(t)[z_{j-1}(t) - z_j(t)] a_j^N - 2 v_1^T(t) v_1(t) a_1^N \\ &\quad + 2 \sum_{j=2}^N v_j^T(t)[v_{j-1}(t) - v_j(t)] a_j^N. \end{aligned}$$

We now obtain the following bounds for the right-hand side terms:

$$\begin{aligned} \text{(i)} \quad 2 \sum_{j=1}^N z_j^T(t)[z_{j-1}(t) - z_j(t)] a_j^N &\le -2 \sum_{j=1}^N |z_j(t)|^2 a_j^N \\ &\quad + \sum_{j=1}^N |z_j(t)|^2 a_j^N + \sum_{j=1}^N |z_{j-1}(t)|^2 a_j^N \\ &= -\sum_{j=1}^N |z_j(t)|^2 a_j^N + \sum_{j=0}^{N-1} |z_j(t)|^2 a_{j+1}^N \\ &\le |z_0(t)|^2 a_1^N + \sum_{j=1}^N (a_{j+1}^N - a_j^N) |z_j(t)|^2. \end{aligned}$$

Similarly,

$$\text{(ii)} \quad -2 v_1^T(t) v_1(t) a_1^N + 2 \sum_{j=2}^N v_j^T(t)[v_{j-1}(t) - v_j(t)] a_j^N \le \sum_{j=1}^N (a_{j+1}^N - a_j^N) |v_j(t)|^2.$$

(iii) Define $(A_{01})_i = \frac{N}{h}\int_{t_i}^{t_{i-1}} A_{01}(\theta)d\theta, i = 1,\dots,N$. Then

$$2z_0^T(t)[A_0z_0(t) + \sum_{j=1}^{N} A_j^N z_j(t)] = 2z_0^T(t)[A_0z_0(t) - \sum_{i=1}^{q} A_i z_{j_i}(t) + \sum_{i=1}^{N} \frac{h}{N}(A_{01})_i z_i(t)]$$

$$\leq (1 + \sum_{i=0}^{q} |A_i|^2)|z_0(t)|^2 + \sum_{i=1}^{q} |z_{j_i}(t)|^2 + \frac{h}{N}\sum_{i=1}^{N} |(A_{01})_i|^2|z_0(t)|^2$$

$$+ \frac{h}{N}\sum_{i=1}^{N} |z_i(t)|^2$$

$$\leq (1 + \sum_{i=0}^{q} |A_i|^2 + \|A_{01}\|_{L^2}^2)|z_0(t)|^2 + \frac{h}{N}\sum_{i=1}^{N} |z_i(t)|^2 + \sum_{i=1}^{q} |z_{j_i}(t)|^2,$$

where we have used Schwartz inequality. Similarly,

$$\text{(iv)} \qquad 2z_0^T(t)\sum_{j=1}^{N} B_j^N v_j(t) \leq \left(\sum_{i=1}^{q} |B_i|^2 + \|B_{01}\|_{L^2}^2\right)|z_0(t)|^2 + \frac{h}{N}\sum_{i=1}^{N} |v_i(t)|^2$$

$$+ \sum_{i=1}^{q} |v_{j_i}(t)|^2$$

Combining estimates (i)–(iv), and noting that

$$\sum_{i=1}^{N} (a_{i+1}^N - a_i^N)|z_i(t)|^2 + \sum_{i=1}^{q} |z_{j_i}(t)|^2 = 0,$$

and

$$\sum_{i=1}^{N} (a_{i+1}^N - a_i^N)|v_i(t)|^2 + \sum_{i=1}^{q} |v_{j_i}(t)|^2 = 0,$$

we have

$$\frac{d}{dt}\|z(t)\|_{\tau^N}^2 \leq \omega\left\{|z_0(t)|^2 + \frac{h}{\omega N}\sum_{i=1}^{N} |z_i(t)|^2 + \frac{h}{\omega N}\sum_{i=1}^{N} |v_i(t)|^2\right\},$$

where $\omega = 1 + a_1^N + \sum_{i=0}^{q} |A_i|^2 + \sum_{i=1}^{q} |B_i|^2 + \|A_{01}\|_{L^2}^2 + \|B_{01}\|_{L^2}^2$. Since $\omega > 1$, and $\|\cdot\|_N \leq \|\cdot\|_{\tau^N} \leq \gamma\|\cdot\|_N$ where γ is independent of N [B1], then

$$\frac{d}{dt}\|z(t)\|_{\tau^N}^2 \leq \omega\|z(t)\|_{\tau^N}^2.$$

Hence,

$$\|z(t)\|_{\tau^N}^2 \leq \|z\|_{\tau^N}^2 + \omega\int_0^t \|z(\theta)\|_{\tau^N}^2 d\theta,$$

where we used $z = z(0)$. By Gronwall's inequality, $\|z(t)\|_{\tau^N}^2 \leq \|z\|_{\tau^N}^2 e^{\omega t}$, which concludes our proof. □

For the remainder of this section we will characterize the approximating resolvent operators $(\lambda I - [A^N])^{-1}$ where $[A^N]$ is the matrix representation of A^N (see Remark

3.2 below). In doing so we will introduce several approximating operators that are similar to well-known operators introduced in §2.

Let $\Delta^N(\lambda)$ be an $n \times n$ complex matrix defined by

$$\Delta^N(\lambda) = \lambda I - L^N(\lambda),$$

where $L^N(\lambda) = \sum_{j=0}^{N} A_j^N(\lambda/(\lambda h + N))^j$, for $\lambda \neq -\frac{N}{h}$, and where $A_0^N = A_0$. For $\lambda \in \mathbf{C}$ and $\lambda \neq -\frac{N}{h}$, we define the following mappings. The approximating "exponential" map $E_\lambda^N : \mathbf{C}^n \to \mathbf{C}^{n(N+1)+mN}$,

$$E_\lambda^N x = z = \text{col } (z_0, \ldots, z_N, v_1, \ldots, v_N),$$

where $z_0 = x, z_i = (N/(\lambda h + N))^i x, v_i = 0, i = 1, 2, \ldots, N$. A similar definition holds for $(E_\lambda^T)^N : \mathbf{C}^n \to \mathbf{C}^{n(N+1)+pN}$. Using the induced inner product in X^N, the dual mappings $(E_\lambda^*)^N$ and $(E_\lambda^{T^*})^N$ are given by

$$(E_\lambda^*)^N z = z_0 + \frac{h}{N} \sum_{i=1}^{N} \left(\frac{N}{\lambda h + N} \right)^i z_i,$$

$$(E_\lambda^{T^*})^N \xi = \xi_0 + \frac{h}{N} \sum_{i=1}^{N} \left(\frac{N}{\lambda h + N} \right)^i \xi_i,$$

where $\xi = \text{col } (\xi_0, \ldots, \xi_N, \nu_1, \ldots, \nu_N) \in \mathbf{C}^{n(N+1)+pN}$. We also define the linear mapping T_λ^N on $\mathbf{C}^{n(N+1)+mN}$ to be

$$T_\lambda^N z = y = \text{col } (y_0, \ldots, y_N, w_1, \ldots, w_N),$$

where

$$y_0 = 0, \quad y_j = \frac{h}{N} \sum_{i=1}^{j} \left(\frac{N}{\lambda h + N} \right)^{j+1-i} z_i, \quad w_j = \frac{h}{N} \sum_{i=1}^{j} \left(\frac{N}{\lambda h + N} \right)^{j+1-i} v_i,$$

for $j = 1, \ldots, N$. Finally, we introduce the "extended" approximating structural mapping $F^N : \mathbf{C}^{n(N+1)+mN} \to \mathbf{C}^{n(N+1)+pN}$

$$F^N z = y,$$

where

$$y_0 = z_0, \quad y_i = \sum_{j=i}^{N} A_j^N z_{j+1-i} + \sum_{j=i}^{N} B_j^N v_{j+1-i}, \quad w_i = \sum_{j=i}^{N} C_j^N z_{j+1-i},$$

$$\text{for } i = 1, 2, \ldots, N.$$

Using the same arguments as in [S1], we can easily show that the resolvent $(\lambda I - [A^N])^{-1}$ is given by

$$(\lambda I - [A^N])^{-1} = E_\lambda^N \Delta^N(\lambda)^{-1} (E_\lambda^{T^*})^N F^N + T_\lambda^N$$

for $\lambda \neq -\frac{N}{h}$ and satisfies $\det \Delta^N(\lambda) \neq 0$. Furthermore, the spectrum of A^N is the set $\{\lambda \in \mathbf{C}/\lambda \neq -\frac{N}{h},\ \det \Delta^N(\lambda) = 0\}$.

Remark 3.1. From the above formulas and those in §2 of the operators characterizing the approximating resolvents $R(\lambda, A^N)$ and $R(\lambda, A)$, respectively, it follows immediately from [S1] that $\wp^N(\lambda I - [A^N])^{-1}\pi^N$ converges to $R(\lambda, A)$ uniformly in the operator topology as $N \to \infty$ on bounded subsets of the complex plane that are uniformly bounded away from the zeros of $\det \Delta(\lambda)$.

Remark 3.2. To implement the averaging scheme on the computer, we must calculate the matrix representations for the operators A^N, B^N, and C^N with respect to the basis $\{(1,0,0),(0,\chi_j,0),(0,0,\chi_j)\}$, $j = 1,\ldots,N$. The matrix representation $[A^N]$ of A^N is given by

$$[A^N] = \begin{bmatrix} A_{11}^N & A_{12}^N \\ 0 & A_{22}^N \end{bmatrix},$$

where

$$A_{11}^N = (Q_{11}^N)^{-1}H_{11}^N, \qquad H_{11}^N = \begin{bmatrix} A_0 & A_1^N & \cdots & A_N^N \\ I_n & -I_n & & \\ & \ddots & \ddots & \\ & & I_n & -I_n \end{bmatrix},$$

$$A_{12}^N = (Q_{11}^N)^{-1}H_{12}^N, \qquad H_{12}^N = \begin{bmatrix} B_1^N & B_2^N & \cdots & B_N^N \\ 0 & & & 0 \\ & & & \\ 0 & & & 0 \end{bmatrix},$$

$$A_{22}^N = (Q_{22}^N)^{-1}H_{22}^N, \qquad H_{22}^N = \begin{bmatrix} -I_m & & & \\ I_m & -I_m & & \\ & \ddots & \ddots & \\ & & I_m & -I_m \end{bmatrix}.$$

The matrix representation $[B^N]$ of B^N is given by

$$[B^N] = \operatorname{col}\left[B_0, 0, \ldots, 0, \frac{N}{h}I_m, 0, \ldots, 0\right].$$

The matrix representation $[C^N]$ of C^N as an operator $X^N \to R^p$ is given by

$$[C^N] = [C_0\ C_1^N\ C_2^N\ \ldots\ C_N^N\ 0 \ldots 0].$$

3.2. Uniform differentiability and stability of approximating semigroups. Using the same notation as in Lasiecka and Manitius [L1], let $a, b > h$ and μ be fixed real numbers and let us define the set

$$\Sigma_a = \{\lambda \in \mathbf{C}/|I_m\lambda| \geq e^{ab}e^{-b\mathrm{Re}\lambda}, \quad \mathrm{Re}\lambda \leq \mu\}.$$

Let Σ_a^1 be the complement of Σ_a, and define

$$\mathcal{L} = \Sigma_a^1 \cap \{\lambda/\mathrm{Re}\lambda < \mu\}.$$

We will now obtain the following two results, which are essential for establishing the uniform differentiability of the approximating semigroups $S^N(t)$. At this point, we note that the spectra of A^N, $\sigma(A^N)$, is the same as the spectra of A_0^N, $\sigma(A_0^N)$, for each N, where A_0^N is the Nth averaging approximating generator corresponding to the RFDE (2.1) with no delays in control. Hence from [L1], the spectra of the approximating generators A^N corresponding to the averaging approximations are all contained in the logarithmic sector $\mathcal{L}$ for a suitable choice of a, b and μ. For the remainder of the paper we will denote Var ζ to be the total variation of ζ on $[-h, 0]$.

LEMMA 3.2. *Let $\mathcal{L}$ be given as above with $\mu >$ Var $\eta_0, b > h$ and $a > \max(a_0, a_1)$ where*

$$a_0 = \frac{1}{b}\log\frac{1+\sqrt{2}}{b-h} \quad \text{and} \quad a_1 = \gamma + \frac{\log \text{Var } \eta_0}{b}.$$

The eigenvalues of the approximating generators A^N are contained in $\mathcal{L}$ for all $N \in \mathbf{N}$.

Our next lemma shows that outside the logarithmic sector $\mathcal{L}$ the resolvents of the generators A^N are uniformly bounded by $|\text{Im}\lambda|$.

LEMMA 3.3. *Let μ, a, and b be given as in Lemma 3.2. Then there exists a positive constant c, independent of N, such that*

$$\|R(\lambda, A^N)\| \le c|\text{Im}\lambda|,$$

for all $\lambda \in \Sigma_a$, for all $N \in \mathbf{N}$.

Proof. By the representation of $R(\lambda, A^N)$, we have

$$\begin{aligned}\|R(\lambda, A^N)P^N\|_N &= \|\wr^N (\lambda I - [A^N])^{-1}\pi^N\| \\ &\le \|E_\lambda^N\|\,\|\Delta^N(\lambda)^{-1}\|\,\|(E_\lambda^{T^*})^N\|\,\|F^N\| + \|T_\lambda^N\|.\end{aligned}$$

For all λ in Σ_a and all N in $\mathbf{N}$ we have the following estimates of these norms with respect to $\|\cdot\|_N$.

(i) $\|(E_\lambda^{T^*})^N\| = \|E_\lambda^N\| \le \sqrt{1+h}Ke^{-ab}|\text{Im}\lambda|$, where $K = e^{\gamma b}$ (see [L1]).

(ii) $\|\Delta^N(\lambda)^{-1}\| \le c_0\frac{1}{|\lambda|}$, $c_0 = (1 - K\text{Var } \eta_0 e^{-ab})^{-1}$ (see [L1]).

(iii) $\|F^N\| \le \max\{1, \sqrt{2}\text{Var}\eta + \text{Var}\gamma, \sqrt{2}\text{Var}\beta\} = c_1$, for every N in $\mathbf{N}$.

For $z = \text{col}(z_0, \dots, z_N, v_1, \dots, v_N) \in R^{n(N+1)+mN}$

$$\begin{aligned}\|F^N z\|_N^2 = |z_0|^2 &+ \frac{h}{N}\sum_{i=1}^N \Big|\sum_{j=i}^N A_j^N z_{j+1-i} + \sum_{j=i}^N B_j^N v_{j+1-i}\Big|^2 \\ &+ \frac{h}{N}\sum_{i=1}^N \Big|\sum_{j=i}^N C_j^N z_{j+1-i}\Big|^2.\end{aligned}$$

By the well-known convolution inequality, we have

$$\begin{aligned}\sum_{i=1}^N \Big|\sum_{j=i}^N C_j^N z_{j+1-i}\Big|^2 &\le \left\{\sum_{j=1}^N |C_j^N|\right\}^2 \sum_{j=1}^N |z_j|^2 \\ &\le \text{Var}^2\gamma \sum_{j=1}^N |z_j|^2.\end{aligned}$$

Using standard argument, we can show that

$$\sum_{i=1}^{N} |\sum_{j=i}^{N} A_j^N z_{j+1-i} + \sum_{j=i}^{N} B_j^N v_{j+1-i}|^2 \leq 2\sum_{i=1}^{N} |\sum_{j=i}^{N} A_j^N z_{j+1-i}|2.$$

$$+ 2\sum_{i=1}^{N} |\sum_{j=i}^{N} B_j^N v_{j+1-i}|2$$

$$\leq 2\mathrm{Var}^2\eta \sum_{j=1}^{N} |z_j|^2 + 2\mathrm{Var}^2\beta \sum_{j=1}^{N} |v_j|^2.$$

Also,

(iv) $\|T_\lambda^N\| \leq hKe^{-ab}|\mathrm{Im}\lambda|$ (see [L1]).

Combining estimates (i)–(iv)

$$\|R(\lambda, A^N)P^N\|_N \leq c_2|\mathrm{Im}\lambda|$$

where

$$c_2 = c_0c_1Ke^{-ab}(1+h)h\left(\frac{Ke^{-ab}}{h} + \frac{1}{c_0c_1(1+h)}\right).$$

From [L1, App.],

$$\|R(\lambda, A^N)\| \leq \frac{\|I - P^N\|}{|\lambda|} + \|R(\lambda, A^N)P^N\|$$

$$\leq \frac{2}{|\lambda|} + c_2|\mathrm{Im}\lambda|$$

$$\leq (c_2 + c_3)|\mathrm{Im}\lambda|$$

where we used the estimate $\frac{2}{|\lambda|} \leq c_3|\mathrm{Im}\lambda|$ for some constant c_3 since $\lambda \in \Sigma_a$. □

From [L1], the above two lemmas imply immediately that the approximating semigroups $S^N(t)$ are uniformly differentiable for sufficiently large time. More precisely, we have the following theorem, which has been proved in [L1].

THEOREM 3.4. *For all N in $\mathbf{N}$ the approximating semigroups $S^N(t)$ are uniformly differentiable in the sense that there exist constants $t_0 = 3h, \mu > \mathrm{Var}\ \eta_0$ such that*

$$\|A^N S^N(t)\| \leq Me^{\mu t} \quad \forall t > t_0, \quad \forall N \in \mathbf{N}.$$

Furthermore, for $b > h$ and all $k = 0, 1, 2, \ldots$, and for any $t_k > (2+k)b$ there exist positive constants M_k (dependent on t_k, but independent of N) such that

$$\|(A^N)^k S^N(t)\| \leq M_k e^{\mu t}, \quad \forall t \geq t_k, \forall N \in \mathbf{N}.$$

Remark 3.3. (i) An important consequence of this theorem is that by considering $\int e^{\lambda t}R(\lambda, A^N)d\lambda$ along a shifted path, we can easily prove the following: if the original semigroup is exponentially stable with decay rate ω_0, $\omega_0 = \sup\{\mathrm{Re}\lambda/\lambda \in \sigma(A)\} < 0$, then the approximating semigroups are uniformly exponentially stable with decay rate $\omega_0 + \epsilon$, $\epsilon > 0$ is arbitrarily small and $N \geq N_1(\epsilon)$. The preservation of exponential stability under approximation is very important in the convergence proof of the approximating solution of the algebraic Riccati equation associated with a retarded system. Gibson [G1] showed that for RFDE without delays in control and observation,

uniform exponential stability of approximating semigroups yields strong convergence of approximating algebraic Riccati operators.

(ii) Another important consequence of the above theorem is that now we can write

$$S^N(t) = \frac{1}{2\pi i}\int_\Gamma e^{\lambda t}R(\lambda, A^N)d\lambda, \qquad t > 2b$$

where Γ is the boundary of $\mathcal{L}$ oriented so that $\mathrm{Im}\lambda$ increases along Γ. This representation of $S^N(t)$ makes it possible to study convergence rates of approximating semigroups by studying convergence rates of approximating resolvent operators (see §4). Furthermore, as will be evident in the next section, this formulation allowed us to prove that the approximating semigroups converge in the uniform operator topology to the original semigroup for sufficiently large time.

4. Convergence results and convergence rates.

4.1. Homogeneous equations. Throughout this section we will restrict our discussion to the homogeneous systems, $u(t) = 0$ for $t \geq 0$,

$$\frac{d}{dt}z(t) = Az(t), \qquad z(0) = \phi,$$

$$y(t) = Cz(t)$$

in the Hilbert space X and

$$\frac{d}{dt}z^N(t) = A^N Z^N(t), \qquad z^N(0) = P^N\phi,$$

$$y^N(t) = C^N z^N(t)$$

in the finite-dimensional subspaces X^N of X. The operators A and C and their approximating operators A^N and C^N corresponding to the averaging scheme are defined in §§2 and 3, respectively.

We will first show that the approximating semigroups converge to the original semigroup strongly for all initial data in X and the convergence is uniform in t for t in bounded intervals. The underlying tool for this convergence proof is the Trotter–Kato semigroup approximation theorem. We will use here a version of this theorem given by Pazy [P1, Thm. 4.5, p. 88]), which is also used by Banks and Burns [B1] and Gibson [G1].

THEOREM 4.1. *Let A generate a C_0-semigroup $S(\cdot)$ on a Hilbert space X and A^N be a sequence of linear operators, each of which generates a C_0-semigroup $S^N(\cdot)$ on X. Assume the following*:

(a) *as $N \to \infty$, $A^N x \to Ax$ for every $x \in D$ where D is dense in X;*

(b) *there exist constants ω, M independent of N such that $\|S^N(t)\| \leq Me^{\omega t}$;*

(c) *there exists a λ_0 with $\lambda_0 > \omega$ for which $(\lambda_0 I - A)D$ is dense in X.*

Then $\lim_{N\to\infty} S^N(t)x = S(t)x$ for all $t \geq 0$, $x \in X$ and the limit is uniform in t for t in bounded intervals.

We extend the definition of A^N to all of X by $A^N x = A^N P^N x$ and define the C_0-semigroup $S^N(t)$, $t \geq 0$, on X by

$$S^N(t)\phi = S^N(t)P^N\phi + \phi - P^N\phi, \qquad t \geq 0, \quad \phi \in X.$$

The desired convergence $S^N(t)P^N\phi \to S(t)\phi$, $\phi \in X$, will then follow directly from the above theorem and the strong convergence of P^N to I. Let us now proceed to

prove that condition (a) holds for the approximating generators A^N constructed in §3.1. Note that the stability result, condition (c), has been obtained in §3.1 (Theorem 3.1) whereas condition (b) has been proved in [B1].

LEMMA 4.2. *Let*

$$D = \{(\phi^1(0), \phi^1, \phi^2)/\phi^1 \in C^1([-h,0]; R^n), \phi^2 \in C^1([-h,0]; R^m)\}.$$

Then D is dense in X and $\lim_{N\to\infty} A^N\phi = A\phi$ for every ϕ in D.

Proof. Let $\phi = (\phi^1(0), \phi^1, \phi^2) \in D$. The following result has been proved in [B1]:

$$\lim_{N\to\infty} \left\|\nabla_1 P^N\phi - \dot{\phi}^1\right\|_{L^2}^2 = \lim_{N\to\infty} \left\|\sum_{j=1}^N \frac{N}{h}(z_{j-1} - z_j)\chi_j - \dot{\phi}^1(\cdot)\right\|_{L^2}^2 = 0,$$

where $P^N\phi = \text{col}\,(z_0, \ldots, z_N, v_1, \ldots, v_N)$. Similarly,

$$\lim_{N\to\infty} \left\|\nabla_2 P^N\phi - \dot{\phi}^2\right\|_{L^2}^2 = \lim_{N\to\infty} \left\|\sum_{j=1}^N \frac{N}{h}(v_{j-1} - v_j)\chi_j - \dot{\phi}^2(\cdot)\right\|_{L^2}^2 = 0.$$

Hence it only remains to prove the following convergence of R^n-component,

$$\lim_{N\to\infty} \left\{\sum_{j=0}^N A_j^N z_j + \sum_{j=1}^N B_j^N v_j\right\} = L\phi^1 + \hat{B}\phi^2.$$

Let us define $\eta_0^N(t) = \lim_{\theta\uparrow t_j} \eta_0(\theta)$ and $\beta^N(t) = \lim_{\theta\uparrow t_j} \beta(\theta)$ for $t \in (t_{j+1}, t_j]$ and $j = 0, 1, \ldots, N-1$. Then

$$\begin{aligned}
\sum_{j=0}^N A_j^N z_j &= \sum_{j=0}^N \eta_0^N(t_{j-1})z_j - \sum_{j=0}^N \eta_0^N(t_j)z_j \\
&= \sum_{j=1}^N \eta_0^N(t_{j-1})z_j - \sum_{j=1}^{N+1} \eta_0^N(t_{j-1})z_{j-1} \\
&= -\eta_0(-h)z_N - \frac{h}{N}\sum_{j=1}^N \eta_0^N(t_{j-1})\frac{N}{h}(z_{j-1} - z_j) \\
&= -\eta_0(-h)z_N - \sum_{j=1}^N \int_{t_j}^{t_{j-1}} \eta_0^N(s)\frac{N}{h}(z_{j-1} - z_j)ds \\
&= -\eta_0(-h)z_N - \int_{-h}^0 \eta_0^N(s)\sum_{j=1}^N \frac{N}{h}(z_{j-1} - z_j)\chi_j(s)ds.
\end{aligned}$$

Similarly,

$$\sum_{j=1}^N B_j^N v_j = -\beta(-h)v_N - \int_{-h}^0 \beta^N(s)\sum_{j=1}^N \frac{N}{h}(v_{j-1} - v_j)\chi_j(s)ds.$$

By integration by parts,

$$L\phi^1 + \hat{B}\phi^2 = -\eta_0(-h)\phi^1(-h) - \int_{-h}^{0} \eta_0(s)\dot{\phi}^1(s)ds - \beta(-h)\phi^2(-h) - \int_{-h}^{0} \beta(s)\dot{\phi}^2(s)ds.$$

Hence,

$$\begin{aligned}
&|L\phi^1 + \hat{B}\phi^2 - (\sum_{j=0}^{N} A_j^N z_j + \sum_{j=1}^{N} B_j^N v_j)| \\
&\quad \le \|\eta_0(-h)\|\ |\phi^1(-h) - z_N| + \|\beta(-h)\|\ |\phi^2(-h) - v_N| \\
&\qquad + \left| \int_{-h}^{0} \eta_0(s)\dot{\phi}^1(s)ds - \int_{-h}^{0} \eta_0^N(s) \sum_{j=1}^{N} \frac{N}{h}(z_{j-1} - z_j)\chi_j(s)ds \right| \\
&\qquad + \left| \int_{-h}^{0} \beta(s)\dot{\phi}^2(s)ds - \int_{-h}^{0} \beta^N(s) \sum_{j=1}^{N} \frac{N}{h}(v_{j-1} - v_j)\chi_j(s)ds \right|.
\end{aligned}$$

Since

$$z_N = \frac{N}{h}\int_{-h}^{-(N-1)\frac{h}{N}} \phi^1(s)ds \quad \text{and} \quad v_N = \frac{N}{h}\int_{-h}^{-(N-1)\frac{h}{N}} \phi^2(s)ds,$$

it follows immediately that $z_N \to \phi^1(-h)$ and $v_N \to \phi^2(-h)$. Furthermore, we have

$$\begin{aligned}
&\left| \int_{-h}^{0} \eta_0(s)\dot{\phi}^1(s)ds - \int_{-h}^{0} \eta_0^N(s) \sum_{j=1}^{N} \frac{N}{h}(z_{j-1} - z_j)\chi_j(s)ds \right| \\
&\quad \le \left| \int_{-h}^{0} (\eta_0(s) - \eta_0^N(s))\dot{\phi}^1(s)ds \right| \\
&\qquad + \left| \int_{-h}^{0} \eta_0^N(s)\Big(\dot{\phi}^1(s) - \sum_{j=1}^{N} \frac{N}{h}(z_{j-1} - z_j)\chi_j(s)\Big)ds \right|
\end{aligned}$$

From [S1, Rem. 4.1(i)], we can show that the first term on the right-hand side is bounded by $\sup_{s\in[-h,0]} |\dot{\phi}^1(s)| \frac{h}{N}\mathrm{Var}\eta_0$. The second term is bounded by $\mathrm{Var}\eta_0\sqrt{h}\|\dot{\phi}^1 - \sum_{j=1}^{N} \frac{N}{h}(z_{j-1} - z_j)\chi_j\|_{L^2}$. Analogously, we can show that

$$\begin{aligned}
&\left| \int_{-h}^{0} \beta(s)\dot{\phi}^2(s)ds - \int_{-h}^{0} \beta^N(s) \sum_{j=1}^{N} \frac{N}{h}(v_{j-1} - v_j)\chi_j(s)ds \right| \\
&\quad \le \sup_{s\in[-h,0]} |\phi^2(s)| \frac{h}{N}\mathrm{Var}\beta + \sqrt{h}\mathrm{Var}\beta \left\| \dot{\phi}^2 - \sum_{j=1}^{N} \frac{N}{h}(v_{j-1} - v_j)\chi_j \right\|_{L^2}.
\end{aligned}$$

The statement of the lemma then follows by the convergence of the C^1-terms [B1]. □

In what follows we will establish convergence rates for the sequence of approximating semigroups in both the strong and uniform operator topology. By using the differentiability of $S(t)$ and $S^N(t)$, i.e., for $t > 2b$ we have

$$P^N S(t) - S^N(t)P^N = \int_\Gamma \frac{1}{2\pi i}[P^N R(\lambda, A) - R(\lambda, A^N)P^N]e^{\lambda t}d\lambda,$$

where Γ is the boundary of some logarithmic sector containing all eigenvalues of A^N, we compute these estimates through technical estimates of the convergence rates of the approximating resolvent operators. The following hypotheses will be considered in various combinations.

(H3)

$$\eta(\theta) = -\sum_{i=1}^{q} A_i \chi_{(-\infty,-h_i]}(\theta) - \int_\theta^0 A_{01}(s)ds,$$

$$\beta(\theta) = -\sum_{i=1}^{q} B_i \chi_{(-\infty,-h_i]}(\theta) - \int_\theta^0 B_{01}(s)ds.$$

(H4) The matrix functions η, β are normalized functions of bounded variation, left continuous of $[-h, 0)$ (in the same sense as in [P2]).

(H5) The delays h_i are commensurate, i.e. there is a positive real number r, and integers $k_j, j = 1, \ldots, q$, such that $h_j = k_j r, j = 1, \ldots, q$.

(H6) The interval $[-h, 0]$ is partitioned into N subintervals $[t_j, t_{j-1})$, where $j = 1, \ldots, N$, and $t_j = -\frac{jh}{N}$.

(H7) In the case where $q > 1$, the set of meshpoints t_j includes $\{h_0, \ldots, h_q\}$, i.e., we partition $[-h, 0]$ into $k_q N$ subintervals $[t_j, t_{j-1})$ where $t_j = -\frac{jr}{N}$, $j = 0, 1, \ldots, k_q N$.

In the results we will point out which assumptions guarantee which convergence rate.

THEOREM 4.3. *Let λ be a fixed real number, $\lambda > \mathrm{Var}\eta_0$. Then $\lambda \in \rho(A) \cap \rho(A^N)$ for all $N \in \mathbf{N}$. Let $N_0 \geq \lambda h$. Then for all $N \geq N_0$:*

(i) *if* (H4) *and* (H6) *hold, then*

$$\|P^N R(\lambda, A) - R(\lambda, A^N)P^N\| \leq D_1 \left(\frac{h}{N}\right)^{\frac{1}{2}};$$

(ii) *if* (H3) *and* (H6) *hold and the system has only one delay, $h_i = h$, then*

$$\|P^N R(\lambda, A) - R(\lambda, A^N)P^N\| \leq D_2 \frac{h}{N};$$

(iii) *if* (H3), (H5)–(H7) *hold, then*

$$\|P^N R(\lambda, A) - R(\lambda, A^N)P^N\| \leq D_3 \frac{r}{N};$$

(iv) *if only* (H4) *and* (H6) *hold, but $x \in R^n \times L^\infty([-h, 0]; R^n) \times L^\infty([-h, 0]; R^m)$, then*

$$\|P^N R(\lambda, A)x - R(\lambda, A^N)P^N x\|_{R^n \times L^\infty \times L^\infty} \leq D_4 \frac{h}{N} \|x\|_{R^n \times L^\infty \times L^\infty}.$$

Proof. From the structures of the resolvents,

$$P^N R(\lambda, A) - R(\lambda, A^N)P^N = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = (P^N E_\lambda - \wr^N E_\lambda^N)\Delta^{-1}(\lambda)E_\lambda^{T^*}F,$$
$$T_2 = \wr^N E_\lambda^N(\Delta^{-1}(\lambda) - \Delta^N(\lambda)^{-1})E_\lambda^{T^*}F,$$
$$T_3 = \wr^N E_\lambda^N \Delta^N(\lambda)^{-1}(E_\lambda^{T^*}F - (E_\lambda^{T^*})^N F^N \pi^N),$$
$$T_4 = P^N T_\lambda - \wr^N T_\lambda^N \pi^N.$$

Let us first obtain the following estimates on $\|F\|$:

$$\|F\phi\|_X^2 \le |\phi^0|^2 + 2\left\|\int_{-h}^{\cdot} d\eta(\theta)\phi^1(\theta - \cdot)\right\|_{L^2}^2 + 2\left\|\int_{-h}^{\cdot} d\beta(\theta)\phi^2(\theta - \cdot)\right\|_{L^2}^2 + \left\|\int_{-h}^{t} d\gamma(\theta)\phi^1(\theta - t)\right\|_{L^2}^2.$$

From [D1]

$$\|F\| \le d_2 = \max(1, \sqrt{2}\mathrm{Var}\eta + \mathrm{Var}\gamma, \sqrt{2}\mathrm{Var}\beta).$$

Now by comparing the operators characterizing the resolvents corresponding to RFDE with delays in control and observation and those same operators corresponding to systems with no delays in input and output, it follows straightforwardly from [L1] that the following estimates on $\|T_i\|, i = 1, 2, 4$, still hold.

(i) $\|T_1\| \le d_1 d_2 d_4 d_5 \frac{h}{N}$, for $\lambda > d_3$, and $N \ge N_0$, where $d_1 = \sqrt{1+h}$, $d_3 = \sup_{\theta\in[-h,0]}\|\eta_0(\theta)\| \le \mathrm{Var}\eta_0$, $d_4 = (\lambda - d_3)^{-1}$, and $d_5 = 2\lambda\sqrt{h}(1+\lambda h)$.

(ii) $\|T_2\| \le d_1^2 d_2 d_4^2 d_6 \frac{h}{N}$ for $\lambda > d_3, N \ge N_0$, where $d_6 = 2d_3\lambda(1+\lambda h)^2 + \lambda h\mathrm{Var}\eta_0$.

(iii) $\|T_4\| \le d_7 \frac{h}{N}$, for $N \ge N_0$, where $d_7 = \sqrt{2}(\lambda\sqrt{h}+1) + d_5\sqrt{h}$.

(iv) For $\|T_3\|$ we have

$$\|T_3\| \le d_1 d_4 \|E_\lambda^{T^*}F - (E_\lambda^{T^*})^N F^N \pi^N\|, \qquad \lambda > d_3.$$

Now

$$\|E_\lambda^{T^*}F - (E_\lambda^{T^*})^N F^N \pi^N\| \le \|E_\lambda^{T^*}(F - \wr^N F^N \pi^N)\| + \|(E_\lambda^{T^*}\wr^N - (E_\lambda^{T^*})^N)F^N\pi^N\|.$$

The second term on the right-hand side is bounded above by $d_2 d_5 \frac{h}{N}$, for $N \ge N_0$. As pointed out in Lasiecka and Manitius [L1], it would seem very difficult to obtain the convergence rate of $\|F\phi - \wr^N F^N \pi^N \phi\|$ valid uniformly for all ϕ in X. However, since $\|E_\lambda^{T^*}(F - \wr^N F^N \pi^N)\| = \|(F^* - \wr^N (F^N)^* \pi^N)E_\lambda^T\|$, the rate of convergence of $\|T_3\|$ to zero depends on the rate with which the approximation $\wr^N (F^N)^* \pi^N$ converges to F^* on the exponential function $E_\lambda^T x$, where x is the unit vector in R^n. This is a crucial step in the proof of the theorem and it is this rate that depends on the particular hypotheses (H3)–(H7) chosen.

From the formula (2.5) for the dual operator F^*, it is easy to show that

$$[F^* E_\lambda^T]^0 - [\wr^N (F^N)^* \pi^N E_\lambda^T]^0 = 0$$

and

$$[F^* E_\lambda^T]^1(\theta) = \int_{-h}^{\theta} d\eta^T(s) e^{\lambda(s-\theta)}$$
$$= \eta^T(\theta) - \eta^T(-h)e^{-\lambda h} - \lambda\int_{-h}^{0}\eta^T(s+\theta)e^{\lambda s}ds,$$
$$[F^* E_\lambda^T]^2(\theta) = \int_{-h}^{\theta} d\beta^T(s) e^{\lambda(s-\theta)}$$
$$= \beta^T(\theta) - \beta^T(-h)e^{-\lambda h} - \lambda\int_{-h}^{0}\beta^T(s+\theta)e^{\lambda s}ds.$$

Let $(F^N)^*\pi^N E_\lambda^T = \operatorname{col}(z_0, \dots, z_N, v_1, \dots, v_N) \in R^{n(N+1)+mN}$ where by using the same arguments as in [S1, Thm. 4.4] we have

$$z_j = \frac{N}{h}\int_{-h}^{0} [\eta^{N^T}(s+t_{j-1}) - \eta^{N^T}(s+t_j)]e^{\lambda s}ds,$$
$$v_j = \frac{N}{h}\int_{-h}^{0} [\beta^{N^T}(s+t_{j-1}) - \beta^{N^T}(s+t_j)]e^{\lambda s}ds, \qquad j = 1, 2, \dots, N.$$

Now

$$z_j = \frac{N}{h}\int_{-h}^{0} \eta^{N^T}(s+t_{j-1})e^{\lambda s}ds - \frac{N}{h}\int_{-h-\frac{h}{N}}^{-\frac{h}{N}} \eta^{N^T}(s+t_{j-1})e^{\lambda(s+\frac{h}{N})}ds$$

and after some simple manipulations on the second term on the right-hand side, we obtain

$$z_j = \frac{N}{\lambda h}[e^{\lambda h/N} - 1]\left\{\eta^{N^T}(t_{j-1}) - \eta^T(-h)e^{-\lambda h} - \lambda\int_{-h}^{0} \eta^{N^T}(s+t_{j-1})e^{\lambda s}ds\right\}.$$

Similarly,

$$v_j = \frac{N}{\lambda h}[e^{\lambda h/N} - 1]\left\{\beta^{N^T}(t_{j-1}) - \beta^T(-h)e^{-\lambda h} - \lambda\int_{-h}^{0} \beta^{N^T}(s+t_{j-1})e^{\lambda s}ds\right\}.$$

Note that

$$\frac{N}{\lambda h}(e^{\lambda h/N} - 1) = \frac{N}{\lambda h}\left(1 + \frac{\lambda h}{N} + \frac{(\lambda h)^2}{2!N^2} + \cdots - 1\right)$$
$$= 1 + \epsilon_N,$$

where

$$\epsilon_N = \frac{\lambda h}{2N} + \frac{(\lambda h)^2}{6N^2} + \cdots,$$

and

$$\epsilon_N = \frac{\lambda h}{2N}\left(1 + \frac{\lambda h}{3N} + \frac{(\lambda h)^2}{12N^2} + \cdots\right)$$
$$\le \frac{\lambda h}{2N}\left(1 + \frac{\lambda h}{2N} + \left(\frac{\lambda h}{2N}\right)^2 + \cdots\right), \qquad N > \lambda h$$
$$= \frac{\lambda h}{2N}\frac{1}{1 - \frac{\lambda h}{2N}} \le \frac{\lambda h}{2N}\frac{1}{1 - \frac{1}{2}} = \frac{\lambda h}{N}.$$

Hence

$$\|(F^*E_\lambda^T - \wr^N(F^N)^*\pi^N E_\lambda^T)^1(\cdot)\|_{L^2} \le \|\eta^T - \eta^{N^T}\|_{L^2} + 3d_3\frac{\lambda}{N}$$
$$+ \lambda\left\|\int_{-h}^{0} [\eta^T(s+\cdot) - \eta^{N^T}(s+t_{j-1})]e^{\lambda s}ds\right\|_{L^2}.$$

From the well-known convolution theorem (see, e.g., Hewitt and Ross [H1]) the last term on the right-hand side is bounded by $\|\eta^T - \eta^{N^T}\|_{L^2}$. Therefore,

$$\|(F^*E_\lambda^T - \wr^N(F^N)^*\pi^N E_\lambda^T)^1(\cdot)\|_{L^2} \le 3\frac{\lambda h}{N}d_3 + 2\|\eta^T - \eta^{N^T}\|_{L^2}.$$

Similarly,

$$\|(F^*E_\lambda^T - \iota^N(F^N)^*\pi^N E_\lambda^T)^2(\cdot)\|_{L^2} \leq 3\frac{\lambda h}{N}d_8 + 2\|\beta^T - \beta^{N^T}\|_{L^2},$$

where $d_8 = \sup_{\theta\in[-h,0]}\|\beta(\theta)\| < \mathrm{Var}\beta$. Therefore,

$$\|T_3\| \leq d_1d_4d_2d_5\frac{h}{N} + d_1d_4\Big\{\Big[3\frac{\lambda h}{N}d_3 + 2\|\eta^T - \eta^{N^T}\|_{L^2}\Big]^2 + \Big[3\frac{\lambda h}{N}d_8 + 2\|\beta^T - \beta^{N^T}\|_{L^2}\Big]^2\Big\}^{1/2}.$$

The rate of convergence of $\|T_3\|$ depends on both $\|\eta^T - \eta^{N^T}\|_{L^2}$ and $\|\beta^T - \beta^{N^T}\|_{L^2}$. The rate of convergence of these bounded variation functions depend on the particular hypotheses (H3)–(H7) chosen. Using the same arguments as in [L1, pp. 41–42], the statements in the theorem then follow straightforwardly. □

We can now state the following theorem, which has been proved by Lasiecka and Manitius [L1].

THEOREM 4.4. *Let $\alpha = \frac{1}{2}$ if* (H4) *and* (H6) *hold and $\alpha = 1$ if* (H3), (H6) *hold and $q = 1$, or* (H3), (H5)–(H7) *hold and h is replaced by r. Then the following estimates are true.*

(i) *If $x \in \mathrm{dom}A^2$, then for each $T > 0$ there is a constant D_5 such that*

$$\|P^NS(t)x - S^N(t)P^Nx\| \leq D_5\left(\frac{h}{N}\right)^\alpha \|x\|_{H^2}$$

uniformly for $t \in [0,T]$.

(ii) *If $x \in \mathrm{dom}A$, then*

$$\|P^NS(t)x - S^N(t)P^Nx\| \leq D_6e^{\gamma t}\left(\frac{h}{N}\right)^\alpha \|x\|_{H^1}, \quad \forall N \in \mathbf{N}, \quad t > 4h.$$

(iii) *For $t > 5h$ and $\forall N > N_0$,*

$$\|P^NS(t) - S^N(t)P^N\| \leq D_7e^{\gamma t}\left(\frac{h}{N}\right)^\alpha.$$

Remark 4.1. By the Trotter–Kato theorem, we have $S^N(t)P^N$ converge to $S(t)$ strongly and uniformly in t for t in bounded interval. By using differentiability results of $S(t)$ and $S^N(t)$, part (iii) shows a stronger convergence result in the sense that the convergence is uniform in the uniform operator topology for sufficiently large time, $t > 5h$.

We conclude this section by focusing on the convergence of the output operators. In particular, we will show that $C^NS^N(\cdot)\pi^N\phi$ converges to $CS(\cdot)\phi$ strongly in $L^2(0,T;R^p)$ for all ϕ in X. This is not a trivial problem, since we are considering the averaging approximation scheme on linear subspaces X^N of X, and on X the output operator C is unbounded. Furthermore, the approximating operators C^N are bounded but not uniformly bounded on X. Before turning to our main result, we will need the following important preliminary result, which is the hypothesis (H2) corresponding to the approximating equations in X^N.

THEOREM 4.5. *For every $T > 0$, there exists a positive constant c, dependent on T but independent of N, such that*

$$\int_0^T |C^N e^{A^N t} x|_{R^p}^2 \, dt \leq c \|x\|_N^2,$$

for all $x \in R^{n(N+1)+mN}$ and for all N in $\mathbf{N}$.

Proof. The whole essence of this proof is the application of the Plancherel's theorem. To this end, let us consider the composition of C^N and $(\lambda I - A^N)^{-1}x$ where $x \in R^{n(N+1)+mN}$. From the characterization of the approximating resolvent operators, we have

$$C^N(\lambda I - A^N)^{-1}x = C^N E_\lambda^N \Delta^N(\lambda)^{-1}(E_\lambda^{T^*})^N F^N x + C^N T_\lambda^N x.$$

Let R_1 and R_2 be the first and second terms of the above right-hand terms. We now prove the statement of the theorem in three steps.

(i) There exist constants $\gamma > 0$ and $D_8 > 0$ such that

$$|R_1| \leq \frac{D_8}{|Im\lambda|} \|x\|_N$$

for all $\lambda \in C$ with $\mathrm{Re}\lambda = \gamma$ and for all N in $\mathbf{N}$. This statement holds from the following elementary estimates.

(i1) $|C^N E_\lambda^N| \leq \mathrm{Var}\,\gamma_0$, $\mathrm{Re}\lambda > 0$.

Let $y \in C^n$, then

$$\begin{aligned} |C^N E_\lambda^N y|_{R^p} &= \left| \sum_{i=0}^N \left(\frac{N}{\lambda h + N} \right)^i C_i^N y \right| \\ &\leq \sum_{i=0}^N \left| \frac{N}{\lambda h + N} \right|^i |C_i^N||y|. \end{aligned}$$

For

$$\mathrm{Re}\lambda > 0, \quad \left| \frac{N}{\lambda h + N} \right| \leq 1$$

and thus the estimate.

(i2) For $|\lambda| > 2\mathrm{Var}\eta_0$, $|\Delta^N(\lambda)^{-1}| \leq \frac{2}{|\mathrm{Im}\lambda|}$.

We have

$$\Delta^N(\lambda)^{-1} = \frac{1}{\lambda} \left(I - \frac{1}{\lambda} \sum_{j=1}^N A_j^N \left(\frac{N}{\lambda h + N} \right)^j \right)^{-1}$$

and for $|\lambda| > \mathrm{Var}\eta_0$,

$$\Delta^N(\lambda)^{-1} = \frac{1}{\lambda} \sum_{i=0}^\infty \frac{(\sum_{j=0}^N A_j^N (\frac{N}{\lambda h + N})^j)^i}{\lambda^i}.$$

Therefore,

$$\begin{aligned} \|\Delta^N(\lambda)^{-1}\| &\leq \frac{1}{|\lambda|} \sum_{i=0}^\infty \frac{\mathrm{Var}^i \eta_0}{|\lambda|^i} \\ &= \frac{1}{|\lambda| - \mathrm{Var}\eta_0} = \frac{1}{\frac{|\lambda|}{2} + \frac{|\lambda|}{2} - \mathrm{Var}\eta_0} \\ &\leq \frac{2}{|\lambda|}, \quad \text{for } |\lambda| > 2\mathrm{Var}\eta_0. \end{aligned}$$

(i3) $\|(E_\lambda^{T^*})^N\|_N \le \sqrt{1+h}$.

(i4) $\|F^N\|_N \le \max(1, \sqrt{2}\mathrm{Var}\eta + \mathrm{Var}\gamma, \sqrt{2}\mathrm{Var}\beta)$ (Lemma 3.3 (iii)).

Next we show that

(ii) $\int_{-\infty}^{\infty} |C^N T_{\alpha+i\omega}^N x|_{C^p}^2 d\omega \le D_9 \|x\|_N^2$, for $x \in R^{n(N+1)+mN}$.

Let $x = \mathrm{col}(x_0, \dots, x_N, v_1, \dots, v_N) \in R^{n(N+1)+mN}$, then the matrix representation of $C^N T_\lambda^N x$ is given by

$$C^N T_\lambda^N = [0\ C_1^N \dots C_N^N\ 0 \dots 0] \frac{h}{N} \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{N}{\lambda h+N} I_n & & & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & (\frac{N}{\lambda h+N})^N I_n & \cdots & \frac{N}{\lambda h+N} I_n & 0 \\ 0 & & \cdots & 0 & 0 \end{bmatrix},$$

which can be rewritten equivalently as

$$= [0\ I_p\ 0 \dots 0] \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & C_1^N & \cdots & C_N^N & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & C_N^N & & & \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \frac{h}{N}$$

$$\times \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{N}{\lambda h+N} I_n & & & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & (\frac{N}{\lambda h+N})^N I_n & \cdots & \frac{N}{\lambda h+N} I_n & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Let us denote by N_1 the first row matrix on the right-hand side, N_2 the second matrix and N_3^λ the third matrix multiply by $\frac{h}{N}$. Then

(ii1) $\|N_1\|_N = 1$.

(ii2) $\|N_2\|_N \le \max(1, \mathrm{Var}\gamma)$, for all $N \in \mathbf{N}$.

Let $z = \mathrm{col}(z_0, \dots, z_N, u_1, \dots, u_N) \in R^{n(N+1)+mN}$, then by convolution inequality

$$\begin{aligned} \|N_2 z\|_N^2 &= \frac{h}{N} \sum_{i=1}^{N} \left| \sum_{j=1}^{N} C_j^N z_{j+1-i} \right|^2 \\ &\le \frac{h}{N} \mathrm{Var}^2\gamma \sum_{j=1}^{N} |z_j|^2 \\ &\le \max^2(1, \mathrm{Var}\gamma) \|z\|_N^2, \quad \forall N \in \mathbf{N}. \end{aligned}$$

(ii3) From Salamon [S1, Thm. 4.9, Steps 1–3], we can show that

$$\int_{-\infty}^{\infty} \|N_3^{\alpha+i\omega} x\|_N^2 \, d\omega \le \frac{\pi h}{N} \frac{\gamma_0^2}{\alpha+\epsilon_0} \sum_{i=1}^{N} |x_i|^2,$$

where $0 \le \gamma_0 \le 1$, $\epsilon_0 > \frac{1}{h}$. Hence

$$\int_{-\infty}^{\infty} \|N_3^{\alpha+i\omega} x\|_N^2 \, d\omega \le \max^2 \left(1, \sqrt{\frac{\pi}{\alpha+\epsilon_0}} \gamma_0\right) \|x\|_N^2.$$

Combining (ii1)–(ii3), we obtain

$$\int_{-\infty}^{\infty} |C^N T^N_{\alpha+i\omega} x|^2 d\omega \leq D_9 \|x\|^2_N,$$

where $D_9 = \max^2(1, \mathrm{Var}\gamma)\max^2(1, \sqrt{\pi/(\epsilon_0+\alpha)}\gamma_0)$.

(iii) Now by steps (i) and (ii),

$$\int_{-\infty}^{\infty} |C^N((\alpha+i\omega)I - A^N)^{-1}x|^2 d\omega \leq 2\int_{-\infty}^{\infty} \frac{D_8^2}{\omega^2}\|x\|^2_N d\omega + 2\int_{-\infty}^{\infty} |C^N T^N_{\alpha+i\omega}x|^2 d\omega$$
$$\leq 2D_9\|x\|^2_N.$$

Since the Fourier transform of $e^{-\alpha t}C^N e^{A^N t}x$ is

$$\frac{1}{\sqrt{2\pi}} C^N((\alpha+i\omega)I - A^N)^{-1}x,$$

by Plancherel's theorem,

$$\int_0^{\infty} e^{-2\alpha t}|C^N e^{A^N t}x|^2 dt \leq \frac{1}{2\pi} 2D_9\|x\|^2_N.$$

Since $e^{2\alpha(T-s)} \geq 1$ for $s \in [0,T]$,

$$\int_0^T |C^N e^{A^N t}x|^2 dt \leq \int_0^T e^{2\alpha(T-s)}|C^N e^{A^N s}x|^2 ds \leq e^{2\alpha T}\frac{D_9}{\pi}\|x\|^2_N. \qquad \square$$

THEOREM 4.6. $\lim_{N\to\infty} \int_0^T |C^N e^{A^N t}\pi^N\phi - CS(t)\phi|^2 dt = 0$, *for all* ϕ *in* X.

Proof. By Theorem 4.5, it suffices to prove this theorem for all ϕ in D where D is some dense subset in X. To this end, we consider the following subset in X:

$$D = \{(\phi^1(0), \phi^1, \phi^2)/\phi^1 \in C^1([-h,0]; R^n), \phi^2 \in C^1([-h,0]; R^m)\}.$$

First let us obtain the following preliminary result. Let

$$z^N(t) = \mathrm{col}(z_1^N(t), \ldots, z_N^N(t), v_1^N(t), \ldots, v_N^N(t)) = e^{A^N t}\pi^N\phi \in R^{n(N+1)+mN},$$

where $\phi \in D$. We have

$$\|\wr^N A^N e^{A^N t}\pi^N\phi - AS(t)\phi\|_X \leq \|\wr^N e^{A^N t}\pi^N\| \|\wr^N A^N\pi^N\phi - A\phi\|$$
$$+ \|\wr^N e^{A^N t}\pi^N A\phi - S(t)A\phi\|_X.$$

From the stability result, Theorem 3.1, Lemma 4.2, and the Trotter–Kato Theorem 4.1, it follows straightforwardly

$$\lim_{N\to\infty} \|\wr^N A^N e^{A^N t}\pi^N\phi - AS(t)\phi\|_X = 0,$$

and the limit is uniform in t for t in bounded interval $[0,T]$.

Now, we consider

$$C^N z^N(t) = \sum_{i=0}^{N} C_i^N z_i^N(t).$$

Using the same arguments as in the proof of Lemma 4.2, we have

$$C^N z^N(t) = -\gamma_0(-h) z_N^N(t) - \int_{-h}^{0} \gamma_0^N(s) \sum_{i=1}^{N} \frac{N}{h} (z_{i-1}^N(t) - z_i^N(t)) \chi_i(s) ds.$$

Also by integration by parts

$$CS(t)\phi = -\gamma_0(-h)x(t-h) - \int_{-h}^{0} \gamma_0(s) \frac{d}{ds} x(t+s) ds.$$

However,

$$\begin{aligned}
\lim_{N\to\infty} z_N^N(t) &= \lim_{N\to\infty} \left[z_0^N(t) - \sum_{j=1}^{N} (z_{j-1}^N(t) - z_j^N(t)) \right] \\
&= \lim_{N\to\infty} \left[z_0^N(t) - \sum_{j=1}^{N} \int_{t_j}^{t_{j-1}} \frac{N}{h} (z_{j-1}^N(t) - z_j^N(t)) ds \right] \\
&= \lim_{N\to\infty} \left[z_0^N(t) - \sum_{j=1}^{N} \int_{-h}^{0} \frac{N}{h} (z_{j-1}^N(t) - z_j^N(t) \chi_j ds \right] \\
&= x(t) - \int_{-h}^{0} [AS(t)\phi]^1(s) ds \\
&= x(t-h),
\end{aligned}$$

and this limit is uniform in t for t in $[0, T]$. Also,

$$\begin{aligned}
&\left| \int_{-h}^{0} \gamma_0^N(s) \sum_{i=1}^{N} \frac{N}{h} (z_{i-1}^N(t) - z_i^N(t)) \chi_i ds - \int_{-h}^{0} \gamma_0(s) \frac{d}{ds} x(t+s) ds \right| \\
&\quad \le \left| \int_{-h}^{0} [\gamma_0^N(s) - \gamma_0(s)] dx(t+s) \right| \\
&\qquad + \left| \int_{-h}^{0} \gamma_0^N(s) \left[\sum_{i=1}^{N} \frac{N}{h} (z_{i-1}^N(t) - z_i^N(t)) \chi_i ds - \frac{d}{ds} x(t+s) \right] ds \right| \\
&\quad \le \sup_{\theta \in [-h, T]} |\dot{x}(\theta)| \frac{h}{N} \mathrm{Var}\gamma_0 + \mathrm{Var}\gamma_0 \int_{-h}^{0} \left| [\lambda^N A^N z^N(t)]^1(s) - \frac{d}{ds} x(t+s) \right| ds,
\end{aligned}$$

where we use the inequality [S1]. From the preliminary result of this proof, the right-hand side converges to zero uniformly in t for $t \in [0, T]$. Therefore,

$$\lim_{N\to\infty} C^N e^{A^N t} \pi^N \phi = CS(t)\phi, \qquad \phi \in D$$

and the limit is uniform in t for t in bounded interval $[0, T]$. The statement in the theorem now follows easily. □

4.2. Nonhomogeneous equations. Having obtained convergence results and estimates of the convergence rate associated with the homogeneous equations we now pose the following question, if—and in what sense—the operators associated with the

nonhomogeneous equations $u \neq 0$ converge. This problem will be considered in this section.

First we will consider the convergence of the output associated with the nonhomogeneous equations. We will show that $C^N S^N(\cdot)B^N$ converges to $CS(\cdot)B$ as a function in $L^2(0,t;R^p)$, uniformly for all $t > 0$. The underlying tool for this convergence proof is the Plancherel's theorem. Before turning to this main result, we will prove the following two lemmas. The first lemma shows that the sequence of operators $C^N R(\lambda, A^N)B^N$ and $CR(\lambda, A)B$ decay on the infinite semi-axis $\alpha+i\omega$, $|\omega| > \beta_1$, where $\beta_1 > 0$. And in the second lemma, it is shown that the sequence $C^N R(\lambda, A^N)B^N$ converges uniformly to $CR(\lambda, A)B$ on some compact set in the complex plane of the form $\{\lambda \in \mathbf{C}/\mathrm{Re}\lambda = \alpha, |\mathrm{Im}\lambda| \leq \beta_1\}$.

LEMMA 4.7. *There exist positive constants* α, β_1, D_{10}, *and* D_{11} *such that*

(i) $|C^N R(\lambda, A^N)B^N| \leq \frac{D_{10}}{|\mathrm{Im}\lambda|}$,

(ii) $|CR(\lambda, A)B| \leq \frac{D_{11}}{|\mathrm{Im}\lambda|}$,

for all $\lambda \in \mathbf{C}$ *with* $\mathrm{Re}\lambda = \alpha, |\mathrm{Im}\lambda| > \beta_1$.

Proof. (i) Since $C^N R(\lambda, A^N)B^N = C^N E_\lambda^N \Delta^N(\lambda)^{-1}(E_\lambda^{T^*})^N F^N B^N$, part (i) follows from the following estimates.

From part (i1) and (i2) of Theorem 4.5,

$$|C^N E_\lambda^N| \leq \mathrm{Var}\gamma_0 \quad \text{for } \mathrm{Re}\lambda > 0$$

and

$$|\Delta^N(\lambda)^{-1}| \leq \frac{2}{|\mathrm{Im}\lambda|}, \qquad |\lambda| > 2\mathrm{Var}\,\eta_0.$$

For $u \in \mathbf{C}^m$,

$$(E_\lambda^{T^*})^N F^N B^N u = B_0 u + \frac{h}{N}\sum_{i=1}^{N}\left(\frac{N}{\lambda h + N}\right)^i B_i^N \frac{N}{h} u;$$

it follows that

$$|(E_\lambda^{T^*})^N F^N B^N u| \leq (|B_0| + \mathrm{Var}\beta)|u|, \quad \mathrm{Re}\lambda > 0.$$

(ii) Since $CR(\lambda, A)B = CE_\lambda \Delta^{-1}(\lambda) E_\lambda^{T^*} FB$, similarly part (ii) follows from the following estimates.

Let $x \in \mathbf{C}^n$, then

$$\begin{aligned} |CE_\lambda x| &= \left|\int_{-h}^{0} d\gamma_0(s) e^{\lambda s} x\right| \\ &\leq \max_{s\in[-h,0]} e^{sRe\lambda}\mathrm{Var}\gamma_0 |x| \leq \mathrm{Var}\gamma_0 |x|, \quad \mathrm{Re}\lambda > 0. \end{aligned}$$

Using the same arguments as part (i2) in Theorem 4.5, we can show that, for $|\lambda| > 2\mathrm{Var}\eta_0$

$$|\Delta^{-1}(\lambda)| \leq \frac{2}{|\mathrm{Im}\lambda|}.$$

Since $|E_\lambda^{T^*} FB| = |(FB)^* E_\lambda^T|$. By [S1, Prop. 5.12]

$$\begin{aligned} (FB)^* E_\lambda^T &= B^T E_\lambda^T \\ &= \int_{-h}^{0} d\beta_0^T(s) e^{\lambda s}, \qquad \beta_0 = -B_0 \chi_{(-\infty,0)}(\theta) + \beta(\theta). \end{aligned}$$

It is easy to show that for $x \in \mathbf{C}^n$, $|(FB)^* E_\lambda^T x| \leq \mathrm{Var}\beta^T |x|$. Hence $|E_\lambda^{T^*} FB| \leq \mathrm{Var}\beta_0^T$. □

LEMMA 4.8. *There exist positive constants α and β_1 such that*

$$\lim_{N\to\infty} C^N R(\lambda, A^N) B^N = CR(\lambda, A)B$$

uniformly on every compact set of the form $S = \{\lambda \in \mathbf{C}/\mathrm{Re}\lambda = \alpha, |\mathrm{Im}\lambda| \leq \beta_1\}$.

Proof. The statement in the lemma is a direct consequence of the following convergence results.

(i) $\lim_{N\to\infty} |CE_\lambda - C^N E_\lambda^N| = 0$ exists uniformly on S.

Let y be an arbitrary vector in C^N. By integration by parts

$$CE^\lambda y = -e^{-\lambda h}\gamma_0(-h)y - \lambda \int_{-h}^{0} e^{\lambda s}\gamma_0(s)ds.$$

Let e_λ^N denote the approximate exponential function given by

$$e_\lambda^N(t) = \left(\frac{N}{\lambda h + N}\right)^i, \qquad t \in [t_i, t_{i-1}).$$

From [S1, Lem. 4.2]

$$C^N E_\lambda^N y = -\gamma_0(-h)e_\lambda^N y - \lambda \int_{-h}^{0} e_\lambda^N(s)\gamma_0^N(s)yds.$$

Hence

$$\begin{aligned} |CE_\lambda y - C^N E_\lambda^N y \leq\ & |\gamma_0(-h)y(e_\lambda^N(-h) - e^{-\lambda h})| \\ & + |\lambda| \int_{-h}^{0} |e^{\lambda s}||\gamma_0^N(s) - \gamma_0(s)||y|ds \\ & + |\lambda| \int_{-h}^{0} |\gamma_0^N(s)||y||e^{\lambda s} - e_\lambda^N(s)|ds. \end{aligned}$$

From the inequality [S1],

$$\begin{aligned} \leq\ & \mathrm{Var}\gamma_0 |e_\lambda^N(-h) - e^{-\lambda h}||y| + |\lambda||y|\frac{h}{N}\mathrm{Var}\gamma_0 \\ & + |\lambda| h \mathrm{Var}\gamma_0 |y| \sup_{s\in[-h,0]} |e^{\lambda s} - e_\lambda^N(s)| \end{aligned}$$

From [S1, Rem. 4.1(ii)], given $\epsilon > 0$ there exists an $N_0 \in \mathbf{N}$ such that

$$\sup_{s\in[-h,0]} |e^{\lambda s} - e_\lambda^N(s)| < \epsilon,$$

for all $\lambda \in S$ and all $N > N_0$. Therefore given $\epsilon_0 > 0$ there exists some N_1 such that for all $N > N_1$,

$$|CE_\lambda y - C^N E_\lambda^N y| < \epsilon_1 |y|$$

or

$$|CE_\lambda - C^N E_\lambda^N| < \epsilon, \qquad \lambda \in S, \quad N > N_1.$$

(ii) $\Delta^N(\lambda)$ converges to $\Delta(\lambda)$ uniformly on every bounded subset of $\mathbf{C}$ ([S1, Lem. 4.2]).

(iii) $\lim_{N\to\infty} |(E_\lambda^{T^*})^N F^N B^N - E_\lambda^{T^*} FB| = 0$ exists uniformly on S.

Let y be an arbitrary vector in $\mathbf{C}^n$. Consider the following adjoint operators

$$(E_\lambda^{T^*} FB)^* y = \int_{-h}^{0} d\beta^T(s) e^{\lambda s} y$$

and after some simple manipulations

$$(B^N)^*(F^N)^*(E_\lambda^T)^N y = \sum_{j=0}^{N} (B_j^N)^T \left(\frac{N}{\lambda h + N}\right)^j y.$$

Hence by the same reasoning as in part (i)

$$\lim_{N\to\infty} |(B^N)^*(F^N)^*(E_\lambda^T)^N - (FB)^* E_\lambda^T| = 0$$

exists uniformly on S and thus completes our proof. □

Using the above two lemmas and the Fourier–Plancherel Theorem, we now obtain the following convergence result of the output corresponding to the nonhomogeneous equations.

THEOREM 4.9. $\lim_{N\to\infty} \int_0^t |C^N S^N(\theta) B^N - CS(\theta)B|^2 d\theta = 0$ *for all* $t > 0$.

Proof. Let denote $G(\lambda) = CR(\lambda, A)B$ and $G^N(\lambda) = C^N R(\lambda, A^N)B^N$. Let $\epsilon > 0$ be given and choose $\beta_1 > 9D_{12}^2/\epsilon$,where the constant $D_{12} = \max\{D_{10}, D_{11}\}$ and D_{10}, D_{11} are the bounds in Lemma 4.7. Then from Lemma 4.7,

$$\begin{aligned}
\int_{-\infty}^{\infty} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega &= \int_{-\infty}^{-\beta_1} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega \\
&\quad + \int_{-\beta_1}^{\beta_1} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega + \int_{\beta_1}^{\infty} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega \\
&\le \int_{-\beta_1}^{\beta_1} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega + 8\int_{\beta_1}^{\infty} \frac{D_{12}^2}{\omega^2} d\omega \\
&= \int_{-\beta_1}^{\beta_1} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega + \frac{8D_{12}^2}{\beta_1}.
\end{aligned}$$

From Lemma 4.8, $\lim_{N\to\infty} |G^N(\lambda) - G(\lambda)| = 0$ exists uniformly in S. Hence there exists an $N_0 \in \mathbf{N}$ such that for all $N \ge N_0$,

$$\int_{-\beta_1}^{\beta_1} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega < \frac{\epsilon}{9}.$$

Therefore,

$$\int_{-\infty}^{\infty} |G(\alpha+i\omega) - G^N(\alpha+i\omega)|^2 d\omega < \frac{\epsilon}{9} + \frac{8D_{12}^2}{\beta_1} < \epsilon.$$

Since the Fourier transform of $e^{-\alpha t} CS(t)B$ is $G(\alpha+i\omega)1/\sqrt{2\pi}$, by Plancherel's theorem

$$\int_0^{\infty} e^{-2\alpha t} |C^N S^N(t) B^N - CS(t)B|^2 dt < \frac{\epsilon}{2\pi}.$$

Now since $e^{2\alpha(t-s)} \geq 1$ for $s \in [0,t]$

$$\int_0^t |C^N S^N(\theta)B^N - CS(\theta)B|^2 d\theta \leq \int_0^t e^{2\alpha(t-\theta)}|C^N S^N(\theta)B^N - CS(\theta)B|^2 d\theta$$
$$\leq e^{2\alpha t}\frac{\epsilon}{2\pi}$$

for $N > N_0$. This completes our proof. □

For the remainder of this section, we will consider the convergence of

$$\int_0^t S^N(t-\theta)B^N u(\theta)d\theta \quad \text{to} \quad \int_0^t S(t-\theta)Bu(\theta)d\theta \ \text{ in } X,$$

where $u \in L^2_{\text{loc}}([0,\infty); R^m)$ and B is an unbounded operator in X.

Let us first recall that for any $u \in R^m$, $Bu = (B_0 u, 0, T_{\delta_0} u)$. After some simple manipulations, it is easily seen that

$$A^{-1}Bu = \left(\Delta^{-1}\left(B_0 u + \int_{-h}^0 d\beta(\theta)u\right), \Delta^{-1}\left(B_0 u + \int_{-h}^0 d\beta(\theta)u\right), -u\right) \tag{6}$$

provided $\Delta = \int_{-h}^0 d\eta_0(\theta)$ is invertible, i.e., 0 is not in the spectrum of A. If Δ is not invertible, we can choose any $\lambda \in \rho(A)$ and consider $x_\lambda(t) = e^{-\lambda t}x(t)$ where $x(t)$ is the solution to the initial-value problem (1). As shown in [I2], the corresponding generator A_λ has $0 \in \rho(A_\lambda)$. Therefore, without loss of generality, we can assume the above formulation (6) of $A^{-1}Bu$.

Since $Gu = A^{-1}Bu \in \text{dom } A$,

$$\int_0^t S(t-\theta)Bu(\theta)d\theta = \int_0^t AS(t-\theta)Gu(\theta)d\theta.$$

If u is continuously differentiable, i.e. $u \in C^1(0,t,R^m)$, then it follows from Kato ([K1, pp. 488–489]) that

$$\int_0^t S(t-\theta)Bu(\theta)d\theta = S(t)Gu(0) - Gu(t) + \int_0^t S(t-\theta)G\dot{u}(\theta)d\theta. \tag{7}$$

From the formulation of A^N and B^N (§3.1), it is straightforward to show that

$$(A^N)^{-1}B^N u = Gu.$$

Hence by applying the above arguments to

$$\int_0^t A^N S^N(t-\theta)Gu(\theta)d\theta$$

we get

$$\int_0^t A^N S^N(t-\theta)Gu(\theta)d\theta = S^N(t)Gu(0) - Gu(t) + \int_0^t S^N(t-\theta)G\dot{u}(\theta)d\theta. \tag{8}$$

From the Trotter–Kato Theorem 4.1,

$$\lim_{N\to\infty} S^N(t)Gu = S(t)Gu \quad \text{in } X$$

for all $t \geq 0$ and the limit is uniform in t for t in bounded intervals. Since $u \in R^m$, the linear map $S(t)G : R^m \to X$ is continuous. Hence

$$\lim_{N\to\infty} \|(S^N(t) - S(t))G\|_{L(R^m,X)} = 0.$$

Therefore, it is now clear that from (7) and (8), for $u \in C^1(0,t;R^m)$

$$\lim_{N\to\infty} \int_0^t S^N(t-\theta)B^N u(\theta)d\theta = \int_0^t S(t-\theta)Bu(\theta)d\theta$$

in X, and the limit is uniform in t for t in bounded intervals. To obtain our desired convergence result for all $u \in L^2_{\text{loc}}([0,\infty);R^m)$ we need the following result.

THEOREM 4.10. *There exists positive constant D_{12}, dependent on t but independent of N, such that for every $u(\cdot) \in L^2_{\text{loc}}(0,t;R^m)$ and all N in* $\mathbf{N}$,

$$\left\| \int_0^t S^N(t-\theta)B^N u(\theta)d\theta \right\|_{X^N} \leq D_{12}\|u\|_{L^2(0,t;R^m)}$$

for each $t \geq 0$.

Proof. The proof uses the same arguments as the stability proof of Theorem 3.1. To this end, we consider the following equation:

$$\langle \dot{z}^N(t), z^N(t)\rangle_{\tau^N} = \langle A^N z^N(t) + B^N u(t), z^N(t)\rangle_{\tau^N},$$

where $z^N(t) = \text{col}\,(z_0^N(t),\ldots,z_N^N(t),v_1^N(t),\ldots,v_N^N(t)) \in R^{n(N+1)+mN}$ is the solution to the approximating system

$$\dot{z}^N(t) = A^N z^N(t) + B^N u(t)$$

with zero initial condition. The right-hand side is equivalent to

$$\begin{aligned}
&\left(A_0 z_0(t) + \sum_{j=1}^N A_j^N z_j(t) + \sum_{j=1}^N B_j^N v_j(t) + B_0 u(t), z_0(t) \right) \\
&\quad + \sum_{i=1}^N (z_{i-1}(t) - z_i(t))^T z_i(t) a_i^N + \sum_{i=2}^N (v_{i-1}(t) - v_i(t))^T v_i(t) a_i^N \\
&\quad - |v_1(t)|^2 a_1^N + u^T(t) v_1(t) a_1^N.
\end{aligned}$$

From part (i)–(iv) of Theorem 3.1,

$$\langle \dot{z}^N(t), z^N(t)\rangle_{\tau^N} \leq \frac{1}{2}\beta\|z^N(t)\|_N^2 + |u(t)|^2,$$

where $\beta = 1 + a_1^N + \sum_{i=0}^q |A_i|^2 + \sum_{i=0}^q |B_i|^2 + \|A_{01}\|_{L^2}^2 + \|B_{01}\|_{L^2}^2$. Since the left-hand side is $\frac{1}{2}\frac{d}{dt}\|z^N(t)\|_{\tau^N}^2$ and $\|\cdot\|_N \leq \|\cdot\|_{\tau^N} \leq \gamma\|\cdot\|_N$,

$$\frac{d}{dt}\|z^N(t)\|_{\tau^N}^2 \leq \beta\|z^N(t)\|_{\tau^N}^2 + 2|u(t)|^2.$$

For zero initial condition

$$\|z^N(t)\|^2_{\tau^N} \leq \int_0^t (\beta \|z^N(s)\|^2_{\tau^N} + 2|u(s)|^2) ds.$$

By Gronwall's ineqaulity, $\|z^N(t)\|^2_{\tau^N} \leq 2e^{\beta t} \int_0^t |u(s)|^2 ds$. The result then follows from the equivalence of the norm in X^N and the weighting norm τ^N. □

5. Concluding Remarks. We have considered the averaging approximation scheme for linear retarded functional differential equations with delays in control and observation. We have shown that known results on averaging approximation of retarded systems with only delays in the state can be extended to include delays in input and output. These are the convergence results of the approximating semigroups in the strong operator topology in the context of Trotter–Kato approximation theorem [B1] and estimates of its rate of convergence using the concept of differentiable semigroups [L1]. In addition, we also gave new convergence results of the output and the state corresponding to both homogeneous and nonhomogeneous equations in the presence of unbounded input and output operators. Finally, the present development gives a basis for numerical investigations of control problems, in particular using the operator Riccati equations.

Acknowledgments. The second author is very pleased to acknowledge discussions with K. Ito of the Center for Applied Mathematical Sciences, University of Southern California (current address, Center for Research in Scientific Computation, North Carolina State University) for the proof of Theorem 4.10, and thanks D. Salamon of the Control Theory Centre, University of Warwick, for his invaluable help and suggestions.

REFERENCES

[B1] H. T. BANKS AND J. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169–208.

[B2] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.

[D1] M. C. DELFOUR AND J. KARRAKCHOU, *State space theory of linear time invariant systems with delays in state, control and observation variables*, Part I: J. Math. Anal. Appl., 125 (1987), pp. 361–399; Part II: J. Math. Anal. Appl., 125 (1987), pp. 400–450.

[D2] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded systems*, Part I: J. Math. Anal. Appl., 73 (1980), pp. 466–490, Part II: J. Math. Anal. Appl., 74 (1980), pp. 359–381.

[G1] J. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.

[H1] E. HEWITT AND K. A. ROSS, *Abstract Harmonic Analysis I*, 2nd ed., Springer-Verlag, New York, 1979.

[I1] A. ICHIKAWA, *Quadratic control of evolution equations with delays in control*, SIAM J. Control Optim., 20 (1982), pp. 645–668.

[I2] K. ITO AND R. TEGLAS, *Legendre-Tau approximations for functional differential equations*, SIAM J. Control Optim., 24 (1986), pp. 737–759.

[I3] K. ITO AND H. T. TRAN, *Linear Quadratic Optimal Control Problem for Linear Systems with Unbounded Input and Output Operators: Numerical Approximations*, in Internat. Ser. Numer. Math., Vol. 91, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser, Boston, 1989, pp. 171–196.

[K1] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[K2] N. N. KRASOVSKII, *Approximation of an optimal control problem for a system with delay*, Soviet Phys. Dokl., 11 (1966), pp. 219–221.

[L1] I. LASIECKA AND A. MANITIUS, *Differentiability and Convergence Rates of Approximating Semigroups for Retarded Functional Differential Equations*, SIAM J. Numer. Anal., 25 (1988), pp. 883–907.

[P1] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci., Vol. 44, Springer-Verlag, New York, 1983.

[P2] A. J. PRITCHARD AND D. SALAMON, *The Linear Quadratic Optimal Control Problem for Infinite Dimensional Systems with Unbounded Input and Output Operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[R1] Y. M. REPIN, *On the Approximate Replacement of Systems with Lag by Ordinary Differential Equations*, J. Appl. Math. Mech., 29 (1965), pp. 254–264.

[S1] D. SALAMON, *Structure and Stability of Finite Dimensional Approximations for Functional Differential Equations*, SIAM J. Control Optim., 23 (1985), pp. 928–951.

[S2] D. SALAMON, *On Control and Observation of Neutral Systems*, Research Notes in Mathematics 91, Pitman, London, 1984.

[T1] H. T. TRAN, *Numerical Studies of the Linear Quadratic Control Problem for Retarded Systems with Delay in Control*, in Progress in Systems and Control Theory, K. Bowers and J. Lund, eds., Birkhäuser, Boston, 1991, pp. 307–324.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1364–1427, September 1994

009

ROOT-LOCUS AND BOUNDARY FEEDBACK DESIGN FOR A CLASS OF DISTRIBUTED PARAMETER SYSTEMS*

CHRISTOPHER I. BYRNES†, DAVID S. GILLIAM‡, AND JIANQIU HE‡

This paper is presented in memory of the short but promising mathematical career of Jagath Chandrawansa.

Abstract. In this paper, a fairly complete parallel of the finite-dimensional root locus theory is presented for quite general, nonconstant coefficient, even order ordinary differential operators on a finite interval with control and output boundary conditions representative of a choice of collocated point actuators and sensors. Root-locus design methods for linear distributed parameter systems have also been studied for some time and the primary difficulties in rigorously interpreting root-locus conclusions for distributed parameter systems are well known. First, the transfer function of a distributed parameter system may not be meromorphic at infinity so that many of the standard Rouche arguments, required even in the lumped case to determine the asymptotic behavior of the root loci, are not generally valid. Another difficulty is that the infinitesimal generator in the state-space model for a closed-loop system may not be selfadjoint, accretive or even satisfy the spectrum determined growth condition. Thus, regardless of whether the root loci—interpreted as closed-loop eigenvalues—lie in the open left half-plane, additional analysis would be required to conclude that the closed-loop system would be asymptotically stable. Formulating the systems in the classical format of a boundary control problem, the asymptotic analysis of the root loci can be based on the pioneering work by Birkhoff on eigenfunction expansions for boundary value problems, work that predated and indeed motivated the development of spectral theory in Hilbert space. Birkhoff's work also contains an asymptotic expansion of eigenfunctions in the spatial variable, generalizing the earlier Sturm–Liouville theory for second-order operators. By further extending this general asymptotic analysis to also include expansions in the gain parameter, a rigorous treatment of the open- and closed-loop transfer functions and of the corresponding return difference equation can be presented. The asymptotic analysis of the return difference equation forms the basis for both the rigorous formulation of the basic problem and its solution.

Key words. root locus, distributed parameter systems, boundary feedback, control and sensors, transfer function, impulse response, Riesz basis

AMS subject classification. 93

1. Introduction.

Root-locus plots were invented by Evans over forty years ago as a simple graphical tool for analyzing closed-loop stability properties for feedback systems. Based on a few simple rules, we can sketch the evolution of the closed-loop poles of certain feedback systems as the feedback gain is varied from zero to either plus or minus infinity. While originally derived for proportional error (PE) feedback systems, the root-locus methodology is also a useful tool for the design of stabilizing proportional rate (PD) controllers and more general dynamic compensators. The graphical appeal of this "back of the envelope" design tool, now available in standard control software packages, was responsible for the widespread and continued use of root-locus methods in industrial control system design.

Not surprisingly, root-locus design methods for linear distributed parameter systems have also been studied for some time. For example, motivated by the desire to control rigid spacecraft with flexible appendages, Bryson and several of his students (see, e.g., [4], [28], [42]) developed closed-loop root-locus plots for PE and PD bound-

* Received by the editors November 25, 1991; accepted for publication (in revised form) March 19, 1993. This work was supported in part by grants from the Air Force Office of Scientific Research, the National Science Foundation, and the Texas Advanced Research Program.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

‡ Department of Mathematics, Texas Tech University, Lubbock, Texas, 79409.

ary controllers for various wave and beam equations. Pohjolainen [31], [32], [33] has also presented some root-locus arguments for distributed parameter systems. Even though there are infinitely many branches of the root loci for distributed parameter systems, such feedback analysis and design methods retain the relative simplicity and intuitive appeal of classical automatic control and no doubt lead to stabilizing, low-dimensional controllers when used correctly. Nonetheless, the rigorous analysis of the asymptotic behavior of the infinitely many branches of the root-locus and of the apparently resulting closed-loop stability properties for non-selfadjoint boundary value problems have been open problems. Moreover, because of the engineering appeal of root-locus plots as a graphical basis for feedback design, it is important to resolve these questions in a systematic fashion. In this paper, we present a fairly complete parallel of the finite-dimensional root-locus theory for quite general, nonconstant coefficient, even order ordinary differential operators on a finite interval with boundary conditions whose highest-order terms are separated (i.e., occur at one end point or the other, as in the case of colocated actuators and sensors).

The primary difficulties in rigorously interpreting root locus conclusions for distributed parameter systems are well known. First, the transfer function of a distributed parameter system may not be meromorphic at infinity so that many of the standard Rouche arguments, required even in the lumped case to determine the asymptotic behavior of the root loci, are not generally valid. Related to this, but also important in its own right, is the apparent difficulty of obtaining a well-defined notion of "high frequency" or "instantaneous" gain, the sign of which is crucial in determining the direction of the root locus as well as in a variety of stability considerations throughout classical automatic control. More explicitly, this response gain is classically computed either as the residue of the transfer function at infinity, which is an essential singularity for the systems considered here, or as the value of the impulse response at time zero, which can also be seen to be a singular point by an elementary Dirichlet series argument. Nonetheless, many physical examples suggest a preferred choice of the sign of the feedback gain, reflecting the existence of what should be an instantaneous gain—at least between colocated actuators and sensors. These examples also display closed-loop root loci with markedly different asymptotic behavior as the gain parameter goes to either plus or minus infinity, differences far more exotic than in finite dimensions. Thus, one important corollary to a reasonably complete root locus theory would be a rigorous formulation and treatment of a "high frequency" or "instantaneous" gain.

Another difficulty is, of course, that the infinitesimal generator in the state-space model for a closed-loop system may not be selfadjoint, accretive or even satisfy the spectrum determined growth condition. Thus, regardless of whether the root loci—interpreted as closed-loop eigenvalues—lie in the open left half-plane, additional analysis would be required to conclude that the closed-loop system would be asymptotically stable. Compounding this problem is the fact that, for systems arising from boundary value problems for partial differential equations, we are often interested in point actuators and sensors; i.e., input and output sources occurring at points in the spatial domain. The operators representing these sensors and actuators are seldom bounded (in fact, not even closable) in the standard Hilbert state space defined by the boundary conditions. Thus, to formulate such problems in the state-space form it is required to introduce auxiliary spaces as in [11], [15], [16], [18], [21], [22], [24], [38], [35], [39]–[41]. While such a state-space representation is esthetically appealing, it is nevertheless true that obtaining explicit formulas for the operators B and C can

generally be quite difficult and is in any case not explicitly required in the formulation of a rigorous root locus methodology for boundary feedback controller design.

In §2 we have formulated our systems in the more classical format of a boundary control problem but we also indicate how this class of problems can be cast in terms of abstract boundary control systems, as in [14], [21]. This choice of a classical formulation enables us to base our asymptotic analysis of the root loci on the pioneering work by Birkhoff [1], [2], which consisted essentially of the development of a spectral theory for not necessarily selfadjoint boundary value problems. Of course, Birkhoff's work on eigenfunction expansions for boundary value problems predated, and indeed motivated (see, e.g., [36]), the development of spectral theory in Hilbert space. In particular, with some work in the non-selfadjoint case, by viewing changes in the feedback gain as perturbations of the boundary conditions an alternative approach to root locus analysis could also be carried out by appealing to the modern perturbation theory of unbounded spectral operators, which, however, can lead to subtle technical problems even in the selfadjoint case (cf. Example 3.4). Birkhoff's work also contains an asymptotic expansion of eigenfunctions in the spatial variable, generalizing the earlier Sturm–Liouville theory for second-order operators. By further extending this general asymptotic analysis to also include expansions in the gain parameter, we are able to present a rigorous treatment of the open- and closed-loop transfer functions and of the corresponding return difference equation, all familiar objects in classical automatic control. We remark that, by emphasizing the analysis of the roots of the return difference equation rather than just the closed-loop poles, we avoid difficulties arising in pole-zero cancellations and are thereby able to ultimately obtain results on internal stability of the original boundary value problems. Thus, the asymptotic analysis of the return difference equation forms the basis for both our rigorous formulation of the basic problem and its solution, as described in §§2 and 3, respectively.

In §2, we introduce a hypothesis on the relative orders of the input and the output boundary operators. Informally, we have found it useful to think of this as a "causality" condition, asserting that the "relative degree" of the system is nonnegative. Indeed, we show that the transfer functions for this class of systems exist, lie in the Callier–Desoer class, are strictly proper, and thus are holomorphic and, in fact, vanish at infinity in a right half-plane. We then develop an asymptotic expansion of the transfer function, in a half-plane, in terms of a fractional power series with exponent depending on the orders of the differential operator and of the boundary conditions. One very important corollary of this analysis is the formulation of the concept of "high frequency" or "instantaneous" gain, defined as the leading coefficient in this asymptotic expansion. Moreover, a formula for the sign of the instantaneous gain can be given in terms of the boundary conditions. As shown in §3, the sign of the instantaneous gain plays the expected fundamental role in our subsequent analysis of the root loci and of the spectral properties of the closed-loop system, an analysis that also depends rather heavily on the asymptotics of the return difference equation.

As a preliminary to the statements of the main results on root locus plots and spectral properties for closed-loop distributed parameter systems (DPS), we begin §3 with a series of examples illustrating some of the important contrasts with root locus theory for lumped systems. These differences arise, of course, from the fact that for the systems we consider, the transfer function is not rational but rather always has an essential singularity at infinity. Example 3.1 illustrates the possibility, in the absence of our hypotheses, that the closed-loop infinitesimal generator can have, for various values of the feedback gain, either discrete spectrum, or a continuum (the entire

complex plane) of point spectrum, or empty spectrum. Examples 3.2 and 3.3 are more subtle and illustrate the effect of the essential singularity at infinity on the asymptotic behavior of the root-locus plot for DPS, a situation clarified by the definition of the instantaneous gain and the statement of our main results. Among these, we note in Theorem 3.1 that the closed-loop operators form a holomorphic family. However, as an example of Rellich (adapted to our setting in Example 3.4) shows, from this it can only be argued that finite systems of eigenvalues vary continuously. In this context, the real difficulty in establishing a root-locus theory for this class of distributed systems lies in showing that all branches of the root locus vary continuously as the gain is varied from zero to either plus or minus infinity. With our additional "causality" hypothesis on the relative orders of the input and output boundary conditions, our asymptotic analysis of the return difference equation allows us to prove more refined continuity results for the root loci (cf. Theorem 3.2, Corollary 3.3, and Theorem 3.4) by verifying that we have "separation of the spectrum" for the one-parameter family of operators.

We conclude §3 with an outline of the asymptotic analysis of the return difference equation. In §4 we provide a detailed asymptotic analysis of the return difference equation, except for the proof of Proposition 4.2, which is tedious and therefore appended as §6. Section 5, based on the previous section, contains the complete proofs of the results announced in §2 and the proof of Theorem 3.4. In general, we recommend first reading the outline provided in §3, then skipping to §5 after which the reader can return to the proofs of the asymptotics in §§4 and 6.

Finally, we wish to address the general scope of applicability of these methods. Our immediate goal is to demonstrate that it is possible to obtain a fairly complete analogue of the finite-dimensional root-locus theory—at least for proportional error feedback laws and for a class of parabolic boundary control problems. On the one hand, using standard variational methods, an extension of this analysis to higher-dimensional selfadjoint problems can be obtained [10]. However, as far as we are aware, general methods for spectral analysis for non-selfadjoint problems are available only in one spatial dimension (see, e.g., [20]) or for abstract operators that are either maximal dissipative or for which the eigenfunctions form a Riesz basis. For nth-order ordinary differential operators, the corresponding boundary conditions considered in [1], [2] are now known as "Birkhoff regular" boundary conditions and, parameterizing a system of boundary conditions by an n by $2n$ matrix of rank n, we see that Birkhoff regular boundary conditions are generic among all boundary conditions. In this paper, we consider the case of separated boundary conditions, corresponding to colocated actuators and sensors. Separated boundary conditions are automatically Birkhoff regular. In [23] general Birkhoff regular boundary conditions are considered for parabolic systems. We have also investigated the extension of these methods to the hyperbolic case (see [8]), where our preliminary analysis indicates that it is also possible to obtain a fairly complete analogue of the finite-dimensional root-locus theory. Moreover, as in the classical case, the "root-locus" plots developed for distributed parameter systems can be used to study a broader set of design problems, e.g., PD controller design (see [8]), than just that of stabilization by proportional error feedback gain.

2. Formulation of the problem in the time and frequency domains. In this section, we describe the interpretation of root loci in both the state space and the frequency domain. All proofs are deferred to §4. In what follows, we consider a

distributed parameter control system

$$\begin{aligned} \frac{\partial w}{\partial t}(x,t) &= \mathcal{A}w(x,t), \\ \mathcal{B}w(t) &= u(t), \\ w(x,0) &= f(x) \in L^2(0,1), \\ y(t) &= \mathcal{C}w(t), \end{aligned} \tag{2.1}$$

where $\mathcal{A}$ is an even order, $n = 2\mu$, ordinary differential operator, with $C^\infty[0,1]$, real coefficients, of the form

$$\begin{aligned} \mathcal{A} &= L_0 + L, \\ L_0 &= (-1)^{(\mu-1)}D^n, \qquad D = \frac{d}{dx} \\ L &= \sum_{j=0}^{n-2} p_j(x)D^j \end{aligned} \tag{2.2}$$

acting in the state space $L^2(0,1)$ with domain

$$\mathcal{D}(\mathcal{A}) = \{f \in H^{(n)}(0,1) : \mathcal{W}_i(f) = 0,\ i = 2,\ \ldots, n\}, \tag{2.3}$$

where $H^{(n)}$ is the usual Sobolev space and the operators $\{\mathcal{W}_i\}_{i=0}^n$ are boundary operators providing homogeneous boundary conditions defining $\mathcal{A}$ for $i = 2,\ldots,n$ and defining the output and input operators $\mathcal{C}$ and $\mathcal{B}$ for $i = 0$ and $i = 1$, respectively (cf. (2.5), (2.6) below). Acting on C^{n-1} functions f by these operators are given by

$$\begin{aligned} \mathcal{W}_i(f) &\equiv \alpha_i f^{(m_i)}(0) + \sum_{j=0}^{m_i-1} \left\{\alpha_{ij} f^{(j)}(0) + \beta_{ij} f^{(j)}(1)\right\}, \\ \mathcal{W}_{\mu+i}(f) &\equiv \beta_i f^{(m_{\mu+i})}(1) + \sum_{j=0}^{m_{\mu+i}-1} \left\{\alpha_{(\mu+i)j} f^{(j)}(0) + \beta_{(\mu+i)j} f^{(j)}(1)\right\}, \end{aligned} \tag{2.4}$$

where $i = 1,\ldots,\mu$, $\alpha_i \neq 0$, $\beta_i \neq 0$, α_{ij}, β_{ij} are real and

$$m_1,\ m_{\mu+1} \leq (n-1), \quad m_1 > \cdots > m_\mu, \quad m_{\mu+1} > \cdots > m_n$$

and $\mathcal{W}_0$ is defined in (2.6).

The input $u(t)$ is assumed to occur through the boundary as in [5] and is assumed to occur through the first boundary condition $\mathcal{W}_1$, i.e.,

$$\mathcal{B}w(t) \equiv \mathcal{W}_1(w)(t) = u(t). \tag{2.5}$$

It is further assumed that the output sensor has the form

$$\begin{aligned} y(t) &= \mathcal{C}(w)(t) \equiv \mathcal{W}_0(w)(t), \\ \mathcal{W}_0(w)(t) &= \alpha_0 w^{(m_0)}(0,t) + \sum_{j=0}^{m_0-1} \left\{\alpha_{0j} w^{(j)}(0,t) + \beta_{0j} w^{(j)}(1,t)\right\}. \end{aligned} \tag{2.6}$$

ASSUMPTION 2.1. *The order m_0 of $\mathcal{W}_0$ is not equal to the orders $\{m_i\}_{i=1}^\mu$ of $\{\mathcal{W}_i\}_{i=1}^\mu$ and $\ell = m_1 - m_0 > 0$.*

Remark 2.1. As illustrated in Example 3.3, the assumption $\ell > 0$ is a time domain analogue of strict properness of a transfer function representation. In the time domain, Assumption 2.1 is consistent with the situation found for admissible controls as described in [24], [27], where for the heat equation it is seen that Dirichlet boundary input leads to an inadmissible input operator (cf. Example 3.3).

Returning to (2.1)–(2.6), the uncontrolled system (i.e., $u = 0$ in (2.1)) has the form

$$\begin{aligned} \frac{\partial w}{\partial t}(x,t) &= A_0 w(x,t), \\ \mathcal{B}w(t) &= 0, \\ w(x,0) &= f(x) \in L^2(0,1), \\ y(t) &= \mathcal{C}w(t), \end{aligned} \tag{2.7}$$

where we have introduced the open-loop spatial operator $A_0 = \mathcal{A}$ with domain

$$\mathcal{D}(A_0) = \mathcal{D}(\mathcal{A}) \cap \ker(\mathcal{B}), \tag{2.8}$$

which we can write as

$$\mathcal{D}(A_0) = \{f \in H^{(n)}(0,1) : \mathcal{W}_i(f) = 0,\ i = 1,\dots,n\}. \tag{2.9}$$

Remark 2.2. The operator A_0 with domain $\mathcal{D}(A_0)$ is a special case of the class of Birkhoff regular operators considered in [1], [30]. It follows from [29] that in the special case of boundary conditions (2.4), which are separated in the highest-order terms, the operator (2.2) with domain (2.9) is always a discrete spectral operator whose eigenfunctions and associated functions form a Riesz basis in $L^2(0,1)$. In fact as will be seen in what follows, much more can be said in this case. Furthermore, with significant additional work, the analysis presented here can be applied to more general Birkhoff regular boundary conditions and also to conditions that are not regular (e.g., boundary conditions of the type (2.4) with an unequal number of conditions at the end points). The details are considerably more complicated and the results not as easily stated as in the present case.

DEFINITION 2.1. *The zero dynamics associated with the system* (2.1)–(2.6) *is the system obtained by constraining the output to zero*

$$\begin{aligned} \dot{w}(x,t) &= \mathcal{A}w(x,t), \\ \mathcal{W}_0(w)(t) \equiv \mathcal{C}w(t) &= 0, \\ w(x,0) &= f(x) \in L^2(0,1). \end{aligned} \tag{2.10}$$

The zero dynamics (2.10) can be expressed in terms of the notation introduced in (2.8) by defining the operator $A_\infty = \mathcal{A}$ with domain

$$\mathcal{D}(A_\infty) = \mathcal{D}(\mathcal{A}) \cap \ker(\mathcal{C}), \tag{2.11}$$

which we can write as

$$\mathcal{D}(A_\infty) = \{f \in H^{(n)}(0,1) : \mathcal{W}_i(f) = 0,\ i = 0,2,\dots,n\}. \tag{2.12}$$

A closed-loop system is obtained via a simple scalar boundary output feedback law of the form

$$u(t) = -ky(t). \tag{2.13}$$

Formulated in terms of perturbation of spectra of unbounded spectral operators, we would define a family of operators depending on the parameter k by introducing the spatial operator A_k as the operator $\mathcal{A}$ subject to perturbed boundary conditions obtained from the feedback law (2.13), i.e., $\mathcal{B}(w) + k\mathcal{C}(w) = 0$ or

$$\mathcal{W}_1(w) + k\,\mathcal{W}_0(w) = 0.$$

Thus we define

$$\begin{aligned} A_k &= A \\ D(A_k) &= \{f \in L^2(0,1) \,:\, f \in H^{(n)}(0,1), \\ &\qquad \mathcal{W}_i(f) = 0,\ i = 2,\ \ldots, n,\ \mathcal{W}_1(f) + k\,\mathcal{W}_0(f) = 0\} \end{aligned} \tag{2.14}$$

and the resulting closed-loop system has the form

$$\begin{aligned} \dot{w}(x,t) &= A_k w(x,t), \\ w(x,0) &= f(x). \end{aligned} \tag{2.15}$$

We could, of course, allow more general input and consider a closed-loop system corresponding to a feedback law of the form

$$u(t) = -ky(t) + v(t) \tag{2.16}$$

by defining a closed-loop system with additional input by

$$\begin{aligned} \dot{w}(x,t) &= \mathcal{A}w(x,t), \\ (B(w) + kC(w))(t) &= v(t), \\ y(t) &= C(w), \\ w(x,0) &= f(x). \end{aligned} \tag{2.17}$$

The stability of particular examples of system (2.17) for judicious choices of the gain parameter k has long been an object of study in the area of boundary feedback stabilization and control. The main problem that motivates the present work is the development of a systematic methodology, similar in scope to automatic control, for determining general dynamical properties of (2.17) as a function of the gain parameter. While such a methodology would include in its goals the capability of shaping the system response by tuning one or more gain parameters, for linear systems many problems of regulation and control repose upon the ability to design stabilizing feedback laws. In this context, one particular problem of interest would be to deduce stability properties of A_k for k sufficiently large, from stability properties of the zero dynamics. More generally, we would certainly like to know whether the solution semigroups would vary continuously as a function of the gain parameter in the extended real line.

To this end, it will be important to analyze the behavior of the family of semigroups defined by (2.17); hence our study of this problem incorporates a state space analysis of this class of systems. On the other hand, most of the graphical stability tests in classical automatic control were developed in the frequency domain. In this spirit, the present work provides the rigorous development of a graphical criterion for stability analysis based on a generalization of the finite-dimensional root-locus theory, which provides a simple set of rules for determining the evolution of the closed-loop

poles as functions of the gain parameter. This methodology reposes, of course, on the development of a transfer function description of the underlying distributed parameter system that is both rigorous and easy to relate to the time domain or state space representation.

By a transfer function for the system (2.1)–(2.6) we adopt the following definition, which is more restrictive than it need be but suffices for the systems considered here. First we define the Hilbert space of inputs

$$U_a = \{f : (0,\infty) \to \mathbb{C} : \exp(-a\cdot)f(\cdot) \in L_2(0,\infty),\ a \in \mathbb{R}\} \tag{2.18}$$

with inner product

$$< f, g >_a = \int_0^\infty f(t)g(t)e^{-2at}\,dt.$$

For the single input-single output systems considered in the present work, the space of outputs coincides with the space of inputs. In this context, one of our reasons for adopting Assumption 2.1 was to obtain a system transfer function that lies in $H^\infty(C_+^a)$ (where $C_+^a = \{\lambda \in C | \operatorname{Re}(\lambda) > a\}$) for some $a \in R$.

DEFINITION 2.2. *A transfer function for* (2.1)–(2.6) *is a function* $\mathcal{G}_0(\cdot) \in H^\infty(C_+^a)$ (*where* $C_+^a = \{\lambda \in C | \operatorname{Re}(\lambda) > a\}$) *such that for* $u \in U_a$ *there is a corresponding output* $y \in U_a$ *related via the Laplace transform by*

$$\hat{y}(\lambda) = \mathcal{G}_0(\lambda)\hat{u}(\lambda).$$

PROPOSITION 2.1. *The system* (2.1)–(2.6) *with* $w(x,0) = 0$ *has a transfer function with the form*

$$\mathcal{G}_0(\lambda) = \frac{\mathcal{N}(\lambda)}{\mathcal{D}(\lambda)}, \tag{2.19}$$

where $\mathcal{D}$, $\mathcal{N}$ *are entire functions of* λ *of order* $(1/n)$ *with infinitely many zeros diverging to infinity; these zeros are denoted by* $\lambda_j(0)$, *and* $\lambda_j(\infty)$, *respectively, and are here referred to as the open-loop poles and open-loop zeros. Moreover, the transfer function* $\mathcal{G}_0$ *is real, i.e.,*

$$\mathcal{G}_0(\bar{\lambda}) = \overline{\mathcal{G}_0(\lambda)},$$

from which we conclude that the complex poles and zeros occur in conjugate pairs.

The closed-loop transfer function, i.e., the transfer function for the system (2.17) *corresponding to the feedback law*

$$u = -ky + v$$

is

$$\mathcal{G}_k = \frac{\mathcal{N}(\lambda)}{\mathcal{D}(\lambda) + k\mathcal{N}(\lambda)}. \tag{2.20}$$

Furthermore, for every k *the solutions of the return difference equation*

$$\mathcal{D}(\lambda) + k\mathcal{N}(\lambda) = 0$$

correspond exactly to the spectrum of the closed-loop operator A_k *defined in* (2.14).

Remark 2.3. According to Proposition 2.1, the problem of showing that the spectra of the operators A_k varies continuously as k varies from $k = 0$ to $k = \infty$ is equivalent to showing that the roots of the return difference equation vary from the zeros of $\mathcal{D}$ to the zeros of $\mathcal{N}$. The root-locus theory we seek to develop would provide graphical information on the variation of either the eigenvalues or the roots of the return difference equation as functions of the gain parameter, as well as information on stability of the closed-loop system.

Remark 2.4. As in the open-loop case, we refer to the zeros of $\mathcal{N}$ as the closed-loop zeros and to the zeros of $\mathcal{D}+k\mathcal{N}$ as the closed-loop poles. This convention is potentially different from the zeros and poles of the transfer function $\mathcal{G}_k$ treated as a meromorphic function. Our convention is adopted from the point of view of internal stability relative to the realizations (2.17). For example, the assertion in Proposition 2.1 is potentially stronger and preferable to the corresponding assertion for the poles and zeros of the closed-loop transfer function. More precisely, the fact that for some k the poles of the closed-loop transfer function might lie in the open left half-plane would not imply that the spectrum of A_k would also lie in the open left half-plane if there were a pole-zero cancellation in $\mathcal{G}_k$. Internal stability relative to the realization (2.17) would require information on the full spectrum rather than on those poles (eigenvalues) that are not cancelled especially if the cancellation would occur in the closed right half-plane.

Remark 2.5. Pole-zero cancellation for this class of problems is, however, far from mysterious. First, from the asymptotic analysis developed in §4 it follows that all but a finite number of the solutions of the return difference equation vary with k. On the other hand, to say a pole-zero cancellation occurs in $\mathcal{G}_k$, for some k, is to say that both $\mathcal{N}$ and $\mathcal{D}+k\mathcal{N}$ vanish at some λ_0 in the complex plane. That is, both $\mathcal{N}$ and $\mathcal{D}$ vanish at λ_0 so that a pole-zero cancellation occurs at λ_0 for every k. In particular these already occur $k = 0$ and are finite in number counting multiplicity. We finally note that, in the light of this discussion, to say that the closed-loop poles tend to the open-loop zeros as k tends to infinity is to say that the poles of the closed-loop transfer function tend to the zeros of the open-loop transfer function, and conversely. The actual convergence of these poles will be discussed in the next section.

PROPOSITION 2.2. *For any system* (2.1)–(2.6), *the transfer function* $\mathcal{G}(\lambda)$ *is in* $H^\infty(C_+^a)$ (*where* $C_+^a = \{\lambda \in C | \operatorname{Re}(\lambda) > a\}$) *for some* $a \in R$ *and is strictly proper* (*cf.* [17]); *i.e.,*

$$\lim_{\lambda \to \infty} \mathcal{G}(\lambda) = 0$$

for $\lambda \in C_+^a$. *Furthermore, the impulse response satisfies*

$$h(t) = h_1(t) + h_2(t)$$

with $h_1 \in L^1_{loc}(0, \infty)$, $e^{-at}h_1 \in L^1(0, \infty)$ *and* $h_2 \in L^1(0, \infty)$. *Finally, the input-output map given by* $y(t) = (h * u)(t)$ *defines a bounded map on* U_a.

This result states in particular that the transfer function not only exists but lies in the Callier–Desoer class [17]. A particular consequence of this fact is of immense importance for developing a root-locus theory for this class of distributed parameter systems. More explicitly, one invariant of a rational transfer function that plays an important role in classical automatic control is its high frequency gain; i.e., its residue at infinity. This is also expressible in the time domain as the "instantaneous gain," which can be computed as the system response to a delta function at time zero. In

the classical case, the instantaneous gain can also be computed as the value of the impulse response function at time zero; namely, $h(0)$. Formally, it would appear that for this class of distributed parameter systems neither the high frequency gain nor the instantaneous gain would exist. Indeed, infinity is an essential singularity of the transfer function. In the time domain, we can only assert that h lies in $L^1_{\text{loc}}(0,\infty)$ and, in fact, a straightforward Dirichlet series argument shows that h is singular at $t = 0$. Nonetheless, physically motivated examples (cf. Example 3.2) indicate that, at the very least, there exists the analogue of the signum of the instantaneous gain, which is in fact the quantity used in classical control design. We can now describe this in the frequency domain in terms of an asymptotic expansion in C^a_+

PROPOSITION 2.3. *In C^a_+ we have*

$$\lim_{\lambda\to\infty} \lambda^{\ell/n}\, \mathcal{G}_0(\lambda) = \widehat{\tau}$$

for some nonzero real number $\widehat{\tau}$ where the limit is taken on the positive real axis.

DEFINITION 2.3. *For a system* (2.1)–(2.6) *satisfying Assumption* 2.1, *we refer to the number $\widehat{\tau}$ defined in Proposition* 2.3 *as the "instantaneous gain."*

As in the finite-dimensional case, it will be important for control system design to be able to compute the sign of $\widehat{\tau}$. To this end, we set the notation

$$\ell_0 = \min\{j : m_j < m_0, \quad j = 2, \dots, \mu\},$$

with the convention that ℓ_0 is taken to be zero if no such j exists. Note that in the case of second-order operators ℓ_0 is always zero.

The next result gives a formula for signum of $\widehat{\tau}$ in terms of the input and output boundary operators.

THEOREM 2.4. *Denote by $\mathcal{K}$ the signum of the instantaneous gain, $\widehat{\tau}$. Then $\mathcal{K}$ satisfies*

$$\mathcal{K} = (-1)^s,$$

where

$$s = \arg(\alpha_0/\alpha_1)/\pi + (\ell_0 + \ell), \tag{2.21}$$

with $\ell = m_1 - m_0 > 0$, and α_0, α_1 and m_0, m_1 are the coefficients of the highest-order terms and orders of the output and input operators, respectively.

The instantaneous gain also plays a fundamental role in the finite behavior of $\mathcal{G}_0(\lambda)$.

PROPOSITION 2.5. *For the system* (2.1)–(2.6) *we have the following facts concerning the transfer function.*

1. *The residues corresponding to poles of large modulus are real and of the same sign.*

2. *Suppose λ_j is a pole of large modulus of $\mathcal{G}_0(\lambda)$, $\lambda_j = i^n z_j^n$ where $n = 2\mu$ is the order of $\mathcal{A}$, $\ell > 0$ is the difference of the orders of the input and output operators, s is the integer determining the sign of the instantaneous gain* (*given in* (2.21)). *Then*

$$\text{Res}_{\lambda=\lambda_j}\, \mathcal{G}_0(\lambda) = (-1)^s n|\tau| \sin(\ell\pi/n)|z_j|^{n-\ell-1}[1],$$

where for a complex a, we use the notation $[a] = a + O(1/z)$ for large $|z|$.

3. *The inverse Laplace transform of $\mathcal{G}_0(\lambda)$ can be computed by the method of residues and produces an impulse response function as described in Proposition 2.2. Indeed, all but finitely many poles are real and simple; we denote the real simple poles by λ_j for $j > N+1$. Then we obtain a formula*

$$
\begin{aligned}
h(t) &\equiv \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} e^{st}\,\mathcal{G}_0(s)\,ds \\
&= \sum_{j=1}^{\infty} \mathrm{Res}_{\lambda=\lambda_j}\left(\mathcal{G}_0(\lambda)e^{\lambda t}\right) \qquad (2.22)\\
&= \left\{ \sum_{j=1}^{N} \mathrm{Res}_{\lambda=\lambda_j}\left(\mathcal{G}_0(\lambda)e^{\lambda t}\right) \right. \\
&\quad \left. + \sum_{j=N+1}^{\infty} (-i)^{n-\ell}(-1)^s n|\tau|\sin(\ell\pi/n)|\lambda_j|^{(n-\ell-1)/n} e^{\lambda_j t}[1] \right\} \\
&= \left\{ \sum_{j=1}^{N} \mathrm{Res}_{\lambda=\lambda_j}\left(\mathcal{G}_0(\lambda)e^{\lambda t}\right) + C\sum_{j=N+1}^{\infty} \lambda_j^{(n-\ell-1)/n} e^{\lambda_j t}[1] \right\} \\
&\equiv h_1(t) + h_2(t).
\end{aligned}
$$

Naturally, the derivation of the detailed results discussed above requires a fairly explicit representation of the system transfer function. However, the existence of a transfer function for this class of distributed parameter systems can also be deduced from general principles. We conclude this section with an outline of an existence proof in the context of abstract boundary control systems [21], [11], [14], illustrating the relationship between our approach and another common approach found in the literature. First note that the system (2.1)–(2.6) defines an abstract boundary control system in the sense of [21] since the operator $(-A_0)$ given in (2.8) can be shown to be an accretive operator and hence A_0 generates an analytic semigroup on $Z = L_2(0,1)$. Furthermore, we can construct a polynomial $b \in \mathcal{D}(\mathcal{A})$ that satisfies $\mathcal{W}_1(b) = 1$, $\mathcal{W}_j(b) = 0$, $j = 2,\dots,n$. In fact we can choose the degree of b to be at most $2p+1$ where $p = \max_{1\le j\le n}\{m_j\}$. In particular, we have the following result which is proved in §5.

PROPOSITION 2.6. *There exists a polynomial*

$$b(x) = \sum_{j=0}^{2p+1} a_j x^j$$

such that $b(x)$ satisfies

$$\mathcal{W}_1(b) = 1, \quad \mathcal{W}_j(b) = 0, \qquad j = 2,\dots,n,$$

where $p = \max_{1\le j\le n}\{m_j\}$.

Now define a bounded operator B from the input space U of complex numbers to Z, i.e., $B \in \mathcal{L}(U,Z)$, by multiplication by the function b. Note that

$$Bu \in \mathcal{D}(\mathcal{A}), \quad \mathcal{A}B \in \mathcal{L}(U,Z), \quad \mathcal{B}Bu = u, \quad \forall u \in U.$$

These constructions for the boundary control system (2.1)–(2.6) lead to the abstract Cauchy problem

$$\begin{aligned} \dot{v}(t) &= A_0 v(t) - B\dot{u}(t) + \mathcal{A}Bu(t), \\ v(0) &= v_0. \end{aligned} \tag{2.23}$$

Assume that $u \in C^2([0,\tau],U)$ for all $\tau > 0$. Then if $v_0 = f - Bu(0) \in \mathcal{D}(A_0)$, the classical solutions of (2.1)–(2.6) and (2.23) are related by

$$v(t) = z(t) - Bu(t).$$

Furthermore, the classical solution of (2.1)–(2.6) is unique. On the extended state-space $Z^e = U \oplus Z$, let

$$A^e = \begin{pmatrix} 0 & 0 \\ \mathcal{A}B & A_0 \end{pmatrix}, \quad B^e = \begin{pmatrix} 1 \\ -B \end{pmatrix}, \quad z^e = \begin{pmatrix} u \\ v \end{pmatrix}. \tag{2.24}$$

THEOREM 2.7. *Consider the extended system*

$$\begin{aligned} \dot{z^e}(t) &= A^e z^e(t) + B^e \tilde{u}(t), \\ z^e(0) &= \begin{pmatrix} u(0) \\ v_0 \end{pmatrix} = \begin{pmatrix} u(0) \\ f(x) - b(x)u(0) \end{pmatrix}. \end{aligned} \tag{2.25}$$

If $u \in C^2([0,\tau],U)$ and $v_0 \in \mathcal{D}(A_0)$, the system (2.25) with $\tilde{u} = \dot{u}$ has the unique classical solution

$$z^e(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix},$$

where $v(t)$ is the classical solution of (2.23). Furthermore, if $f = v_0 + Bu(0)$, then the classical solution of (2.1)–(2.6) is given by

$$z(t) = C^e z^e(t),$$

where $C^e = (B, 1)$ is bounded.

This result is well known and can be found, for example, in [14]. The importance of this result, in the present context, is that for "state linear systems" with bounded input and output operators of the form

$$\begin{aligned} \dot{w} &= Aw + Bu, \\ y &= Cw \end{aligned} \tag{2.26}$$

it is shown in [14] that the transfer function and impulse response function are well defined and in fact the transfer function has the explicit representation

$$G(s) = C(sI - A)^{-1}B, \qquad s \in \rho_\infty(A), \tag{2.27}$$

where $\rho_\infty(A)$ is the connected component of the resolvent set of A containing a semi-axis of the positive real line. Moreover, the impulse response function can be represented by

$$h(t) = CT(t)B, \tag{2.28}$$

where A is the infinitesimal generator of the C_0 semigroup $T(t)$.

More explicitly, it can be shown that if $u \in C^2([0,\infty], U)$ and $u(0) = 0$ then

$$\widehat{y}(s) = \mathcal{W}_0 \left\{ I - (sI - A_0)^{-1}(sI - \mathcal{A}) \right\} b(\cdot)\widehat{u}(s),$$

while our earlier calculations show that

$$\widehat{y}(s) = \frac{\mathcal{N}(s)}{\mathcal{D}(s)}\widehat{u}(s).$$

Thus for u in a dense set we obtain

$$\mathcal{W}_0 \left\{ I - (sI - A_0)^{-1}(sI - \mathcal{A}) \right\} b(\cdot)\widehat{u}(s) = \frac{\mathcal{N}(s)}{\mathcal{D}(s)}\widehat{u}(s). \tag{2.29}$$

It now follows from (2.29) and Proposition 2.1 that the transfer function defined in this way can be extended to all $u \in U_a$, yielding the alternative representation

$$\mathcal{W}_0 \left\{ I - (sI - A_0)^{-1}(sI - \mathcal{A}) \right\} b(\cdot) = \frac{\mathcal{N}(s)}{\mathcal{D}(s)},$$

and that, remarkably, this representation is independent of the choice of b.

As a final comment in this section we note that simple examples can be given (cf. Remark 2.1 and [27]) to show that if Assumption 2.1 does not hold then an L^2 boundary input can give rise to a solution with infinite energy at a finite time. In this case we say that the input operator is not admissible. We now present, without proof, a brief summary of the results in [9] regarding admissibility of the boundary control systems considered here. The proofs of these results, while considerably more complicated, exactly parallel those given in §5 for Propositions 2.1–2.5. First we consider a system (2.1)–(2.5) and replace the output operator (2.6) by a point evaluation sensor at a point $x_0 \in [0,1]$, i.e., let

$$y_{x_0}(t) = w(x_0, t). \tag{2.30}$$

Note that this corresponds to the boundary operator

$$\mathcal{W}_0(w)(t) = w(x_0, t)$$

so that in the notation of Assumption 2.1 the order of the output is $m_0 = 0$ and $\ell = m_1 - m_0 = m_1 > 0$. For a fixed x_0 it can be shown that the resulting system has a transfer function of the form

$$\mathcal{G}_{x_0}(\lambda) = \frac{\mathcal{N}_{x_0}(\lambda)}{\mathcal{D}(\lambda)},$$

where the denominator is exactly the same as in (2.19). The asymptotic form of the transfer function is very similar to that for a general open-loop transfer function given in (4.22). The main difference is that the terms $e^{\pm\omega_\mu z}$ are replaced by terms $e^{\pm\omega_\mu z(x_0-1)}$ and the asymptotic constants in the numerator are different. Thus for every x_0 the system has the same open-loop poles as described in Propositions 2.1 and 2.5. The main result of [9] concerning admissibility of the boundary control systems in (2.1)–(2.5) with Assumption 2.1 is that L^2 inputs produce finite energy outputs in $L^2(0,1)$.

PROPOSITION 2.8. *For the system* (2.1)–(2.5) *with Assumption* 2.1 *and for* $x_0 \in [0,1]$, *we have the following results.*

1. *There exist constants a and C such that for all* $\operatorname{Re}\lambda > a$

$$|\mathcal{G}_{x_0}(\lambda)| \le C|\lambda|^{-\ell/n} \exp\left(-x_0|\lambda|^{1/n} \sin\left(\frac{\pi}{2n}\right)\right).$$

In fact, there are constants N_{01}, N_{02}, D_1, D_{-1} so that for large modulus of $\lambda = i^n z^n$ the transfer function has the asymptotic representation (cf. (4.22))

$$\mathcal{G}_{x_0}(\lambda) = \frac{\mathcal{N}_{x_0}(\lambda)}{\mathcal{D}(\lambda)} = \frac{\left(-[N_{01}]e^{\omega_\mu z(x_0-1)} + [N_{02}]e^{-\omega_\mu z(x_0-1)}\right)}{z^\ell\left(-[D_1]e^{\omega_\mu z} + [D_{-1}]e^{-\omega_\mu z}\right)}. \tag{2.31}$$

Here ω_μ is a particular nth root of minus one (cf. Remark 4.1, *part* 6).

2. *Suppose λ_j is a pole of large modulus of $\mathcal{G}_{x_0}(\lambda)$, $\lambda_j = i^n z_j^n$ where $n = 2\mu$ is the order of $\mathcal{A}$, $\ell > 0$ is the order of the input operator. Then*

$$\operatorname{Res}_{\lambda=\lambda_j} \mathcal{G}_{x_0}(\lambda) = (-1)^{t_2-t_1} n|z_j|^{n-\ell-1} \left|\frac{N_{02}}{D_{-1}}\right| \sin\left(\frac{\pi((1-x_0)m - M)}{n}\right)[1],$$

where t_1, t_2 are determined from

$$\arg D_{-1} = t_1\pi - m\pi/n, \qquad \arg N_{02} = t_2\pi - M\pi/n$$

with

$$m = \sum_{j=1}^{n} m_j, \qquad M = \sum_{j=\mu+1}^{n} m_j.$$

3. *The inverse Laplace transform of $\mathcal{G}_{x_0}(\lambda)$ can be computed by the method of residues and as in Proposition* 2.5 *there is a constant C for which*

$$\begin{aligned} h(x_0,t) &\equiv \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} e^{st}\,\mathcal{G}_{x_0}(s)\,ds \\ &= \left\{\sum_{j=1}^{N} \operatorname{Res}_{\lambda=\lambda_j}\left(\mathcal{G}_{x_0}(\lambda)e^{\lambda t}\right)\right. \\ &\qquad \left. + C\sin\left(\frac{\pi((1-x_0)m-M)}{n}\right)\sum_{j=N+1}^{\infty} |\lambda_j|^{(n-\ell-1)/n} e^{\lambda_j t}[1]\right\}. \end{aligned}$$

4. *$w(x,t) = y_x(t) = (h(x,\cdot) * u(\cdot))(t) \in L^2(0,1)$ for every $t > 0$ and $u \in U_a$.*

3. Statement and illustration of the main results. Deferring the detailed proofs to §4, in this section we describe a fairly complete analogue of finite-dimensional root-locus theory for systems (2.1)–(2.6) satisfying Assumption 2.1. In particular, closed-loop poles tend to open-loop zeros, all but a finite number of which are simple and real. The familiar real axis loci results are established from which we deduce some stability results. With some effort, this analysis can be extended to more general inputs and outputs. However, even in the case considered here, where the highest-order derivatives in the boundary conditions are separated—i.e., the actuators and sensors are colocated in the highest-order derivatives—there are some important differences from the finite-dimensional case. These differences mainly stem from the fact that for systems (2.1)–(2.6) the transfer function is not rational but always has an essential

singularity at infinity, a phenomenon which we encountered in §2 in formulating the definition of the instantaneous gain.

The following simple examples are given to demonstrate some of the subtle differences that can occur in the infinite-dimensional case and to motivate assumptions that were made to eliminate some of these difficulties. Example 3.1 shows that for completely general boundary inputs and outputs not of the type given in (2.4) the situation is very different from the finite-dimensional case and underscores the importance of certain of our hypotheses.

Example 3.1. Consider the controlled heat equation

$$\dot{w} = Aw, \quad A = \frac{d^2}{dx^2}, \qquad x \in (0,1),$$

$$w(0,t) = u(t),$$

$$\frac{\partial w}{\partial x}(0,t) - \beta \frac{\partial w}{\partial x}(1,t) = 0, \qquad \beta \in \mathbb{R},$$

$$y(t) = w(1,t).$$

Using the change of variables $\lambda = -z^2$ the characteristic equation for the closed-loop operator $A_k = A$ with boundary conditions

$$f(0) + kf(1) = 0, \qquad f'(0) - \beta f'(1) = 0$$

can be written as

$$(k - \beta)\cos(z) = (k\beta - 1)$$

and a straightforward analysis provides the following possibilities:

1. For $k \neq \beta$, $\beta = \pm 1$, we have that A_k is a discrete spectral operator with all double eigenvalues.
2. If $k = \beta = \pm 1$, the point spectrum consists of the entire complex plane.
3. For $k = \beta \neq \pm 1$ the spectrum is empty.
4. When $k \neq \beta \neq \pm 1$ A_k is a discrete spectral operator with all simple eigenvalues.

The next two examples show that even for simple examples with boundary conditions of the type (2.4) the root loci exhibits somewhat different behavior than the finite-dimensional case. Namely, we should expect very different asymptotic limits, not just angles of approach, as the gain goes either to plus or to minus infinity.

Example 3.2. Consider the controlled heat equation

$$\begin{aligned}
\dot{w} = Aw, \quad A &= \frac{d^2}{dx^2}, \qquad x \in (0,1), \\
u(t) = B(w)(t) &= \mathcal{W}_1(w)(t) = -\frac{\partial w}{\partial x}(0,t), \\
y(t) = C(w)(t) &= \mathcal{W}_0(w)(t) = w(0,t), \\
\mathcal{W}_2(w)(t) &= \frac{\partial w}{\partial x}(1,t) = 0.
\end{aligned}$$

For this problem, of course, a good strategy to stabilize the temperature at zero would be to heat the rod if it is cold and to cool the rod if it is hot. This suggests employing the simple scalar boundary feedback control law

$$u(t) = -ky(t).$$

The open-loop transfer function is easily computed and given by

$$\mathcal{G}_0(\lambda) = \frac{\cosh(\sqrt{\lambda})}{\sqrt{\lambda}\sinh(\sqrt{\lambda})}$$

from which a straightforward calculation provides the open-loop poles $\lambda_j(0) = -j^2\pi^2$, $j = 0, 1, \cdots$ and open-loop zeros $\lambda_j(\infty) = -(j+1/2)^2\pi^2$, $j = 0, 1, \cdots$ which interlace on the negative real axis.

For the feedback law

$$u(t) = -ky(t) + v(t),$$

the closed-loop transfer function is

$$\mathcal{G}_k(\lambda) = \frac{\cosh(\sqrt{\lambda})}{\sqrt{\lambda}\sinh(\sqrt{\lambda}) + k\cosh(\sqrt{\lambda})}$$

and the return difference equation can be written as

$$1 + k\frac{\cosh(\sqrt{\lambda})}{\sqrt{\lambda}\sinh(\sqrt{\lambda})} = 0.$$

To more easily describe geometrically the behavior of the closed-loop poles, we introduce the change of variables $\lambda = -z^2$ so that the return difference equation can be written as

$$-\frac{1}{k} = g(z) \equiv \frac{\cos(z)}{-z\sin(z)},$$

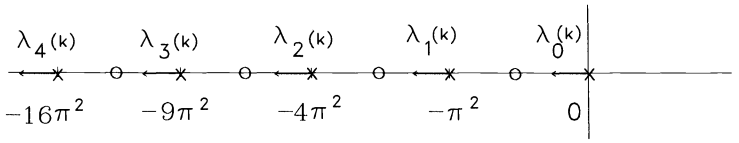
where

$$g(\bar{z}) = \overline{g(z)}.$$

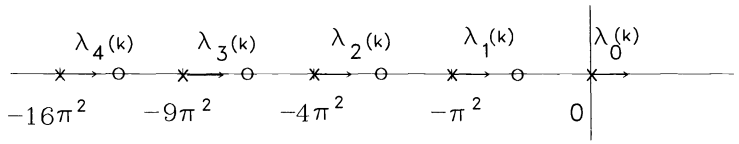
Since the operator A_k in this case is selfadjoint the closed-loop poles (i.e., points of the spectrum of A_k) are real and are given by $\{\lambda_j(k)\}_{j=0}^{\infty}$, which for $k > 0$, satisfy

$$-j^2\pi^2 = \lambda_j(0) > \lambda_j(k) \to \lambda_j(\infty) = -(j + 1/2)^2\pi^2, \qquad j = 0, 1 \ \ldots.$$

We can readily compute the instantaneous gain in this case. We have $\alpha_1 = -1$, $\alpha_0 = 1$, $m_1 = 1$, $m_0 = 0$ so that $\ell = m_1 - m_0 = 1 > 0$ and in Theorem 2.4 $\ell_0 = 0$ so that $s = 1 + 1 = 2$ and $\mathcal{K} = (-1)^s = 1$ and we would take $k > 0$. As k goes from zero to plus infinity the closed-loop poles move to left from the open-loop poles to the open-loop zeroes. All the branches of the root locus are bounded as k goes to plus infinity. For $k < 0$ we have the same functions $\{\lambda_j(k)\}_{j=0}^{\infty}$ defined as distinct branches of a single analytic function. As expected from finite-dimensional root locus the eigenvalues in this case move to the right. But in this case the eigenvalue $\lambda_0(k)$



(a) "x" *open-loop pole*, "o" *open-loop zero, gain* $k > 0$.



(b) "x" *open-loop pole*, "o" *open-loop zero, gain* $k < 0$.

FIG. 1

goes from zero to plus infinity. Thus in this case there is one unbounded branch of the root locus and again we find that unlike the finite-dimensional case the asymptotic behavior of the root loci is different as k goes to plus or minus infinity. (See Fig. 1.)

Example 3.3. In this example, originally due to Rellich [34], we illustrate the importance of Assumption 2.1 in constructing a system transfer function and in analyzing the behavior of the root loci. In particular, the candidate transfer function is unbounded in any right half-plane, the instantaneous gain formula no longer applies, and the conclusions we would draw about closed-loop stability are invalid for negative values of the gain.

Consider the controlled heat equation

$$\begin{aligned} \dot{w} &= Aw, \quad A = \frac{d^2}{dx^2}, \qquad x \in (0,1), \\ u(t) &= B(w)(t) = \mathcal{W}_1(w)(t) = w(0,t), \\ y(t) &= C(w)(t) = \mathcal{W}_0(w)(t) = -\frac{\partial w}{\partial x}(0,t), \\ \mathcal{W}_2(w)(t) &= w(1,t) = 0. \end{aligned}$$

Again we consider the scalar boundary feedback

$$u(t) = -ky(t) + v(t),$$

in which case the closed-loop transfer function is given by

$$\mathcal{G}_k(\lambda) = \frac{\sqrt{\lambda}\cosh(\sqrt{\lambda})}{\sinh(\sqrt{\lambda}) + k\sqrt{\lambda}\cosh(\sqrt{\lambda})}$$

and the open-loop poles are $\lambda_j(0) = -j^2\pi^2$, the open-loop zeros are $\lambda_j(\infty) = -(j - 1/2)^2\pi^2$, $j = 1, 2, \cdots$ and they interlace on the negative real axis.

On introducing the change of variables $\lambda = -z^2$, the return difference equation can be written as

$$-\frac{1}{k} = g(s) \equiv \frac{z\cos(z)}{\sin(z)},$$

where

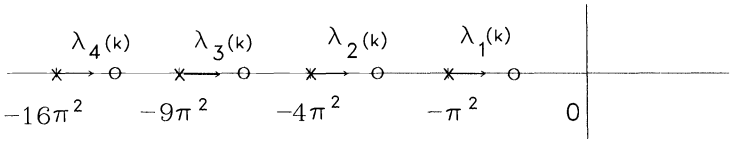
$$g(\bar{z}) = \overline{g(z)}.$$

Just as in the last example, the operator A_k is selfadjoint and it is easy to see that the closed-loop poles (i.e., the spectrum of A_k) given by $\{\lambda_j(k)\}_{j=1}^{\infty}$ is real and for $k \geq 0$, the poles move to the right and as k goes from zero to plus infinity the closed-loop poles move from the open-loop poles to the open-loop zeroes. For $k < 0$ we have the same branches $\{\lambda_j(k)\}_{j=1}^{\infty}$ that move to the left but in addition there is an unstable eigenvalue $\lambda_0(k)$ that begins at plus infinity, moves to zero as $k \to -1$ and as $k \to -\infty$ moves to the first open-loop zero. This eigenvalue satisfies

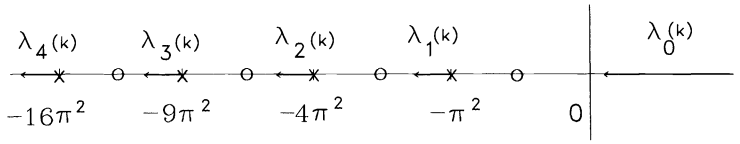
$$\tanh(\sqrt{\lambda_0}) = -k\sqrt{\lambda_0}.$$

Once again in this example, unlike the finite-dimensional case, we see a very different behavior for k positive and negative. (See Fig. 2.)

Note that for this example the "transfer function" is not bounded in the right half-plane. This is to be expected whenever the order of the input is less than the order of the output; this point will become clear once we establish the asymptotic form of the return difference equation.



(a) "x" *open-loop pole*, "o" *open-loop zero, gain* $k > 0$.



(b) "x" *open-loop pole*, "o" *open-loop zero, gain* $k < 0$.

FIG. 2

These examples indicate that some care must be taken in developing a root locus theory for distributed systems. For general boundary conditions the functions $\mathcal{N}$ and $\mathcal{D}$ are entire functions of λ and hence for each k the return difference equation either has a discrete set of zeros with no finite accumulation point, is identically zero (spectrum the entire complex plane) or does not vanish at all (spectrum empty). For the boundary conditions in (2.4) it can be shown that the first alternative always holds; i.e., the operator A_k is always a discrete spectral operator. This enables us to establish a root-locus theory much like the finite-dimensional case.

THEOREM 3.1. *The operators A_k of* (2.14) *are discrete* (*Riesz*) *spectral operators that generate analytic semigroups $T_k(t)$. Furthermore they form a holomorphic family in k* (*in the sense of norm resolvent convergence*), *which satisfies the "separation of the spectrum" condition uniformly in k and the "spectrum determined growth condition" for real k such that $k \cdot \mathcal{K} > 0$. In particular, the spectrum of A_k can be written as $\{\lambda_j(k)\}_{j=1}^{\infty}$ for which the eigenfunctions and associated functions form a Riesz basis in $L^2(0,1)$ for all $k \cdot \mathcal{K} > 0$ and all but finitely many of the eigenvalues are simple so that the associated projections have rank one.*

As stated in [25], [34] it is not so easy to conclude that an infinite set of eigenvalues (or closed-loop poles) vary continuously in k as is illustrated in Example 3.3. Nonetheless, under the additional hypothesis in Assumption 2.1 such an infinite-dimensional root locus result is valid.

THEOREM 3.2. *For the systems* (2.1)–(2.6), *which satisfy Assumption* 2.1, *all but a finite number of the open-loop poles $\{\lambda_j(0)\}_{j=1}^{\infty}$ and zeros $\{\lambda_j(\infty)\}_{j=1}^{\infty}$ are real, interlace on the negative real axis, and tend to minus infinity as $j \to \infty$. Choosing $k \cdot \mathcal{K} > 0$, the closed-loop poles $\{\lambda_j(k)\}_{j=1}^{\infty}$ vary continuously from the open-loop poles to the open-loop zeros. More specifically, the closed-loop poles corresponding to the infinitely many real open-loop poles and zeros $(\{\lambda_j(0)\}_{j=N+1}^{\infty}, \{\lambda_j(\infty)\}_{j=N+1}^{\infty})$ are real, simple, and move to the left from an open-loop pole to an open-loop zero. The remaining finitely many closed-loop poles lie inside a fixed simple closed curve that also contains an equal number of open-loop poles and zeros and these closed-loop poles vary continuously in k. Furthermore, all branches of the root locus are bounded. In general there are at most a fixed finite number of common open-loop poles and zeros that correspond to stationary eigenvalues in the spectrum of the operators A_k for $k \cdot \mathcal{K} \in [0, \infty]$.*

COROLLARY 3.3. *For the systems* (2.1)–(2.6) *satisfying Assumption* 2.1 *and an initial condition $f \in L^2(0,1)$, if $k \cdot \mathcal{K} > 0$ and $k_0 \cdot \mathcal{K} > 0$, then*

$$\|T_k(t)f - T_{k_0}(t)f\| \to 0, \quad \text{as} \quad k \to k_0$$

uniformly for t in compact subintervals of $(0, \infty)$.

From Theorem 3.2 and its proof follow a number of statements about the root loci that are similar to well-known rules in classical automatic control. As an example of our extension of finite-dimensional root locus theory, we show there is a version of the "real axis loci" test.

THEOREM 3.4. *For systems* (2.1)–(2.6) *satisfying Assumption* 2.1, *if $k \cdot \mathcal{K} > 0$, then a real point on the root locus always lies to the left of an odd number of poles and zeros.*

Remark 3.1. For both the case at hand and more general cases, it is interesting to analyze the finitely many "exceptional" branches of the root locus, which may be complex. To this end, we write

$$\mathcal{G}_0(\lambda) = \frac{n_1(\lambda)n_2(\lambda)}{d_1(\lambda)d_2(\lambda)},$$

where n_1, d_1 are polynomials of degree d, having all roots inside a fixed curve Γ and n_2, d_2 are entire functions having no zeros inside Γ. Then, from Rouche's theorem, Proposition 2.1, and Theorem 3.2 we may conclude that the branches of the root locus inside Γ coincide, for $k = 0$, with the roots of d_1, and converge, as $k \cdot \mathcal{K} \to \infty$, to the roots of n_1. Moreover, multiple arriving branches of the root-loci at a root λ_0 of n_1 form a Butterworth pattern having order determined by the rational transfer function

$n_1(\lambda)/d_1(\lambda)$. However, the angles of arrival may not coincide with those computed from $n_1(\lambda)/d_1(\lambda)$, nor will breakaway points for $\mathcal{G}_0(\lambda)$ coincide with those computed for $n_1(\lambda)/d_1(\lambda)$.

Choosing $k_0 = \infty$ in Corollary 3.3 we might expect that, if the zero dynamics is exponentially stable, then for $k \cdot \mathcal{K}$ sufficiently large the trajectories of the closed-loop system would tend to zero. Alternatively, exponential stability of the closed-loop system follows from Theorem 3.2 and the results in [13], [20] on discrete spectral operators, since the operators A_k generate analytic semigroups and are Riesz spectral operators. In particular, since the zero dynamics is exponentially stable we have an estimate for the growth constant

$$w_\infty = \sup_j \{\mathrm{Re}(\lambda_j(\infty))\} < 0.$$

Finally, using Theorem 3.2 and choosing $k_0 \cdot \mathcal{K} > 0$ sufficiently large we can find a positive σ so that for $k \cdot \mathcal{K} > k_0 \cdot \mathcal{K}$ we have

$$\|T_k(t)\| \leq Ce^{-\sigma t}.$$

COROLLARY 3.5. *For systems* (2.1)–(2.4) *satisfying Assumption* 2.1, *if the zero dynamics* (2.10) *is exponentially stable and the sign of the gain is chosen so that* $k \cdot \mathcal{K} > 0$, *then there exists a* k_0 *such that for* $k \cdot \mathcal{K} > k_0 \cdot \mathcal{K}$ *the closed-loop system is exponentially stable.*

Example 3.4. Consider the controlled heat equation

$$\begin{aligned} \dot{w} &= Aw, \quad A = \frac{d^2}{dx^2}, \qquad x \in (0,1), \\ u(t) &= B(w)(t) = \mathcal{W}_1(w)(t) = -\frac{\partial w}{\partial x}(0,t) - w(1,t), \\ y(t) &= C(w)(t) = w(0,t), \\ \mathcal{W}_2(w)(t) &= \frac{\partial w}{\partial x}(1,t) = 0. \end{aligned}$$

Again we consider the scalar boundary feedback

$$u(t) = -ky(t) + v(t).$$

On introducing the change of variables $\lambda = -z^2$, the return difference equation can be written as

$$\frac{1}{k} = g(s) \equiv \frac{\cos(z)}{1 + z\sin(z)}.$$

In this case the instantaneous gain is easily computed from $m_0 = 0$, $m_1 = 1$ and $\ell_0 = 0$, $\ell = 1$, which implies $s = 1 + \arg(-1)/\pi = 2$, so that

$$\mathcal{K} = (-1)^s = 1$$

so we take $k > 0$. It is easy to see that there are real open-loop poles $\{\lambda_j(0)\}_{j=1}^{\infty}$ that are asymptotic to the zeros of $\sin(z)$. In fact, the open-loop poles satisfy

$$-(2j\pi)^2 < \lambda_{2j}(0) < -((2j - 1/2)\pi)^2 < \lambda_{2j-1}(0) < -((2j-1)\pi)^2, \qquad j = 1, 2, \ldots,$$

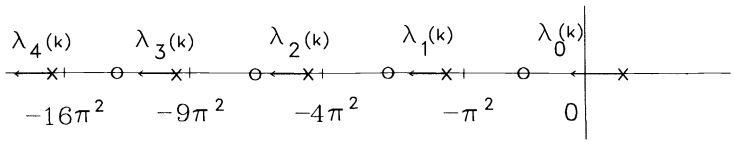


FIG. 3. "x" *open-loop pole*, "o" *open-loop zero, gain* $k > 0$.

where the open-loop zeros are

$$\lambda_j(\infty) = -((j - 1/2)\pi)^2$$

so the open-loop poles and zeros interlace.

In the region $S_0 = \{z : 0 \leq \arg(z) \leq \pi/2\}$ there is another root of the return difference equation having the form $\lambda_0(0) = r^2$, where r is the positive root of the equation

$$\tanh(r) = 1/r.$$

The corresponding branch of the root locus then moves to the first open-loop zero $\lambda_1(\infty) = -(\pi/2)^2$ as k goes from 0 to ∞. (See Fig. 3.)

We conclude this section with a brief outline of the proofs of Theorems 3.1 and 3.2, since the proofs are somewhat long and draw on a combination of methods from asymptotic analysis, complex analysis, and functional analysis. In effect, this combination of methods allows us to establish a spectral theory for certain non-selfadjoint boundary value problems.

More explicitly, the proofs of Theorems 3.1 and 3.2 are based on an analysis of the zeros $\lambda = \lambda(k)$ of the return difference equation

$$\mathcal{D}(\lambda) + k\mathcal{N}(\lambda) = 0,$$

which is also exactly the characteristic equation providing the spectra of the closed-loop operators A_k. In the proof of Proposition 2.1 given in §5, it is shown that the numerator and denominator of the open-loop transfer function are given explicitly by

$$\mathcal{N}(\lambda) = \det\left(\{\mathcal{W}_i(f_j(\cdot,\lambda))\}_{i=0,2,j=1}^{n,n}\right), \tag{3.1}$$

$$\mathcal{D}(\lambda) = \det\left(\{\mathcal{W}_i(f_j(\cdot,\lambda))\}_{i=1,j=1}^{n,n}\right),$$

where $\{f_j\}$ denotes a basis of solutions of the ordinary differential equation

$$Af - \lambda f = 0, \tag{3.2}$$

which are analytic functions of λ. The existence of such a basis is well known and is discussed in Proposition 5.1. Unfortunately it is extremely difficult in general to obtain this basis explicitly. So to obtain qualitative information concerning the zeros of the closed-loop transfer function, we employ the asymptotic techniques developed by Birkhoff in [1], [2] to obtain asymptotic formulas for a special basis of solutions

(given in Proposition 5.1 below) that are only analytic in the cut plane and have simple asymptotic representations for modulus of λ large. With this the analysis of the return difference equation is divided into two main parts. First an asymptotic development is employed to analyze the poles and zeros of large modulus. One important corollary of the analysis is that the operators A_k satisfy the separation of the spectrum condition of Kato [25] uniformly in $k \cdot \mathcal{K} > 0$. This then allows us to carry out the second part of the proof, which is to analyze the variation of the remaining finitely many closed-loop poles, not considered in the asymptotic development, based on classical perturbation techniques for finite systems of eigenvalues (cf. [25]). In particular we show that the operators A_k are holomorphic in k in the generalized sense of norm resolvent convergence.

To carry out the asymptotic analysis we employ techniques introduced by Birkhoff [1] for ordinary differential operators on a finite interval. This work predated the spectral theory for unbounded selfadjoint operators in Hilbert space by many years and, while his efforts were restricted to ordinary differential operators, the analysis is applicable to a wide variety of non-selfadjoint problems, providing in particular the basis of the present work. It is worth commenting that Birkhoff's work was followed by a lively literature related to his work as well as to generalizations. In 1912, Tamarkin [37] presented a paper questioning the validity of Birkhoff's work in the case of even order operators, evoking a paper [2] published by Birkhoff in response to these criticisms. In 1926, Stone [36] related the expansions of Birkhoff and Fourier. Birkhoff and Langer [3] later extended these results to systems of ordinary differential equations and Wilder [43] extended the analysis to problems with boundary conditions at points other than the end points (in the context of control theory this would correspond to interior point control problems). For a more complete reference to the many subsequent extensions and refinements we refer to the references contained in [20], [30].

For the asymptotic analysis, we first introduce a change of variables in the return difference equation by letting $\lambda = i^n z^n$ and then consider z in the region

$$S = \{z \in \mathbb{C} \mid -\pi/n \leq \arg(z) < /n\}.$$

In the region S of the z plane, the return difference equation can be written in the form (cf. (4.13))

$$\Delta(z,k) = h(z)\left(z^{\ell}\delta(z,0) + k\delta(z,\infty)\right), \tag{3.3}$$

where $h(z)$ described in (4.14) is not zero and $\delta(z,0)$ and $\delta(z,\infty)$ are given explicitly in (4.17) and (4.18), respectively. As is shown below, the cases in which $\mu = n/2$ is even or odd are somewhat different. Nevertheless, on introducing appropriate rotations, it is possible to treat both cases together. In particular, after considerable computation and simplification it is shown that the closed-loop poles of large modulus are the zeros of a function in the asymptotic form

$$F(z,k) = e^{-2iz} - [v]\left(\frac{z^{\ell} + k[\overline{\tau}]}{z^{\ell} + k[\tau]}\right).$$

Here the asymptotic notation $[a]$ introduced by Birkhoff [1] is used

$$[a] = a + O(1/z).$$

Moreover, $\tau \in \mathbb{C}$ satisfies

$$\arg(\tau) = \begin{cases} s\pi + \pi\ell/n, & \mu \text{ odd}, \\ s\pi + 2\pi\ell/n, & \mu \text{ even}, \end{cases}$$

where $s = \arg(\alpha_0/\alpha_1)/\pi + (\ell_0 + \ell)$ is defined in Theorem 2.4 and v satisfies

$$v = e^{2\pi m i/n}, \qquad m = \sum_{j=1}^{n} m_j.$$

The analysis proceeds by comparing the zeros of $F(z,k)$ with the zeros of the function

$$f(z,k) = e^{-2iz} - v\left(\frac{z^\ell + k\overline{\tau}}{z^\ell + k\tau}\right)$$

using Rouche's theorem.

It is first shown that for $z \in S$ and $k \cdot \mathcal{K} > 0$, there exists $M, y_0 > 0$ such that $F(z,k) \neq 0$ for $z \in S_{y_0,M}$

$$S_{y_0,M} = S \cap \{z \in \mathbb{C} \mid |\operatorname{Im}(z)| > y_0,\ |z| > M\}.$$

Next we show that the zeros of $f(z,k)$ are all real and simple and that the nonzero zeros of $f(z,0)$ and $f(z,\infty)$ interlace on the positive real axis. Indeed for $k = 0,\ \infty$ we obtain simple explicit formulas for the zeros of f.

Exploiting the periodicity of the factor $\exp(-2iz)$ in both F and f, we next decompose the complementary region $S \backslash S_{y_0,M}$ into rectangular regions $V_p = V + p\pi$ for $p \in \mathbb{Z}$ and

$$V = \{\tilde{z} \mid |\operatorname{Im}(\tilde{z})| < y_0,\ a_1 < \operatorname{Re}(\tilde{z}) < a_2,\ a_1 = \frac{\pi - \arg(v) + \pi\ell/n}{2},\ a_2 = a_1 + \pi\}.$$

This decomposition allows us to reduce the arguments for calculations on the regions V_p to the single region V. In particular, for a fixed p, we introduce the new gain parameter

$$g = \frac{k}{(p\pi)^\ell}$$

and the functions

$$F_p(\tilde{z},g) = e^{-2i\tilde{z}} - [v]\left(\frac{(1+\tilde{z}/(p\pi))^\ell + g[\overline{\tau}]}{(1+\tilde{z}/(p\pi))^\ell + g[\tau]}\right),$$

$$f_p(\tilde{z},g) = e^{-2i\tilde{z}} - v\left(\frac{(1+\tilde{z}/(p\pi))^\ell + g\overline{\tau}}{(1+\tilde{z}/(p\pi))^\ell + g\tau}\right),$$

and

$$h(\tilde{z},g) = e^{-2i\tilde{z}} - v\left(\frac{1+g\overline{\tau}}{1+g\tau}\right).$$

Using these functions defined on the single rectangle V, we show that for sufficiently large p, f_p has only one zero in $\overline{V}$ which, since the roots occur in conjugate pairs, must be real. To this end, we show that

$$|f_p(\tilde{z}, g) - h(\tilde{z}, g)|$$

goes to zero uniformly in g for $\tilde{z} \in \overline{V}$ as $p \to \infty$. Then we show that for $\tilde{z} \in \partial V$ (the boundary of V) and all g there is a $C > 0$ so that

$$|h(\tilde{z}, g)| > C,$$

which implies that there exists a P_1 such that for $p > P_1$

$$|f_p(\tilde{z}, g)| > C/2$$

for $\tilde{z} \in \partial V$ and all g.

Next we show, in a similar way, that

$$|F_p(\tilde{z}, g) - f_p(\tilde{z}, g)| = O(1/p)$$

uniformly for $\tilde{z} \in \overline{V}$ and $g \cdot \mathcal{K} > 0$. Hence there exists $P_2 > P_1$ such that for $p > P_2$

$$|F_p(\tilde{z}, g)| \geq C/3$$

for $\tilde{z} \in \partial V$ and $g \cdot \mathcal{K} > 0$. So for $p > P_2$, $\tilde{z} \in \partial V$ and $g \cdot \mathcal{K} > 0$,

$$|F_p(\tilde{z}, g) - f_p(\tilde{z}, g)| < |f_p(\tilde{z}, g)|$$

and we can apply Rouche's theorem to conclude that $F_p(\tilde{z}, g)$ has only one zero in V which again must be real since roots occur in conjugate pairs.

The next step is to show that for large $|z|$, the zeros of $F(z, k)$ are continuous monotone increasing functions of k for $|k| \to \infty$, $k \cdot \mathcal{K} > 0$. Once this is established we will know that the infinitely many real closed-loop poles of large modulus vary continuously from real open-loop zeros to real open-loop poles that interlace on the negative real axis and that the operators A_k satisfy the separation of the spectrum condition uniformly in k.

The remainder of the proof of Theorem 3.2 consists in establishing that the resolvent operator $R(\lambda, k) = (A_k - \lambda I)^{-1}$ is a holomorphic family in the generalized sense [25] for $k \cdot \mathcal{K} > 0$ and hence every finite system of eigenvalues vary continuously.

4. Asymptotic analysis of the return difference equation. In this section, we provide detailed information on the asymptotic behavior of the closed-loop systems described in §2. The essential ingredient for carrying out this analysis is an explicit asymptotic representation for a basis of solutions of the ordinary differential equation (3.2) that are analytic in λ. This analysis reposes heavily on work found in [1], [2], [30]. In particular, it can be shown (as in §5) that the spectra of A_0, A_∞, and A_k (open-loop zeros, open-loop poles, and closed-loop poles) are given, respectively (cf. the proof of Proposition 2.1 in §5 and (3.1)), by the zeros of the determinants

$$\mathcal{N}(\lambda) = \det\left(\{\mathcal{W}_i(f_j(\cdot, \lambda))\}_{i=0,2,j=1}^{n,n}\right), \tag{4.1}$$

$$\mathcal{D}(\lambda) = \det\left(\{\mathcal{W}_i(f_j(\cdot, \lambda))\}_{i=1,j=1}^{n,n}\right),$$

$$\mathcal{D}(\lambda) + k\mathcal{N}(\lambda),$$

where $\{f_j\}$ denotes a basis of solutions of (3.2).

As suggested by the examples in §3, it is convenient to introduce the change of variables in the complex parameter λ by

$$z^n = i^n \lambda, \qquad i = \sqrt{-1} \tag{4.2}$$

and consider the eigenvalue problem

$$f^{(n)} + i^{n-2} L_1 f + z^n f = 0 \tag{4.3}$$

$$\mathcal{W}_1(f) + k\,\mathcal{W}_0(f) = 0, \quad \mathcal{W}_i(f) = 0, \qquad i = 2, \ldots, n$$

in a suitable region of the z-plane. Namely for $n = 2\mu$ and for $0 \leq j \leq (2n-1)$ let

$$S_j = \{z \mid j\pi/n \leq \arg(z) < (j+1)\pi/n\} \tag{4.4}$$

and denote by ψ_j, $j = \pm 1, \pm 3, \cdots, \pm(n-1)$ the nth roots of (-1) given by

$$\psi_j = \exp(\pi i + j\pi i/n). \tag{4.5}$$

(See Fig. 4.)

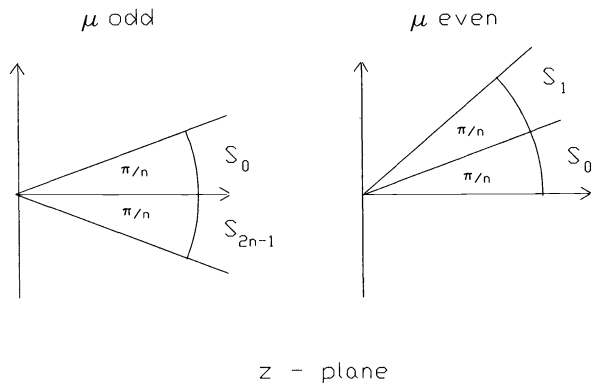


FIG. 4

Remark 4.1.

1. The entire complex λ-plane under the map (4.2) is covered by the image of two adjacent regions S_j from (4.4). In the following analysis it is important to know that the eigenvalues obtained are independent of the pair of regions chosen. This is proven in, for example, [1], [30].

2. Following [1], [30], for each region S_j we prescribe a particular ordering of the roots ψ_k, denoted by ω_k, $k = 1, 2, \ldots, n$. The ordering is chosen so that for all z in S_j, we have

$$\begin{aligned} &\operatorname{Re}(z\omega_1) \leq \operatorname{Re}(z\omega_2) \leq \cdots \leq \operatorname{Re}(z\omega_{\mu-1}) < 0, \\ &\operatorname{Re}(z\omega_\mu) \leq 0, \qquad \operatorname{Re}(z\omega_{\mu+1}) \geq 0, \\ &0 < \operatorname{Re}(z\omega_{\mu+2}) \leq \operatorname{Re}(z\omega_{\mu+3}) \leq \cdots \leq \operatorname{Re}(z\omega_n). \end{aligned} \tag{4.6}$$

In particular, for S_0, let

$$\omega_{2j-1} = \exp\left(\left(1 - \frac{2j-1}{n}\right)\pi i\right), \quad \omega_{2j} = \exp\left(\left(1 + \frac{2j-1}{n}\right)\pi i\right), \qquad j = 1, \ldots, \mu.$$

Then for S_{2n-1} let $\{\omega_j'\}$ denote the appropriate ordering. We have

$$\omega_j' = \overline{\omega_j}, \qquad j = 1, \ldots, n$$

and defining $\{\omega_j''\}$ as the ordering for S_1 we have

$$\omega_j'' = \omega_j' e^{-(2\pi i/n)} = \overline{\omega_j} e^{-(2\pi i/n)}, \qquad j = 1, \ldots, \mu.$$

3. A straightforward calculation based on the definition of ω_j shows that (see Fig. 5)

$$e^{-(2\pi i/n)}\omega_j = \begin{cases} \omega_1, & j = 2, \\ \omega_{j-2}, & j \neq 2 \text{ even}, \\ \omega_{j+2}, & j \neq n-1 \text{ odd}, \\ \omega_n, & j = n-1, \end{cases}$$

$$e^{2\pi i/n}\omega_j = \begin{cases} \omega_2, & j = 1, \\ \omega_{j-2}, & j \neq 1 \text{ odd}, \\ \omega_{j+2}, & j \neq n \text{ even}, \\ \omega_{n-1}, & j = n. \end{cases}$$

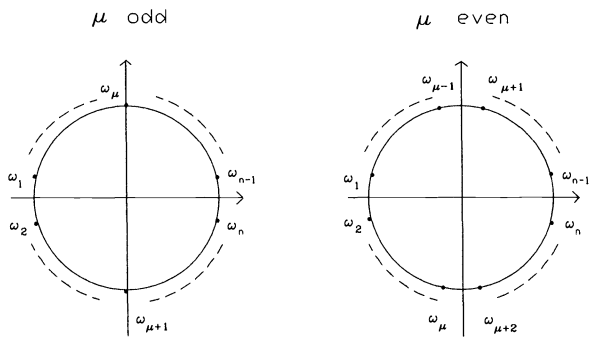


FIG. 5. *Roots of minus one for region* S_0.

4. For $c \in \mathbb{C}$ and $S = S_j$, for some j, we define $T_c = c + S$. With the ordering $\{\omega_j\}$ of the roots of minus one given above for a region S we have the estimates for $z \in T_c$

$$|e^{\omega_j z}| \leq \exp(\mathrm{Re}(z\omega_{\mu-1})) \to 0, \quad |z| \to \infty, \qquad j = 1, \ldots, \mu - 1,$$

$$|e^{\omega_j z}| \geq \exp(\mathrm{Re}(z\omega_{\mu+2})) \to \infty, \quad |z| \to \infty, \qquad j = \mu + 2, \ldots, n.$$

5. Recall the notation of Birkhoff [a] for $a \in \mathbb{C}$ to indicate an asymptotic expression of the form

$$[a] = a + O(1/z),$$

where by $O(1/z)$, as usual, we mean there exists a constant C so that

$$|O(1/z)| \leq C/|z|, \qquad |z| >> 1.$$

6. For $n = 2\mu$ we have
(a) For μ odd, $\omega_\mu = i$ for S_0.
(b) For μ even, $\omega_\mu = i\exp(-i\pi/n)$ for S_1.
For $z \in S_r$, $r = 0, \ldots, (2n-1)$, the functions

$$e^{z\omega_j x}, \qquad 1 \leq j \leq n, \tag{4.7}$$

form a basis of solutions for the equation

$$f^{(n)}(x) + z^n f(x) = 0. \tag{4.8}$$

It was shown by Birkhoff [1] that the asymptotic behavior of the eigenvectors and eigenvalues for the general problem are completely determined by the solutions (4.7) for (4.8). In particular, the following result can be found in [1], [30] regarding the asymptotic representation for a basis of solutions for (4.3).

THEOREM 4.1. *If the functions $p_2, \ldots, p_n$ are continuous in the interval* [0,1], *then the equation*

$$(-1)^{\mu-1}Af + z^n f = 0$$

has, for each region T_c (for any $c \in \mathbb{C}$) of the complex z-plane, n linearly independent solutions $f_1, \ldots, f_n$, that are regular for $z \in T_c$ for sufficiently large $|z|$, and which, with their derivatives, can be expressed in the form

$$\begin{aligned} f_j &= e^{z\omega_j x}[1], \\ \frac{df_j}{dx} &= (z\omega_j)e^{z\omega_j x}[1], \\ &\vdots \\ \frac{d^{n-1}f_j}{dx^{n-1}} &= (z\omega_j)^{n-1}e^{z\omega_j x}[1]. \end{aligned} \tag{4.9}$$

From this result and because of the special ordering chosen for the roots of minus one, we have for $j < \mu$, that the functions $e^{z\omega_j}$ decreases exponentially as $z \to \infty, z \in T_c$; hence

$$\mathcal{W}_i(f_j) = (z\omega_j)^{m_i}[\alpha_i] \quad \text{for } j < \mu. \tag{4.10}$$

Similarly we find that for $j > \mu + 1$

$$\mathcal{W}_i(f_j) = (z\omega_j)^{m_i}e^{z\omega_j}[\beta_i]. \tag{4.11}$$

Appealing to notation introduced in Propositions 5.1 and 5.2 of the proof of Proposition 2.1 in §5, we can use the simple asymptotic formulas (4.10), (4.11) to

obtain an explicit asymptotic form of the return difference equation in the z plane. To this end, again using the notation of §5, let the subscript g and f denote a quantity with respect to the basis of functions given in Proposition 5.1 and Theorem 4.1, respectively. Then for $|z| >> 0$, $z \in T_c$, $c \in \mathbb{C}$, $\lambda = i^n z^n$, the return difference equation can be written as in Proposition 5.2 as

$$\begin{aligned} \mathcal{D}(\lambda) + k\mathcal{N}(\lambda) &= \mathcal{D}_g(\lambda) + k\mathcal{N}_g(\lambda) \\ &= (\mathcal{D}_f(\lambda) + k\mathcal{N}_f(\lambda))\, W_f^{-1}(0,\lambda) \\ &\equiv \Delta(z,k). \end{aligned} \tag{4.12}$$

Substituting the asymptotic formulas in (4.10), (4.11) satisfied by the basis of solutions given in Theorem 4.1 into the formulas for the return difference equation in (4.12) and factoring out the common factors $z^{m_0}, z^{m_2}, \dots, z^{m_n}$ from the rows and also the common factors $e^{z\omega_{\mu+1}}, e^{z\omega_{\mu+2}}, \dots, e^{z\omega_n}$ from the last μ columns of the determinant $\Delta(z,k)$, the equation can be written in the form

$$\Delta(z,k) = h(z)\left(z^{\ell}\delta(z,0) + k\delta(z,\infty)\right), \tag{4.13}$$

where $\delta(z,0)$ and $\delta(z,\infty)$ are discussed in detail below,

$$h(z) = z^{m^0} e^{\omega z} W_f^{-1}(0,z) \tag{4.14}$$

and

$$\begin{aligned} m &= \sum_{j=1}^{n} m_j, \quad m^0 = \sum_{j=0,2}^{n} m_j, \\ \omega &= \sum_{j=1}^{\mu} \omega_{\mu+j}, \\ \ell &= m_1 - m_0 = m - m^0. \end{aligned} \tag{4.15}$$

A straightforward computation shows that $W_f(0,z)$ can be expressed asymptotically in terms of a simple Vandermonde determinant, namely,

$$W_f(0,z) = z^{\mu(n-1)} \det\left(\left\{\left[\omega_j^{i-1}\right]\right\}_{i=1,j=1}^{n,n}\right),$$

where

$$\det\left(\left\{\left[\omega_j^{i-1}\right]\right\}_{i=1,j=1}^{n,n}\right) = \det\left(\left\{\omega_j^{i-1}\right\}_{i=1,j=1}^{n,n}\right) + O(1/z)$$

and the absolute value of the determinant on the right hand side is

$$\left|\det\left(\left\{\omega_j^{i-1}\right\}_{i=1,j=1}^{n,n}\right)\right| = n^{\mu}.$$

Thus the nonzero roots of the return difference equation satisfy a much simpler equation, which we denote by

$$\delta(z,k) = z^{\ell}\delta(z,0) + k\delta(z,\infty) = 0, \tag{4.16}$$

where $\delta(z,0)$ is the determinant

$$\begin{vmatrix} [\alpha](U_1,\cdots,U_{\mu-1}), & \alpha]\, U_{\mu}, & [\alpha]e^{z\omega_{\mu}}\, U_{\mu+1}, & \left[0_{\mu\times(\mu-1)}\right] \\ \left[0_{\mu\times(\mu-1)}\right], & [\beta]e^{z\omega_{\mu}}\, V_{\mu}, & [\beta]V_{\mu+1}, & [\beta](V_{\mu+2},\cdots,V_n) \end{vmatrix} \tag{4.17}$$

and $\delta(z,\infty)$ is the determinant

$$(4.18)\quad \begin{vmatrix} [\alpha^0](U_1^0,\cdots,U_{\mu-1}^0), & [\alpha^0]\,U_\mu^0, & [\alpha^0]e^{z\omega_\mu}\,U_{\mu+1}^0, & [0_{\mu\times(\mu-1)}] \\ [0_{\mu\times(\mu-1)}], & [\beta]e^{z\omega_\mu}\,V_\mu, & [\beta]\,V_{\mu+1}, & [\beta](V_{\mu+2},\cdots,V_n) \end{vmatrix}$$

with

$$\begin{aligned} U_j &= [\omega_j^{m_1},\omega_j^{m_2},\ldots,\omega_j^{m_\mu}]^T, \\ U_j^0 &= [\omega_j^{m_0},\omega_j^{m_2},\ldots,\omega_j^{m_\mu}]^T, \\ V_j &= [\omega_j^{m_{\mu+1}},\omega_j^{m_{\mu+2}},\ldots,\omega_j^{m_n}]^T, \\ \alpha &= \operatorname{diag}(\alpha_1,\ldots,\alpha_\mu), \\ \alpha^0 &= \operatorname{diag}(\alpha_0,\alpha_2,\ldots,\alpha_\mu), \\ \beta &= \operatorname{diag}(\beta_1,\ldots,\beta_\mu). \end{aligned} \tag{4.19}$$

Expanding the determinants $\delta(z,0)$, $\delta(z,\infty)$ using Laplace's expansion method for $\mu\times\mu$ minors, we find that

$$\delta(z,0) = -[D_1]e^{2z\omega_\mu} + [D_{-1}], \tag{4.20}$$

$$\delta(z,\infty) = -[N_1]e^{2z\omega_\mu} + [N_{-1}], \tag{4.21}$$

where D_1, D_{-1}, N_1, N_{-1} are constants. In the notation in (4.19), these constants are given by the determinants

$$\begin{aligned} D_1 &= |\alpha||\beta|\,|U_1,\ldots,U_{\mu-1},U_{\mu+1}|\,|V_\mu,\ V_{\mu+2},\ldots,V_n|, \\ D_{-1} &= |\alpha||\beta|\,|U_1,\ldots,U_\mu|\,|V_{\mu+1},\ldots,V_n|, \\ N_1 &= |\alpha^0||\beta|\,\left|U_1^0,\ldots,U_{\mu-1}^0,U_{\mu+1}^0\right|\,|V_\mu,\ V_{\mu+2},\ldots,V_n|, \\ N_{-1} &= |\alpha^0||\beta|\,\left|U_1^0,\ldots,U_\mu^0\right|\,|V_{\mu+1},\ldots,V_n|. \end{aligned}$$

Remark 4.2. Recalling the form of the open-loop transfer function $\mathcal{G}_0(\lambda)$ in terms of $\mathcal{N}(\lambda)$, $\mathcal{D}(\lambda)$ together with the results in (4.12)–(4.21) and the relation $\lambda = i^n z^n$, we obtain a very useful asymptotic representation for the open-loop transfer function

$$\mathcal{G}_0(\lambda) = \frac{\mathcal{N}(\lambda)}{\mathcal{D}(\lambda)} = \frac{(-[N_1]e^{\omega_\mu z} + [N_{-1}]e^{-\omega_\mu z})}{z^\ell\,(-[D_1]e^{\omega_\mu z} + [D_{-1}]e^{-\omega_\mu z})}. \tag{4.22}$$

To investigate the poles and zeroes of the transfer functions, we need to establish several important properties of the determinants D_1, D_{-1}, N_1, N_{-1}. To this end we first introduce some notation and recall the definition of instantaneous gain given in Definition 2.3.

DEFINITION 4.1. *Let*

1. $\tau = N_{-1}/D_{-1}$, $\tau_1 = N_1/D_1$,

2. $v_d = D_{-1}/D_1$, $v_n = N_{-1}/N_1$.

With this we have the following proposition, whose proof is rather lengthy and technical and is provided in detail in §5.

PROPOSITION 4.2. *If the ordering of the nth roots of* (-1) *is chosen so that for* $z \in S_0$ *when* μ *is odd, for* $z \in S_1$ *when* μ *is even, then we obtain the inequalities in* (4.6) *so that the following relations hold:*

1. $$\overline{D_{-1}} = \begin{cases} D_1, & \mu \text{ odd}, \\ D_{-1}\exp\left(\frac{4m\pi i}{n}\right), & \mu \text{ even}. \end{cases}$$

2. $$\overline{N_{-1}} = \begin{cases} N_1, & \mu \text{ odd}, \\ N_{-1}\exp\left(\frac{4m^0\pi i}{n}\right), & \mu \text{ even}. \end{cases}$$

3. *For any* μ, $D_{-1} = D_1 \exp\left(\dfrac{-2m\pi i}{n}\right)$ *therefore* $v_d = \exp\left(\dfrac{-2m\pi i}{n}\right)$.

4. *For any* μ, $N_{-1} = N_1 \exp\left(\dfrac{-2m^0\pi i}{n}\right)$ *therefore* $v_n = \exp\left(\dfrac{-2m^0\pi i}{n}\right)$.

5. *From parts* 1 *and* 2 *we have*

   (a) $$\overline{\tau} = \begin{cases} \tau_1, & \mu \text{ odd} \\ \tau\exp\left(\frac{-4\ell\pi i}{n}\right), & \mu \text{ even} \end{cases} \quad \text{so for all } \mu\ \tau_1 = \tau\exp\left(-\frac{2\pi i\ell}{n}\right).$$

   (b) *The argument of* τ *is given in terms of* $s = \arg(\alpha_0/\alpha_1)/\pi + (\ell_0 + \ell)$ (*cf. Theorem* 2.4) *by*

$$\arg(\tau) = \begin{cases} s\pi + \frac{\ell\pi}{n}, & \mu \text{ odd}, \\ s\pi + \frac{2\ell\pi}{n}, & \mu \text{ even}. \end{cases}$$

It now follows from (4.13), (4.14) and (4.16)–(4.18), (4.20), (4.21), and Proposition 4.1 that the asymptotic behavior of the closed-loop poles is completely determined by the equation

$$e^{-2\omega_\mu z}\left(z^\ell[D_{-1}] + k[N_{-1}]\right) = \left(z^\ell[D_1] + k[N_1]\right),$$

which, for $\left(z^\ell[D_{-1}] + k[N_{-1}]\right) \neq 0$, is the same as

$$\begin{aligned} e^{-2\omega_\mu z} &= \left(\frac{z^\ell[D_1] + k[N_1]}{z^\ell[D_{-1}] + k[N_{-1}]}\right) \\ &= \frac{[D_1]}{[D_{-1}]}\left(\frac{z^\ell + k[N_1/D_1]}{z^\ell + k[N_{-1}/D_{-1}]}\right) \\ &= [v_d^{-1}]\left(\frac{z^\ell + k[\tau_1]}{z^\ell + k[\tau]}\right). \end{aligned} \tag{4.23}$$

Initially, the cases μ odd and even must be treated separately due to the asymptotic distribution of the roots z of $\delta(z,k)$. Namely, it is shown in [30] that when k is fixed, for μ odd the roots are asymptotic to the real axis (the bisector of S_0 and

S_{2n-1}) while for μ even the roots are asymptotic to the ray $\arg(z) = \pi/n$ (the bisector of the regions S_0 and S_1). Nevertheless the properties described in Proposition 4.2, together with the asymptotic form of the return difference equation, allow us to treat both cases at the same time once we introduce a change of variables (in the μ even case) corresponding to a rotation of the regions S_1, S_0 into S_0, S_{2n-1}, respectively.

First recall the relations

$$\omega_\mu = \begin{cases} i, & \mu \text{ odd}, \\ ie^{-i\pi/n}, & \mu \text{ even}; \end{cases}$$

$$\tau_1 = \begin{cases} \overline{\tau}, & \mu \text{ odd}, \\ \tau e^{-2\pi i\ell/n}, & \mu \text{ even}; \end{cases}$$

$$v_d = \begin{cases} e^{-2\pi mi/n}, & \mu \text{ odd}, \\ e^{-2\pi mi/n}, & \mu \text{ even}; \end{cases}$$

$$\arg(\tau) = \begin{cases} s\pi + \frac{\ell\pi}{n}, & \mu \text{ odd}, \\ s\pi + \frac{2\ell\pi}{n}, & \mu \text{ even}. \end{cases}$$

Now for μ odd, and z in S_0 we have

$$\begin{aligned} e^{-2iz} &= e^{-2\omega_\mu z} \\ &= [v_d^{-1}] \left(\frac{z^\ell + k[\overline{\tau}]}{z^\ell + k[\tau]} \right) \\ &= [e^{2\pi mi/n}] \left(\frac{z^\ell + k[\overline{\tau}]}{z^\ell + k[\tau]} \right). \end{aligned} \tag{4.24}$$

For μ even, and z in S_1 we have

$$\begin{aligned} e^{-2\omega_\mu z} &= \left(\frac{z^\ell [D_1] + k[N_1]}{z^\ell [D_{-1}] + k[N_{-1}]} \right) \\ &= \frac{[D_1]}{[D_{-1}]} \left(\frac{z^\ell + k[N_1/D_1]}{z^\ell + k[N_{-1}/D_{-1}]} \right) \\ &= [v_d^{-1}] \left(\frac{z^\ell + k[\tau_1]}{z^\ell + k[\tau]} \right) \\ &= [e^{2\pi mi/n}] \left(\frac{z^\ell + k[\tau_1]}{z^\ell + k[\tau_1 e^{2\pi i\ell/n}]} \right). \end{aligned} \tag{4.25}$$

If we make the change of variables $z = we^{i\pi/n}$, then $w \in S_0$ and

$$\omega_\mu z = ie^{-\pi i/n} we^{\pi i/n} = iw.$$

Hence the asymptotic form of the return difference equation in this case can again be written in the form

$$e^{-2iw} = e^{-2\omega_\mu z} = [e^{2\pi mi/n}]\left(\frac{w^\ell e^{\pi\ell i/n} + k[\tau_1]}{w^\ell e^{\pi\ell i/n} + k[\tau_1 e^{2\pi i\ell/n}]}\right)$$

$$= [e^{2\pi mi/n}]\left(\frac{w^\ell + k[\tau_1 e^{-\pi\ell i/n}]}{w^\ell + k[\tau_1 e^{\pi\ell i/n}]}\right)$$

$$= [e^{2\pi mi/n}]\left(\frac{w^\ell + k[\widetilde{\tau}]}{w^\ell + k[\widetilde{\tau}]}\right) \tag{4.26}$$

with

$$\widetilde{\tau} = \tau_1 e^{\pi i\ell/n},$$

and

$$\arg(\widetilde{\tau}) = \arg(\tau_1) + \frac{\pi\ell}{n} = s\pi + \pi\ell/n.$$

Therefore to carry out the asymptotic analysis for μ even or odd, we need only consider the asymptotic behavior for μ odd and $z \in T_c = c + S_0$ for an equation in the form

$$e^{-2iz} = [v]\left(\frac{z^\ell + k[\overline{\tau}]}{z^\ell + k[\tau]}\right), \qquad v = e^{2\pi mi/n}. \tag{4.27}$$

Remark 4.3. Recall that for μ odd, the argument of τ is $s\pi + \ell\pi/n$ where $1 \le \ell \le (n-1)$ and $s = \arg(\alpha_0/\alpha_1)/\pi + (\ell + \ell_0)$. Thus s is either an even or odd integer, so that $\tau \notin \mathbb{R}$. Similarly, for μ even we consider (4.26) with $\widetilde{\tau}$ having argument of the same form $\arg(\widetilde{\tau}) = s\pi + \ell\pi/n$ and exactly the same result follows.

The choice of k positive or negative is determined by the sign of $\mathcal{K}$ so that mod 2π, we have

$$\arg(k \cdot \tau) = \frac{\pi\ell}{n} \quad \text{for} \quad k \cdot \mathcal{K} > 0. \tag{4.28}$$

We see that there exists a positive constant c such that for all k satisfying $k \cdot \mathcal{K} > 0$

$$|z^\ell + k\tau| > c \tag{4.29}$$

for all $z \in S_0 \cap \{|z| > M\}$.

LEMMA 4.3. *There exists an $M > 0$ such that*

$$\left|[v]\frac{z^\ell + k[\overline{\tau}]}{z^\ell + k[\tau]}\right| \le 9/2$$

uniformly in k and z for $k \cdot \mathcal{K} > 0$ and $z \in S_0$ with $|z| > M$.

Proof. First we show that

$$\left|\frac{z^\ell + k\overline{\tau}}{z^\ell + k\tau}\right| \le 1 \tag{4.30}$$

for $k \cdot \mathcal{K} > 0$ uniformly in k and $z \in S_0$.

For $z \in S_0$, $z = re^{i\theta}$ the modulus squared of (4.30) is

$$\frac{r^{2\ell} + |k\tau|^2 + 2r^{\ell}|k\tau| \cos(\ell\theta + (\pi\ell/n))}{r^{2\ell} + |k\tau|^2 + 2r^{\ell}|k\tau| \cos(\ell\theta - (\pi\ell/n))}. \tag{4.31}$$

The expression (4.31) is less than or equal to one provided

$$\cos(\ell\theta + (\pi\ell/n)) \leq \cos(\ell\theta - (\pi\ell/n)),$$

which is true provided $0 \leq \theta \leq \pi/n$, $k \cdot \mathcal{K} > 0$.

To complete the proof of the lemma we note that our expression can be written as

$$(v + O(1/z)) \frac{z^{\ell} + k\overline{\tau} + kO(1/z)}{z^{\ell} + k\tau + kO(1/z)}.$$

Divide the numerator and denominator by $z^{\ell} + k\tau$ and define

$$T_1(z,k) = \frac{z^{\ell} + k\overline{\tau}}{z^{\ell} + k\tau},$$

$$T_2(z,k) = \frac{k}{z^{\ell} + k\tau}.$$

Recalling that $|v| = 1$, the modulus of the resulting expression can be written as

$$\left| [v] \frac{T_1(z,k) + T_2(z,k)O(1/z)}{1 + T_2(z,k)O(1/z)} \right|.$$

From (4.30), we see that

$$|T_1(z,k)| \leq 1$$

uniformly in z and k.

Note that for every $M > 0$, the map

$$(k,z) \mapsto \frac{z^{\ell}}{k}$$

takes $(0,\infty) \times \{S_0 \cap \{|z| > M\}\}$ onto $S_0^{\ell} = \{w : 0 \leq \arg(w) \leq \ell\pi/n\}$ and it maps $(-\infty,0) \times \{S_0 \cap \{|z| > M\}\}$ onto $-S_0^{\ell} = \{w : \pi \leq \arg(w) \leq \pi + \ell\pi/n\}$.

For $T_2(z,k)$ we define $w = z^{\ell}/k$ and consider the function

$$\frac{1}{w + \tau}$$

on S_0^{ℓ} or $-S_0^{\ell}$ depending on whether $k > 0$ or $k < 0$. By our assumption $k \cdot \mathcal{K} > 0$ we see that $w + \tau \neq 0$ and hence

$$\left| \frac{1}{w + \tau} \right| \leq C$$

for some constant C. Therefore

$$|T_2(z,k)| \leq C$$

uniformly in k and $z \in S_0$. With this we choose $M > 0$ so that for $|z| > M$ the term $|O(1/z)|$ satisfies

$$|O(1/z)| \leq \frac{1}{2C}$$

and

$$|[v]| \leq (1 + 1/2).$$

So finally we obtain

$$\left|[v]\frac{z^\ell + k[\overline{\tau}]}{z^\ell + k[\tau]}\right| \leq (1+1/2)\frac{1+1/2}{1-1/2} \leq \frac{9}{2}. \qquad \square$$

THEOREM 4.4. *There exists an $M > 0$ and $y_0 > 0$ such that*

$$e^{-2iz} = [v]\left(\frac{z^\ell + k[\overline{\tau}]}{z^\ell + k[\tau]}\right)$$

has no roots for $z \in (S_0 \cup S_{2n-1}) \cap \{z : |\operatorname{Im}(z)| \geq y_0,\ |z| > M\}$ and all k such that $k \cdot \mathcal{K} > 0$.

Proof. For $z \in S_0$, $z = x + iy$ we have

$$\left|e^{-2iz}\right| = e^{2y} > e^{2y_0}$$

for $y > y_0 > 0$ and we see that the modulus of the left side of our equation can be made as large as we like. From the previous lemma, we need only take $y_0 > 0$ so that

$$e^{2y_0} > \frac{9}{2}.$$

So for $|z| > M$ and $\operatorname{Im}(z) > y_0$ there are no roots for all k so that $k \cdot \mathcal{K} > 0$. The result for $\operatorname{Im}(z) < -y_0$ follows by conjugation. $\square$

Remark 4.4. Recall that we have reduced the general problem of the asymptotic behavior of the closed-loop poles to the case in which μ is odd. For μ odd, recall that the zeros of $\mathcal{D}(\lambda) + k\mathcal{N}(\lambda)$ of large modulus and $\pi \leq \arg(\lambda) < 2\pi$ are exactly the set of $\lambda = -z^n$ where z are the zeros of large modulus of $\Delta(z,k)$ (and hence $\delta(z,k)$) in S_0. Indeed, for any $c \in \mathbb{C}$ the zeros of $\Delta(z,k)$ and $\mathcal{D}(-z^n) + k\mathcal{N}(-z^n)$ agree for $z \in T_c$, $|z| >> 0$. The zeros of $\mathcal{D}(\lambda) + k\mathcal{N}(\lambda)$ occur in conjugate pairs and we have shown that for large modulus there are no zeros of $\delta(z,k)$ for $\operatorname{Im}(z) > y_0$ so there are no zeros of $\mathcal{D}(-z^n) + k\mathcal{N}(-z^n)$ for $\operatorname{Im}(z) > y_0$ and hence for $\operatorname{Im}(z) < -y_0$. From this we observe that there are no zeros of $\delta(z,k)$ or $\mathcal{D}(-z^n) + k\mathcal{N}(-z^n)$ for $|z|$ large, $|\operatorname{Im}(z)| > y_0$ in $S_0 \cup S_{2n-1}$.

We now proceed to show that for $|\lambda|$ large the closed-loop poles are real, negative, simple and move to the left for $k \cdot \mathcal{K} > 0$.

LEMMA 4.5. *If $k \cdot \mathcal{K} \geq 0$, then the roots of*

$$e^{-2iz} = (v)\left(\frac{z^\ell + k\overline{\tau}}{z^\ell + k\tau}\right)$$

are all real and simple for $z \in S_0 \cup S_{2n-1}$.

Proof. If $z = r\exp(i\theta) \in S_0$, then $0 \le \theta \le \pi/n$ and

$$e^{4r\sin\theta} = |e^{-2iz}|^2 = \left|\frac{(z^\ell + k\overline{\tau})}{(z^\ell + k\tau)}\right|^2 \le 1$$

by (4.30), from which it follows that $\theta = 0$ and hence a root z must be real.

To see that the roots are simple, let

$$h(z,k) = e^{-2iz} - (v)\left(\frac{z^\ell + k\overline{\tau}}{z^\ell + k\tau}\right).$$

Then for $z \in \mathbb{R}_+$, we have

$$\frac{\partial h}{\partial z}(z,k) = -2ie^{-2iz} - (v)\,\frac{\ell z^{\ell-1}(z^\ell + k\tau) - \ell z^{\ell-1}(z^\ell + k\overline{\tau})}{(z^\ell + k\tau)^2}$$

$$= -2ie^{-2iz} - (v)\,\frac{2i\ell z^{\ell-1} k\,\mathrm{Im}(\tau)}{(z^\ell + k\tau)^2}$$

$$= -2i\left((v)\,\frac{z^\ell + k\overline{\tau}}{z^\ell + k\tau} + (v)\,\frac{\ell z^{\ell-1} k\,\mathrm{Im}(\tau)}{(z^\ell + k\tau)^2}\right)$$

$$= \frac{-2i\,(v)}{(z^\ell + k\tau)^2}\left(|z^\ell + k\overline{\tau}|^2 + |\ell z^{\ell-1} k\,\mathrm{Im}(\tau)|\right) \neq 0.$$

Therefore the real positive roots are simple. □

Assume $k \cdot \mathcal{K} > 0$ which implies $\arg(k\tau) = \pi\ell/n$, and let

$$g = \frac{k}{(p\pi)^\ell}, \qquad p = 1, 2, \ldots,$$

$$a_1 = \frac{\pi - \arg(v) + \pi\ell/n}{2}, \qquad a_2 = \pi + a_1$$

and define

$$V = \{\tilde{z} \;:\; |\,\mathrm{Im}(\tilde{z})| < y_0,\ a_1 < \mathrm{Re}(\tilde{z}) < a_2\}.$$

With this, for any $z \in V + p\pi$ we have $z = \tilde{z} + p\pi$, $\tilde{z} \in V$ and (4.27) can be written as

$$e^{-2i\tilde{z}} = [v]\left(\frac{(\tilde{z} + p\pi)^\ell + k[\overline{\tau}]}{(\tilde{z} + p\pi)^\ell + k[\tau]}\right), \qquad v = e^{2\pi mi/n}. \tag{4.32}$$

Thus we let

$$F_p(\tilde{z}, g) = e^{-2i\tilde{z}} - [v]\left(\frac{(1 + \tilde{z}/(p\pi))^\ell + g[\overline{\tau}]}{(1 + \tilde{z}/(p\pi))^\ell + g[\tau]}\right),$$

$$f_p(\tilde{z}, g) = e^{-2i\tilde{z}} - v\left(\frac{(1 + \tilde{z}/(p\pi))^\ell + g\overline{\tau}}{(1 + \tilde{z}/(p\pi))^\ell + g\tau}\right),$$

$$h(\tilde{z}, g) = e^{-2i\tilde{z}} - v\frac{1+g\overline{\tau}}{1+g\tau},$$

for $\tilde{z}$ in a neighborhood of $\overline{V}$ (recall $[a] = a + O(1/p)$).

Our goal now is to show that for p sufficiently large, $F_p(\tilde{z}, g)$ has only one zero in $\overline{V}$ that, since the roots must occur in conjugate pairs, must be real. To this end we first show that

$$|f_p(\tilde{z}, g) - h(\tilde{z}, g)|$$

goes to zero uniformly in g for $\tilde{z} \in \overline{V}$ as p goes to infinity. We have

$$\begin{aligned}
|f_p(\tilde{z}, g) - h(\tilde{z}, g)| &= \left| \frac{(1+\tilde{z}/(p\pi))^{\ell} + g\overline{\tau}}{(1+\tilde{z}/(p\pi))^{\ell} + g\tau} - \frac{1+g\overline{\tau}}{1+g\tau} \right| \\
&= \left| \frac{2g\,\mathrm{Im}(\tau)\left\{1 - (1+\tilde{z}/(p\pi))^{\ell}\right\}}{\left\{(1+\tilde{z}/(p\pi))^{\ell} + g\tau\right\}(1+g\tau)} \right| \\
&\leq O(1/p),
\end{aligned}$$

where we have used the fact that

$$\left| \frac{2g\,\mathrm{Im}(\tau)}{\left\{(1+\tilde{z}/(p\pi))^{\ell} + g\tau\right\}(1+g\tau)} \right|$$

is uniformly bounded for $\tilde{z} \in \overline{V}$ and for $g \in \mathbb{R}_{\pm}$ (depending on the requirement that $g \cdot \mathcal{K} > 0$).

Now we define θ_g by

$$\frac{1+g\overline{\tau}}{1+g\tau} \equiv e^{-i\theta_g}$$

and by our choice of a_j, $j = 1, 2$ and hence V we see that $h(\tilde{z}, g)$ has exactly one real root in V that can be written in terms of θ_g as

$$\tilde{z}(g) = -\frac{\arg(v)}{2} + \frac{\theta_g}{2} + \pi$$

and we now show there exists a constant C such that

$$|h(\tilde{z}, g)| > C, \qquad \tilde{z} \in \partial V$$

for all g.

If $\tilde{z} = a_j + iy \in \partial V$, $j = 1, 2$ then

$$\begin{aligned}
|h(\tilde{z}, g)|^2 &= \left| \exp(-2ia_j + 2y) - v\frac{1+g\overline{\tau}}{1+g\tau} \right|^2 \\
&= |\exp(-i(\arg(g\tau) - \arg(v)) + 2y) + \exp(i(\arg(v) - 2\arg(1+g\tau)))|^2 \\
&= |\exp(2y - i\arg(g\tau)) + \exp(-2i\arg(1+g\tau))|^2 \\
&= |\exp(2y) + \exp(i\arg(g\tau) - 2i\arg(1+g\tau))|^2.
\end{aligned}$$

Now

$$\arg(g\tau) = \frac{\ell\pi}{n} \in \left[\frac{\pi}{n}, \frac{(n-1)\pi}{n}\right],$$

which implies that

$$\arg(g\tau) > \arg(1+g\tau) > 0.$$

This in turn implies

$$\begin{aligned} -\frac{(n-1)}{n}\pi \le -\arg(g\tau) &< -\arg(1+g\tau) \\ &< \arg(g\tau) - 2\arg(1+g\tau) < \arg(g\tau) \le \frac{(n-1)}{n}\pi; \end{aligned}$$

that is,

$$|\arg(g\tau) - 2\arg(1+g\tau)| \le \frac{(n-1)}{n}\pi.$$

Let $\theta = \arg(g\tau) - 2\arg(1+g\tau)$, so that $|\theta| \le \dfrac{(n-1)}{n}\pi$ and we have

$$\begin{aligned} |h(\tilde{z}, g)|^2 &= |e^{2y} + e^{i\theta}|^2 \\ &= e^{4y} + 1 + 2e^{2y}\cos(\theta) \\ &= e^{4y} - 2e^{2y} + 1 + 2e^{2y} + 2e^{2y}\cos(\theta) \\ &= (e^{2y}-1)^2 + 2e^{2y}(1+\cos(\theta)) \\ &\ge (e^{2y}-1)^2 + 2e^{2y}\left(1 + \cos\left(\frac{(n-1)}{n}\pi\right)\right) \\ &= (e^{2y}-1)^2 + 2e^{2y}\left(1 - \cos\left(\frac{\pi}{n}\right)\right) \\ &\ge 2e^{-2y_0}\left(1 - \cos\left(\frac{\pi}{n}\right)\right) > 0. \end{aligned}$$

If $\tilde{z} = x \pm iy_0 \in \partial V$, then

$$\begin{aligned} |h(\tilde{z}, g)| &= |e^{\pm 2y_0 - 2ix} - v\frac{1+g\overline{\tau}}{1+g\tau}| \\ &\ge \left| |e^{\pm 2y_0}| - |v\frac{1+g\overline{\tau}}{1+g\tau}| \right| \\ &= |e^{\pm 2y_0} - 1| > 0 \end{aligned}$$

for $y_0 \neq 0$. Therefore there exists a constant C such that

$$|h(\tilde{z}, g)| > C, \qquad \tilde{z} \in \partial V$$

for all g.

Let P_1 be a number such that for $p > P_1$

$$|h(\tilde{z}, g) - f_p(\tilde{z}, g)| < C/2$$

and hence for $\tilde{z} \in \partial V$ and all g we have

$$|f_p(\tilde{z}, g)| \geq |h(\tilde{z}, g)| - |h(\tilde{z}, g) - f_p(\tilde{z}, g)| > C/2.$$

In a similar manner we have

$$\begin{aligned}
|F_p(\tilde{z}, g) - f_p(\tilde{z}, g)| &= \left| [v] \frac{(1 + \tilde{z}/(p\pi))^{\ell} + g[\overline{\tau}]}{(1 + \tilde{z}/(p\pi))^{\ell} + g[\tau]} - (v) \frac{(1 + \tilde{z}/(p\pi))^{\ell} + g\overline{\tau}}{(1 + \tilde{z}/(p\pi))^{\ell} + g\tau} \right| \\
&\leq \left| \frac{(1 + \tilde{z}/(p\pi))^{\ell} + g[\overline{\tau}]}{(1 + \tilde{z}/(p\pi))^{\ell} + g[\tau]} - \frac{(1 + \tilde{z}/(p\pi))^{\ell} + g\overline{\tau}}{(1 + \tilde{z}/(p\pi))^{\ell} + g\tau} \right| \\
&\quad + \left| \frac{(1 + \tilde{z}/(p\pi))^{\ell} + g[\overline{\tau}]}{(1 + \tilde{z}/(p\pi))^{\ell} + g[\tau]} \right| |O(1/p)| \\
&= \left| \frac{g}{(1 + \tilde{z}/(p\pi))^{\ell} + g[\tau]} \right| |O(1/p)| + |O(1/p)| \\
&= O(1/p)
\end{aligned}$$

uniformly for $\tilde{z} \in \overline{V}$ and $g \cdot \mathcal{K} > 0$. Hence there exists a $P_2 > P_1$ such that for $p > P_2$

$$|F_p(\tilde{z}, g) - f_p(\tilde{z}, g)| \leq C/6$$

for $\tilde{z} \in \overline{V}$ and $g \cdot \mathcal{K} > 0$. Therefore

$$|F_p(\tilde{z}, g)| \geq |f_p(\tilde{z}, g)| - |f_p(\tilde{z}, g) - F_p(\tilde{z}, g)| > C/3$$

for $p > P_2$ and $\tilde{z} \in \partial V$, $g \cdot \mathcal{K} > 0$. The inequalities

$$|f_p| > C/2, \quad |F_p| > C/3$$

guarantee that f_p and F_p have no zeros on ∂V so that we can apply Rouche's theorem.

For P_2, we have

$$|F_p(\tilde{z}, g) - f_p(\tilde{z}, g)| \leq C/6 < C/2 < |f_p(\tilde{z}, g)|$$

for all $p > P_2$, $\tilde{z} \in \partial V$ and $g \cdot \mathcal{K} > 0$. Thus we can apply Rouche's theorem to F_p and f_p to conclude that $F_p(\tilde{z}, g)$ has only one zero in V, which must be real since the zeros of $f_p(\tilde{z}, g)$ occur in conjugate pairs. □

Remark 4.5. When $k = 0$, (4.27) reduces to

$$e^{-2iz} = [e^{2\pi mi/n}]$$

with zeros

$$\tilde{z}_j(0) = -m\pi/n + j\pi + O(1/j), \qquad j = 0, \pm 1, \ldots. \tag{4.33}$$

Similarly, for $k = \infty$ (4.27) reduces to

$$e^{-2iz} = [e^{2\pi m^0 i/n}]$$

with zeros

$$\tilde{z}_j(\infty) = -m^0\pi/n + j\pi + O(1/j),\ j = 0, \pm 1, \ldots. \tag{4.34}$$

THEOREM 4.6. *For $|z|$ large, the zeros of*

$$F(z,k) \equiv e^{-2iz} - \left[e^{2\pi mi/n}\right] \frac{z^\ell + k[\overline{\tau}]}{z^\ell + k[\tau]}$$

are continuous, monotone increasing functions of k for $|k| \to \infty$ with $k \cdot \mathcal{K} > 0$.

Proof. Because we have shown that the roots of large modulus are real and simple, we know that

$$\frac{\partial F}{\partial z}(z,k) \neq 0$$

at these real zeros of $F(z,k)$, and thus we can apply the implicit function theorem to show that these zeros $z(k)$ of $F(z,k)$ vary continuously in k.

To see that a real zero $z(k)$ is monotone in k, let us suppose that $F(z,k_1) = F(z,k_2)$ and hence

$$\frac{z^\ell + k_1[\overline{\tau}]}{z^\ell + k_1[\tau]} = \frac{z^\ell + k_2[\overline{\tau}]}{z^\ell + k_2[\tau]},$$

which implies

$$k_1[\overline{\tau}] + k_2[\tau] = k_1[\tau] + k_2[\overline{\tau}]$$

or

$$(k_1 - k_2)[\mathrm{Im}(\tau)] = 0, \qquad \Rightarrow k_1 = k_2,$$

and hence $z(k)$ is monotone.

To show that the roots move to the right, we first show that if $z_p(0)$, $z_q(\infty)$ are zeros for $k = 0$ and $k = \infty$, respectively, that lie in a region

$$j\pi + a_1 < z < j\pi + a_2 = (j+1)\pi + a_1$$

then

$$z_p(0) < z_q(\infty).$$

As shown above,

$$F(z,0) = e^{-2iz} - [v] = 0$$

implies

$$z_p(0) = p\pi - \frac{\arg(v)}{2} + o(1/p),$$

while

$$F(z,\infty) = e^{-2iz} - \left[v\frac{\overline{\tau}}{\tau}\right]$$

implies

$$z_q(\infty) = q\pi - \frac{\arg(v)}{2} + \ell\pi/n + o(1/q).$$

If

$$j\pi + a_1 < p\pi - \frac{\arg(v)}{2} + o(1/p) < (j+1)\pi + a_1,$$

then subtracting $j\pi$ and adding $\arg(v)/2$, we obtain

$$0 < \frac{\pi}{2} + \frac{\ell\pi}{2n} < (p-j)\pi + o(1/p) < \pi + \frac{\pi}{2} + \frac{\ell\pi}{2n} < 2\pi.$$

Hence, we see that $p = (j+1)$.

If

$$j\pi + a_1 < q\pi - \frac{\arg(v)}{2} + \frac{\ell\pi}{n} + o(1/q) < (j+1)\pi + a_1,$$

then

$$\frac{\pi}{2} + \frac{\ell\pi}{2n} < (q-j)\pi + \frac{\ell\pi}{n} + o(1/q) < \pi + \frac{\pi}{2} + \frac{\ell\pi}{2n}$$

and

$$\frac{\pi}{2} < (q-j)\pi + \frac{\ell\pi}{2n} + o(1/q) < \pi + \frac{\pi}{2}.$$

Hence, $q = (j+1) = p$ and

$$\begin{aligned} z_p(0) - z_p(\infty) &= -\frac{1}{2}\arg(v) + \frac{1}{2}\arg(v) - \frac{\ell\pi}{n} + o(1/p) \\ &= -\frac{\ell\pi}{n} + o(1/p) < 0. \end{aligned}$$

Thus the zero of $F(z,k)$ on a particular branch must lie to the right of the corresponding zero of $F(z,0)$ and to the left of the zero of $F(z,\infty)$. Since they are monotone, they must move to the right. □

Remark 4.6. Since these roots are simple, the corresponding eigenvalues are simple and must occur in conjugate pairs so they must be real for all k. Thus under the Assumption 2.1 on the order of inputs and outputs and sign of the gain chosen by the instantaneous gain formula, we see that there are no unbounded branches of the root locus; sufficiently large modulus open-loop poles and zeros are real and interlace on the negative real axis; for these open-loop poles and zeros, the closed-loop poles are real and vary from the open-loop poles to the open-loop zeros.

Finally, we consider the possible finite number of multiple and/or complex roots.

THEOREM 4.7. *The resolvent* $R(\lambda,k) = (A_k - \lambda I)^{-1}$ *is holomorphic in* k *in the norm resolvent sense and the spectrum of* A_k *is precisely the set* $\{\lambda_j(k)\}$ *of closed-loop poles for the system* (2.2)–(2.13). *Every finite system of eigenvalues of* A_k *vary continuously for* $k \cdot \mathcal{K} \geq 0$.

Proof. The proof consists of providing an explicit representation (following [30]) for the resolvent in the form

$$R(\lambda,k)f(x)=\int_0^1 G_k(x,\xi;\lambda)f(\xi)\,d\xi,$$

where $G_k(x,\xi;\lambda)$ is the resolvent kernel for A_k. The explicit form of the resolvent kernel will show that the resolvent is analytic.

Let $\{g_j(x,\lambda)\}_{j=1}^n$ denote the unique basis of solutions of $(A-\lambda)f=0$ given in Proposition 5.1 (we will also use the notation $g_j(x)\equiv g_j(x,\lambda)$ and in what follows we suppress the dependence on λ to simplify notation) that are entire functions of λ and satisfy

$$g_j^{(k-1)}(0,\lambda)=\delta_{jk}.$$

Furthermore, define

$$\widehat{W}(x)=\left(\{g_j^{(i-1)}(x)\}_{i=1,j=1}^{n,n}\right)$$

so that the Wronskian is given by

$$W(x)=\det\left(\widehat{W}(x)\right).$$

With this, we can define

$$g(x,\xi)=\frac{\pm 1}{2W(\xi)}\begin{vmatrix} g_1(x) & g_2(x) & \cdots & g_n(x)\\ g_1^{n-2}(\xi) & g_2^{n-2}(\xi) & \cdots & g_n^{n-2}(\xi)\\ \vdots & \vdots & \ddots & \vdots\\ g_1(\xi) & g_2(\xi) & \cdots & g_n(\xi)\end{vmatrix},$$

where the positive sign is taken if $x>\xi$ and the negative sign is taken if $x<\xi$. As in the formula for the closed-loop transfer function in (5.45), we let $\Delta(\lambda,k)=\mathcal{D}(\lambda)+k\mathcal{N}(\lambda)$. Then the resolvent kernel is given by

$$G_k(x,y;\lambda)=\frac{(-1)^n}{\Delta(\lambda,k)}H_k(x,y;\lambda),$$

where

$$H_k(x,\xi;\lambda)=\begin{vmatrix} g_1(x) & \cdots & g_n(x) & g(x,\xi)\\ \mathcal{W}_1(g_1)+k\mathcal{W}_0(g_1) & \cdots & \mathcal{W}_1(g_n)+k\mathcal{W}_0(g_n) & \mathcal{W}_1(g)+k\mathcal{W}_0(g)\\ \mathcal{W}_2(g_1) & \cdots & \mathcal{W}_2(g_n) & \mathcal{W}_2(g)\\ \vdots & \vdots & \ddots & \vdots\\ \mathcal{W}_n(g_1) & \cdots & \mathcal{W}_n(g_n) & \mathcal{W}_n(g)\end{vmatrix}.$$

From this we see that the spectrum of A_k coincides exactly with the closed-loop poles; namely, the zeros of the return difference equation,

$$\Delta(\lambda,k)=\mathcal{D}(\lambda)+k\mathcal{N}(\lambda)=0.$$

Thus the resolvent kernel is an analytic function of λ and k with singularities only at the closed-loop poles.

Since we have shown that there are no roots of Δ for all $k \cdot \mathcal{K} > 0$, $|z| > M$, $z \in S_0 \cup S_{2n-1}$, there is a complex number z_0 such for $\lambda_0 = -z_0^n$

$$\operatorname{dist}(\lambda_0, \sigma(A_k)) > (\text{constant}) > 0$$

for all k where $\sigma(A_k)$ denotes the spectrum of A_k. Thus the resolvent operator $R_k(\lambda_0) = (A_k - \lambda_0)^{-1}$ maps the entire Hilbert space onto the domain of A_k for every k. Therefore by Kato [25, Chap. VII, §1.2, Thm. 1.3] A_k is a holomorphic family for all k. From this we see that, in our case, if the boundary conditions are holomorphic in the parameter k, then the resulting family of operators A_k is holomorphic in the generalized sense of Kato [25].

A further result of the proof of Theorem 3.1 is that the operators A_k satisfy the separation of the spectrum condition (cf. [25, Chap. III, §6.4]). Namely, we can find a rectifiable curve Γ separating the spectrum of A_k into a finite part contained in the interior of Γ and the remainder of the spectrum that consists of an infinite set of real negative values $\lambda_j(k)$ that converge to minus infinity as j tends to infinity. In particular, for $z_j(k)$ a zero of $\delta(z,k)$ of large modulus, we have an integer p so that for all $j' > j$ and the closed-loop poles $z_{j'}(k)$ satisfy

$$a_1 + p\pi < z_j(0) \le z_j(k) \le z_j(\infty) < a_2 + p\pi < z_{j'}(k).$$

So it is easy to find such a curve Γ. Furthermore the decomposition for this curve Γ holds for all $k \cdot \mathcal{K} \ge 0$. Thus for $k \cdot \mathcal{K} \ge 0$, the spectrum of A_k satisfies the spectrum separation property described in Kato [25].

Appealing once again to results found in Kato [25, Chap. VII, §1.3, Thms. 1.7, 1.8] we see that the finite number of eigenvalues of A_k, $k \cdot \mathcal{K} \ge 0$, inside Γ vary continuously in k while the remainder of the spectrum of A_k for all $k \cdot \mathcal{K} \ge 0$ is real and has been shown to varies continuously on the negative real axis. □

5. Proofs of the results in §2 and Theorem 3.4. We begin this section with the proof of Proposition 2.1.

Proof of Proposition 2.1. The proof is based on the Laplace transform, straightforward calculations, and two propositions which are given below in Proposition 5.1 and 5.2. We first obtain a representation for the transfer function in terms of a special basis of solutions of a boundary value problem for an ordinary differential equation. Then we show that this representation is independent of the basis chosen.

Applying the Laplace transform to (2.2)–(2.6) with initial data $w_0 = 0$ we obtain the system

$$\begin{aligned} A\hat{w} &= \lambda\hat{w}, \\ \mathcal{W}_1(\hat{w}) &= \hat{u}, \\ \mathcal{W}_i(\hat{w}) &= 0, \qquad i = 2, \dots, n, \\ \hat{y} &= \mathcal{W}_0(\hat{w}). \end{aligned} \tag{5.35}$$

Computation of a candidate for the closed-loop transfer function can be carried out once we have sufficient knowledge of a basis of eigenfunctions and eigenvalues for the problem

$$Af = \lambda f \tag{5.36}$$

$$\mathcal{W}_1(f) + k\,\mathcal{W}_0(f) = 0, \quad \mathcal{W}_i(f) = 0, \qquad i = 2, \dots, n.$$

PROPOSITION 5.1. *There is a unique basis of solutions* $\{g_j(x,\lambda)\}_{j=1}^n$ *of*

$$g^{(n)} + (-1)^{(\mu-1)}\left(p_{n-2}(x)g^{(n-2)} + \cdots + p_1(x)g^{(1)} + p_0(x)g - \lambda g\right) = 0 \tag{5.37}$$

that are entire functions of λ *and satisfy*

$$g_j^{(k-1)}(0,\lambda) = \delta_{jk}. \tag{5.38}$$

Furthermore, these functions are real in the sense that

$$\overline{g(x,\lambda)} = g(x,\overline{\lambda})$$

under the assumption that the coefficients $\{p_{n-j}\}_{j=2}^n$ *are real.*

The proof of this result is classical; see, for example, [30].

Let $g_j \equiv g_j(x,\lambda)$ denote the basis of solutions of (5.36) given in Proposition 5.1. Then the solution to (5.35) is given by

$$\hat{w} = \sum_{j=1}^n a_j g_j$$

with $\{a_j\}$ determined from the system

$$\sum_{j=1}^n a_j \mathcal{W}_1(g_j) = \hat{u}, \tag{5.39}$$

$$\sum_{j=1}^n a_j \mathcal{W}_i(g_j) = 0, \qquad i = 2,\ldots,n.$$

Applying Cramer's rule we obtain

$$\hat{w} = \det\left(\{\mathcal{W}_i(g_j)\}_{i=1,j=1}^{n,n}\right)^{-1} \sum_{j=1}^n C_{1j} g_j \hat{u},$$

where C_{1j} is the $1j$th cofactor of $\{\mathcal{W}_i(g_j)\}_{i=1,j=1}^{n,n}$. Now apply the observation operator $C = \mathcal{W}_0$ to obtain

$$\hat{y} = \det\left(\{\mathcal{W}_i(g_j)\}_{i=1,j=1}^{n,n}\right)^{-1} \sum_{j=1}^n C_{1j}\, \mathcal{W}_0(g_j)\hat{u}.$$

The sum on the right is precisely $\det(\{\mathcal{W}_i(g_j)\}_{i=0,2,j=1}^{n,n})$ and so we have

$$\mathcal{G}_0 = \frac{\mathcal{N}}{\mathcal{D}}, \tag{5.40}$$

where

$$\mathcal{N}(\lambda) \equiv \det\left(\{\mathcal{W}_i(g_j)\}_{i=0,2,j=1}^{n,n}\right), \tag{5.41}$$

$$\mathcal{D}(\lambda) \equiv \det\left(\{\mathcal{W}_i(g_j)\}_{i=1,j=1}^{n,n}\right), \tag{5.42}$$

$$\hat{y} = \mathcal{G}_0 \hat{u}. \tag{5.43}$$

Now consider the feedback

$$u(t) = -ky(t) + v(t). \tag{5.44}$$

The closed-loop transfer function, denoted by $\mathcal{G}_k$, is obtained using exactly the same argument as above but with $\mathcal{W}_1$ replaced by $\mathcal{W}_1 + k\mathcal{W}_0$. In this case the expression (5.42) is replaced by an expression in which only the first row of the determinant is changed. In particular, $\mathcal{W}_1$ is replaced by $\mathcal{W}_1 + k\mathcal{W}_0$. Expanding this determinant using the first row gives

$$\hat{y} = \mathcal{G}_k \hat{v},$$

where

$$\mathcal{G}_k(\lambda) = \frac{\mathcal{N}(\lambda)}{\mathcal{D}(\lambda) + k\mathcal{N}(\lambda)}. \tag{5.45}$$

The functions $\mathcal{D}(\lambda)$, $\mathcal{N}(\lambda)$ are entire functions with discrete zeros (cf. [30])

$$\{\lambda_j(0)\}_{j=1}^{\infty},\ \{\lambda_j(\infty)\}_{j=1}^{\infty}$$

and provide the spectrum of A_0 and A_∞, respectively. The spectrum of A_k is obtained in the same way from the zeros of

$$\mathcal{D}(\lambda) + k\mathcal{N}(\lambda) = 0.$$

That $\mathcal{G}_k(\lambda)$ is real follows from Proposition 5.1 where it was observed that

$$\overline{g_j(x,\lambda)} = g_j(x,\overline{\lambda}).$$

So far, our calculations of $\mathcal{G}_0$ apparently depend on a choice of basis. In fact, this dependence is superfluous.

PROPOSITION 5.2. *Suppose that* $\{f_j\}$ *and* $\{g_j\}$ *are both bases for* (5.36) *where* $\{g_j\}$ *is the basis of Proposition* 5.1 *and we define*

$$\mathcal{D}_g(\lambda) = \det\left(\{\mathcal{W}_i(g_j)\}_{i=1,j=1}^{n,n}\right), \qquad \mathcal{N}_g(\lambda) = \det\left(\{\mathcal{W}_i(g_j)\}_{i=0,2,j=1}^{n,n}\right),$$

and

$$\mathcal{D}_f(\lambda) = \det\left(\{\mathcal{W}_i(f_j)\}_{i=1,j=1}^{n,n}\right), \qquad \mathcal{N}_f(\lambda) = \det\left(\{\mathcal{W}_i(g_j)\}_{i=0,2,j=1}^{n,n}\right).$$

Let

$$\widehat{W}_f(x) = \left(\{f_j^{(i-1)}(x)\}_{i=1,j=1}^{n,n}\right)$$

and denote the Wronskian for the basis $\{f_j\}$ *by*

$$W_f(x) = \det\left(\widehat{W}_f(x)\right).$$

Then

$$\mathcal{G}_0(\lambda) = \frac{\mathcal{N}_g(\lambda)}{\mathcal{D}_g(\lambda)} = \frac{\mathcal{N}_f(\lambda)W_f^{-1}(0)}{\mathcal{D}_f(\lambda)W_f^{-1}(0)} = \frac{\mathcal{N}_f(\lambda)}{\mathcal{D}_f(\lambda)} \tag{5.46}$$

from which it follows that the definition of our transfer function is invariant with respect to the basis chosen for (5.36).

Proof. We have

$$\widehat{W}_g(x,\lambda) = \widehat{W}_f(x,\lambda)\widehat{W}_f^{-1}(0,\lambda)$$

and from this it can be shown that

$$\mathcal{D}_g(\lambda) = W_f^{-1}(0,\lambda)\,\mathcal{D}_f(\lambda),$$

and

$$\mathcal{N}_g(\lambda) = W_f^{-1}(0,\lambda)\,\mathcal{N}_f(\lambda).$$

This follows from the relation

$$g_k(x) = \sum_{j=1}^{n} c_{kj} f_j(x),$$

which implies

$$\delta_{ik} = g_k^{(i-1)}(0) = \sum_{j=1}^{n} c_{kj} f_j^{(i-1)}(0)$$

and hence

$$C = \left(\{f_j^{(i-1)}(0)\}_{i=1,j=1}^{n,n}\right) = W_f^{(-1)}(0).$$

Let

$$a_{1i} = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{im_i-1}, \alpha_i, \overbrace{0, \ldots, 0}^{n-m_i}), \qquad i = 1, \ldots, \mu,$$

$$b_{1i} = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{im_i-1}, \overbrace{0, \ldots, 0}^{n-m_i+1}), \qquad i = 1, \ldots, \mu,$$

$$a_{2i} = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{im_i-1}, \overbrace{0, \ldots, 0}^{n-m_i+1}), \qquad i = \mu+1, \ldots, n,$$

$$b_{2i} = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{im_i-1}, \beta_i, \overbrace{0, \ldots, 0}^{n-m_i}), \qquad i = \mu+1, \ldots, n$$

and define the four $\mu \times n$ matrices

$$A_j = \begin{pmatrix} a_{j1} \\ a_{j2} \\ \vdots \\ a_{j\mu} \end{pmatrix}, \quad B_j = \begin{pmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{j\mu} \end{pmatrix}, \qquad j = 1, 2.$$

With this notation we have

$$\begin{aligned}\mathcal{D}_g(\lambda) &= \det\left(\begin{bmatrix} A_1 & B_1 \\ A_2 & B_2 \end{bmatrix}\begin{bmatrix} \widehat{W}_g(0) \\ \widehat{W}_g(1) \end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} A_1 & B_1 \\ A_2 & B_2 \end{bmatrix}\begin{bmatrix} \widehat{W}_f(0) \\ \widehat{W}_f(1) \end{bmatrix}\widehat{W}_f^{-1}(0)\right) \\ &= \det\left(\begin{bmatrix} A_1 & B_1 \\ A_2 & B_2 \end{bmatrix}\begin{bmatrix} \widehat{W}_f(0) \\ \widehat{W}_f(1) \end{bmatrix}\right) W_f^{-1}(0) \\ &= \mathcal{D}_f(\lambda)W_f^{-1}(0).\end{aligned}$$

The analogous result for $\mathcal{N}_g$ and $\mathcal{N}_f$ is exactly the same. □

The proof that $\mathcal{D}$ and $\mathcal{N}$ are entire functions of order $1/n$ is based on explicit information on the asymptotic form of the functions $\mathcal{N}(\lambda)$ and $\mathcal{D}(\lambda)$. From Propositions 5.1 and 5.2 it is clear that they are entire functions and simple estimates obtained from the asymptotic form of $\mathcal{N}$ and $\mathcal{D}$ in the z variable (recall that $\lambda = i^n z^n$) will show that these functions have the same order $\sigma = 1/n$. We consider only the case μ odd and note that the case μ even can be handled following the discussion in (4.23)–(4.26). In this case $\omega_\mu = i$ and from (4.13)–(4.22),

$$\begin{aligned}\mathcal{N}(\lambda) &= \frac{z^{m^0-\mu(n-1)}e^{\omega z}\delta(z,\infty)}{\det\left([\omega_j^{i-1}]\right)} \\ &= [C]z^{m^0-\mu(n-1)}e^{\omega z}\left\{-[N_1]e^{2z\omega_\mu} + [N_{-1}]\right\}.\end{aligned}$$

Factor out N_1 and recall that from Proposition 4.2, part 4, $N_{-1}/N_1 = \exp(-2m^0\pi i/n)$; also factor out $\exp(-m^0\pi i/n)$, $\exp(\omega_\mu z)$ and recall from (4.15) and $\omega_\mu = -\omega_{\mu+1}$ that

$$\omega + \omega_\mu \equiv \widehat{\omega} = \sum_{j=2}^{\mu}\omega_{\mu+j}.$$

From this we obtain

$$\mathcal{N}(\lambda) = [-2iN_1C]z^{m^0-\mu(n-1)}e^{-m^0\pi i/n}e^{\widehat{\omega}z}\left[\sin(z + m^0\pi/n)\right],$$

where $|C| = n^{-\mu}$.

Hence for $|z|$ large

$$\begin{aligned}|\mathcal{N}(\lambda)| &= |[2N_1C]|\,|z|^{m^0-\mu(n-1)}|e^{-m^0\pi i/n}||e^{\widehat{\omega}z}|\,\left|\left[\sin(z + m^0\pi/n)\right]\right| \\ &= |[2N_1C]|\,|\lambda|^{(m^0-\mu(n-1))/n}|e^{\widehat{\omega}\lambda^{1/n}}|\left|\left[\sin(i\lambda^{1/n} + m^0\pi/n)\right]\right|,\end{aligned}$$

from which it is easy to see that for large modulus of λ

$$\limsup_{|\lambda|\to\infty}\left(\frac{\log\log\max_{0\le\theta<2\pi}|\mathcal{N}(|\lambda|e^{i\theta})|}{\log(|\lambda|)}\right) = \frac{1}{n},$$

which, with the definition of order of an entire function, shows that the order of $\mathcal{N}$ is $1/n$. A completely analogous argument shows that the order of $\mathcal{D}$ is also $1/n$ and this concludes the proof of Proposition 2.1.

Proof of Proposition 2.3. For μ odd, $z \in S_0 \cup S_{2n-1}$, we have

$$\arg z \in [-\pi/n, \pi/n)$$

and for μ even, $z \in S_0 \cup S_1$ with

$$\arg z \in [0, 2\pi/n).$$

Thus by the relation $\lambda = i^n z^n$, we have for $\lambda \in \mathbb{R}_+$,

$$z^\ell = \begin{cases} \lambda^{\ell/n} e^{-\ell\pi i/n}, & \text{for } \mu \text{ odd}, \\ \lambda^{\ell/n} & \text{for } \mu \text{ even}. \end{cases}$$

Then by the asymptotic expression for the transfer function given in (4.22) we have

$$\lim_{\lambda \to +\infty} \lambda^{\ell/n} \mathcal{G}_0(\lambda) \equiv \widehat{\tau} = \begin{cases} \tau_1 e^{\ell\pi i/n}, & \text{for } \mu \text{ odd}, \\ \tau_1 & \text{for } \mu \text{ even}. \end{cases}$$

Finally using Proposition 4.2, part 5, we know that $\lim_{\lambda \to +\infty} \lambda^{\ell/n} \mathcal{G}_0(\lambda)$ is a real number. □

Proof of Theorem 2.4. From the proof of Propositions 2.3 and 4.2, we see that the signum of the instantaneous gain $\widehat{\tau}$ is given by $(-1)^s$ with s defined in (2.21) and computed in Proposition 4.2. □

Proof of Proposition 2.5. Part 1 of Proposition 2.5 proceeds as follows. Since by Proposition 2.1 both $\mathcal{D}(\lambda)$ and $\mathcal{D}(\lambda)$ are entire function of order $1/n$, by the Hadamard factorization theorem, there exists integers p and q and constants C_1, C_2 so that

$$\mathcal{N}(\lambda) = \lambda^p C_1 \prod_{n=p+1}^{\infty} \left(1 - \frac{\lambda}{\xi_n}\right),$$

$$\mathcal{D}(\lambda) = \lambda^q C_2 \prod_{n=q+1}^{\infty} \left(1 - \frac{\lambda}{p_n}\right),$$

where $\{\xi_n\}$, $\{p_n\}$ are the nonzero zeros of $\mathcal{N}(\lambda)$ and $\mathcal{D}(\lambda)$, respectively, i.e., the nonzero open-loop zeros and poles ordered so that

$$|\xi_n| \leq |\xi_{n+1}|,$$

$$|p_n| \leq |p_{n+1}|.$$

By Theorem 3.2, there is an $N > 0$ so that for $n \geq N$ the poles and zeros are simple and

$$\xi_n < p_n < \xi_{n-1} < p_{n-1} < 0.$$

Hence for $k \geq N$,

$$\text{Res}_{\lambda = p_k} = \lim_{\lambda \to p_k} (\lambda - p_k) \frac{\mathcal{N}(\lambda)}{\mathcal{D}(\lambda)} = \mathcal{N}(p_k) \left[\lim_{\lambda \to p_k} \frac{\mathcal{D}(\lambda)}{(\lambda - p_k)} \right]^{-1}$$

and we can assert

$$\lim_{\lambda\to p_k}\frac{\mathcal{D}(\lambda)}{(\lambda-p_k)} = C_2 p_k^q \prod_{n=q+1}^{k-1}\left(1-\frac{p_k}{p_n}\right)\left(-\frac{1}{p_k}\right)\prod_{n=k+1}^{\infty}\left(1-\frac{p_k}{p_n}\right)$$
$$= C_2(-1)^{k-1}|p_k|^{q-1}\prod_{\substack{n>q\\ n\neq k}}\left|1-\frac{p_k}{p_n}\right|,$$

where we note that some finite number of the terms in the infinite product may correspond to complex poles, which must occur in conjugate pairs, so that the absolute value is still correct.

Again by the Hadamard factorization theorem we have

$$\mathcal{N}(p_k) = C_1 p_k^p \prod_{n=p+1}^{\infty}\left(1-\frac{p_k}{\xi_n}\right)$$
$$= C_1(-1)^{k-1}|p_k|^p \prod_{n=p+1}^{\infty}\left|1-\frac{p_k}{\xi_n}\right|.$$

Here again there may be a finite number of terms corresponding to complex zeros but they occur in conjugate pairs.

Now recalling that $\mathcal{G}_0(\lambda) = \mathcal{N}(\lambda)/\mathcal{D}(\lambda)$ is real, the above calculations show that C_1/C_2 is real. Thus we have that $\operatorname{Res}_{\lambda=p_k}\mathcal{G}_0(\lambda)$ is real and has the same sign as C_1/C_2 and we have established the first part of the proof.

For the second part of the proof recall that $\lambda = i^n z^n$ and note that for any constant a, $d[a]/dz = [0]$ and also for any pole λ_k with k large

$$e^{2\omega_\mu z_k} = \left[\frac{D_{-1}}{D_1}\right] = \left[e^{-2m\pi i/n}\right].$$

With this and the asymptotic form of the transfer function in (4.22), we have

$$\operatorname{Res}_{\lambda=p_k}\mathcal{G}_0(\lambda) = \lim_{\lambda\to p_k}(\lambda-p_k)\,\mathcal{G}_0(\lambda)$$

$$= \lim_{\lambda\to p_k}\frac{\left(-[N_1]e^{2\omega_\mu z}+[N_{-1}]\right)(i^n z^n - i^n z_k^n)/(z-z_k)}{z^\ell\left(-[D_1]e^{2\omega_\mu z}+[D_{-1}]\right)/(z-z_k)}$$

$$= \frac{\left(-[N_1]e^{2\omega_\mu z_k}+[N_{-1}]\right)ni^n z_k^{n-1}}{z_k^\ell\left(-2\omega_\mu[D_1]e^{2\omega_\mu z_k}\right)}$$

$$= \frac{ni^n z_k^{n-\ell-1}}{-2\omega_\mu}\,\frac{-[N_1]\left[e^{-2m\pi i/n}\right]+[N_{-1}]}{[D_{-1}]}$$

$$= \frac{ni^n z_k^{n-\ell-1}}{-2\omega_\mu}\,\frac{N_{-1}}{D_{-1}}\,\frac{-\left[\left[e^{-2m^0\pi i/n}\right]e^{-2m\pi i/n}\right]+[1]}{[1]}$$

$$= \frac{ni^n z_k^{n-\ell-1}\tau}{-\omega_\mu}\,\frac{\left[e^{\ell\pi i/n}\right]-\left[e^{-\ell\pi i/n}\right]}{[2]}\,e^{-\ell\pi i/n}$$

$$= -\frac{ni^{n+1}z_k^{n-\ell-1}\tau}{\omega_\mu}e^{-\ell\pi i/n}\sin\left(\frac{\ell\pi}{n}\right)[1].$$

When μ is odd, $\omega_\mu = i$ and this reduces to

$$-\frac{ni^{n+1}|z_k|^{n-\ell-1}|\tau|}{i} e^{(s+2\ell/n)\pi i} e^{-\ell\pi i/n} \sin\left(\frac{\ell\pi}{n}\right)[1],$$

where s is given (2.21) and when μ is even, $\omega_\mu = ie^{-\pi i/n}$ and

$$-\frac{ni^{n+1}|z_k|^{n-\ell-1}e^{(n-\ell-1)\pi i/n}|\tau|}{ie^{-\pi i/n}} e^{(s+2\ell/n)\pi i} e^{-\ell\pi i/n} \sin\left(\frac{\ell\pi}{n}\right)[1].$$

Simplifying these expressions, we arrive at the conclusion of part 2, namely,

$$\operatorname{Res}_{\lambda=p_k} \mathcal{G}_0(\lambda) = (-1)^s n|z_k|^{n-\ell-1}|\tau| \sin\left(\frac{\ell\pi}{n}\right)[1].$$

For part 3 of the proof we first show that for

$$x_j = j\pi + \frac{\pi}{2} - \frac{m\pi}{n},$$

we have

$$|\mathcal{G}_0(s_j)| \le \frac{3|\tau|}{|\xi_j|^\ell},$$

where $s_j = i^n \xi_j^n$ and $\xi_j = x_j + iy \in S_0 \cup S_{2n-1}$.

Recall the asymptotic expression for $\mathcal{G}_0(s)$ in the z-plane

$$\begin{aligned}
\mathcal{G}_0(s)|_{s=i^n z^n} &= \left.\frac{\mathcal{N}(s)}{\mathcal{D}(s)}\right|_{s=i^n z^n} \\
&= \frac{[-2iN_1 n^{-\mu}]z^{m^0-\mu(n-1)}e^{-m^0\pi i/n}e^{\hat{w}z}[\sin(z+m^0\pi/n)]}{[-2iD_1 n^{-\mu}]z^{m-\mu(n-1)}e^{-m\pi i/n}e^{\hat{w}z}[\sin(z+m\pi/n)]} \\
&= \left[\tau_1 e^{\frac{\ell\pi i}{n}}\right] \frac{[\sin(z+m^0\pi/n)]}{z^\ell[\sin(z+m\pi/n)]}.
\end{aligned}$$

Since

$$\begin{aligned}
\left|\left[\sin\left(\xi_j + \frac{m\pi}{n}\right)\right]\right| &= \left|\left[\sin\left(j\pi + \frac{\pi}{2} + iy\right)\right]\right| \\
&= \left|[(-1)^j \cos(iy)]\right| \\
&= \left|[(-1)^j \cosh(y)]\right| \\
&\ge \cosh(y) - \frac{1}{2},
\end{aligned}$$

$$\begin{aligned}
\left|\left[\sin\left(\xi_j + \frac{m^0\pi}{n}\right)\right]\right| &= \left|\left[\sin\left(j\pi + \frac{\pi}{2} + iy - \frac{\ell\pi}{n}\right) -\right]\right| \\
&= \left|[(-1)^j \cos(iy - \frac{\ell\pi}{n})]\right| \\
&\le \cosh(y) + \frac{1}{2};
\end{aligned}$$

thus

$$|\mathcal{G}_0(s_j)| \le \left| \frac{\tau_1 \left(\cosh(y) + \frac{1}{2}\right)}{\xi_j^{\ell} \left(\cosh(y) - \frac{1}{2}\right)} \right|$$

$$\le \frac{|\tau|}{|\xi_j|^{\ell}} \left(1 + \frac{1}{\cosh(y) - \frac{1}{2}}\right)$$

$$\le \frac{3|\tau|}{|\xi_j|^{\ell}}$$

for j large. The assertion is established.

Note that the impulse response function is given by

$$h(t) = \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} e^{st}\, \mathcal{G}_0(s)\, ds$$

$$= \lim_{R \to \infty} \frac{1}{2\pi i} \int_{\sigma - iR}^{\sigma + iR} e^{st}\, \mathcal{G}_0(s)\, ds$$

$$= \lim_{j \to \infty} \frac{1}{2\pi i} \int_{\sigma - ir_j}^{\sigma + ir_j} e^{st}\, \mathcal{G}_0(s)\, ds.$$

To see that the inverse Laplace transform of $\mathcal{G}_0$ can be computed using the method of residues, we let

$$V = \{s : \operatorname{Re} s = \sigma\}$$

be a vertical line such that all poles of $\mathcal{G}_0(s)$ lie to the left of V. Let

$$L = \{z \in S_0 \cup S_{2n-1} : \operatorname{Re} z = x\},$$
$$W = \{s = -z^n : z \in L, \operatorname{Re} s \le \sigma\},$$
$$l = \{z \in L : -z^n \in W\}.$$

Then W is a simple rectifiable curve in the λ-plane that intersects V at two points, is symmetric with respect to the real axis, and extends to the left of the vertical line V. It is clear that when $x \to \infty$, the curve W contains the whole half-plane left of the vertical line V.

What we need to show now is

$$\left| \int_W e^{st} \mathcal{G}_0(s) ds \right| \to 0$$

as $x \to \infty$.

Let $y_x > 0$ be the end point of l in S_0. Let σ_x be the intersection of W and V corresponding to y_x; then σ_x is in quadrant IV if we assume $\sigma > 0$. Let α be the angle between the negative imaginary axis and σ_x in the λ-plane. Then since for $z = x + iy \in l, s = -z^n \in W$,

$$z = \sqrt{x^2 + y^2} e^{i \arctan(y/x)},$$

$$s = (x^2 + y^2)^{\frac{n}{2}} e^{i(\pi + n \arctan(y/x))},$$

we know

$$\alpha = \arcsin \frac{\sigma}{(x^2 + y_x^2)^{n/2}} \leq \arcsin \frac{\sigma}{x^n}.$$

Let

$$\beta = \arcsin \frac{\sigma}{x^n}.$$

From

$$\frac{\pi}{2} - \beta \leq \pi + n \arctan \frac{y}{x} \leq \frac{3\pi}{2} + \beta$$

we see that

$$-\frac{\pi + 2\beta}{2n} \leq \arctan \frac{y}{x} \leq \frac{\pi + 2\beta}{2n},$$

which implies

$$-x \tan \frac{\pi + 2\beta}{2n} \leq y \leq x \tan \frac{\pi + 2\beta}{2n}.$$

Now, for $C = 3|\tau|$, we have

$$\left| \int_W e^{st} \mathcal{G}_0(s) ds \right| = \left| \int_l e^{-z^n t} \left[\tau_1 e^{\frac{\ell \pi i}{n}} \right] \frac{[\sin(z + m^0 \pi/n)]}{z^\ell [\sin(z + m\pi/n)]} i^n n z^{n-1} dz \right|$$

$$\leq n \int_{-x \tan \frac{\pi+2\beta}{2n}}^{x \tan \frac{\pi+2\beta}{2n}} e^{-(x^2+y^2)^{n/2} t \cos(n \arctan(y/x))} \frac{C}{(x^2 + y^2)^{\ell/2}} (x^2 + y^2)^{(n-1)/2} dy$$

$$= 2Cn \int_0^{x \tan \frac{\pi+2\beta}{2n}} e^{-(x^2+y^2)^{n/2} t \cos(n \arctan(y/x))} (x^2 + y^2)^{(n-\ell-1)/2} dy.$$

Let $y = x \tan u$. Then

$$x^2 + y^2 = x^2 \sec^2 u, \qquad dy = x \sec^2 u du.$$

Therefore

$$\left| \int_W e^{st} \mathcal{G}_0(s) ds \right| \leq 2Cnx^{n-\ell} \int_0^{\frac{\pi+2\beta}{2n}} e^{-x^n t (\sec^n u) \cos(nu)} \sec^{n-\ell+1} u du$$

$$\equiv I_1 + I_2,$$

where

$$I_1 \equiv 2Cnx^{n-\ell} \int_0^{\frac{\pi}{2n}} e^{-x^n t (\sec^n u) \cos(nu)} \sec^{n-\ell+1} u du$$

$$\leq 2Cnx^{n-\ell} \int_0^{\frac{\pi}{2n}} e^{-x^n t (\sec^n 0) \cos(nu)} \left(\sec^{n-\ell+1} \frac{\pi}{2n} \right) du$$

$$\leq 2Cnx^{n-\ell}\left(\sec^{n-\ell+1}\frac{\pi}{2n}\right)\int_0^{\frac{\pi}{2n}} e^{x^n t\frac{2}{\pi}(nu-\frac{\pi}{2})}du$$

$$= C\pi x^{-\ell}t^{-1}\left(\sec^{n-\ell+1}\frac{\pi}{2n}\right)e^{x^n t\frac{2}{\pi}(nu-\frac{\pi}{2})}\Big|_0^{\frac{\pi}{2n}}$$

$$= C\pi x^{-\ell}t^{-1}\left(\sec^{n-\ell+1}\frac{\pi}{2n}\right)\left(1-e^{-x^n t}\right)\to 0$$

as $x\to\infty$, and

$$I_2 \equiv 2Cnx^{n-\ell}\int_{\frac{\pi}{2n}}^{\frac{\pi+2\beta}{2n}} e^{-x^n t(\sec^n u)\cos(nu)}\left(\sec^{n-\ell+1}u\right)du$$

$$\leq 2Cnx^{n-\ell}\int_{\frac{\pi}{2n}}^{\frac{\pi+2\beta}{2n}} e^{-x^n t\left(\sec^n\frac{\pi+2\beta}{2n}\right)\cos\frac{\pi+2\beta}{2}}\left(\sec^{n-\ell+1}\frac{\pi+2\beta}{2n}\right)du$$

$$= 2Cnx^{n-\ell}\left(\sec^{n-\ell+1}\frac{\pi+2\beta}{2n}\right)e^{x^n t\left(\sec^n\frac{\pi+2\beta}{2n}\right)\sin\beta}\left(\frac{\pi+2\beta}{2n}-\frac{\pi}{2n}\right)$$

$$= 2Cx^{-\ell}\left(\sec^{n-\ell+1}\frac{\pi+2\beta}{2n}\right)e^{t\sec^n\frac{\pi+2\beta}{2n}x^n\sin\beta}x^n\beta$$

$$= 2Cx^{-\ell}\left(\sec^{n-\ell+1}\frac{\pi+2\beta}{2n}\right)e^{t\sec^n\frac{\pi+2\beta}{2n}\sigma}\sigma\frac{\beta}{\sin\beta}\to 0$$

as $x\to\infty$. Here we have used the fact that

$$\sin\beta = \frac{\sigma}{x^n}$$

and

$$\lim_{\beta\to 0}\frac{\beta}{\sin\beta} = 1.$$

Also in the second inequality in the expression for I_1 we have used the estimate $-\cos(nu)\leq(2/\pi)(nu-\pi/2)$ for $u\in[0,\pi/(2n)]$. This inequality follows from the well-known estimate $\sin(x)\geq(2/\pi)x$ for $x\in[0,\pi/2]$. Namely, let $y=nu$ and consider $-\cos(y)\leq(2/\pi)(y-\pi/2)$, which is equivalent to $\cos(y)\leq(2/\pi)(\pi/2-y)$. Let $x=\pi/2-y$, then $x\in[0,\pi/2]$, $\sin(x)=\sin(\pi/2-y)=\cos(y)$ and the result follows.

Thus for $t>0$

$$\lim_{j\to\infty}\left|\int_W e^{st}\mathcal{G}_0(s)\,ds\right| = 0$$

and we are justified in evaluating the integral via residues. □

Proof of Proposition 2.2. The proof is similar to those given in Propositions 2.3 and 2.5 and relies heavily on the results of Proposition 2.5. From Theorem 3.2 we know that the poles of large modulus of $\mathcal{G}_0$ are all negative and exactly as in the proof of part 3 of Proposition 2.5 we can show that there exists a constant M so that

$$|\mathcal{G}_0(\lambda)|\leq\frac{M}{|\lambda|^{\ell/n}}$$

for all $\lambda \in \mathbb{C}_+^a$ for some $a \geq 0$ where a is chosen, by Theorem 3.2, so that $\mathcal{G}_0$ is analytic and has no zeros for $\mathrm{Re}(\lambda) > a$. Then clearly, $\mathcal{G}_0 \in H^\infty(\mathbb{C}_+^a)$ and $\mathcal{G}_0$ is strictly proper since it goes to zero as λ goes to infinity in this right half-plane.

Choosing N by Theorem 3.2 large enough so that for $n > N$ all the open-loop poles and zeros interlace and the closed-loop poles are real and simple. Then defining $h_1(t)$ and $h_2(t)$ through the expressions (cf. Proposition 2.5, part 3)

$$h_1(t) = \sum_{j=1}^{N} \mathrm{Res}_{\lambda=p_k} \left(\mathcal{G}_0(\lambda)e^{\lambda t}\right),$$

$$h_2(t) = C \sum_{j=N+1}^{\infty} |\lambda_j|^{(n-\ell-1)/n} e^{\lambda_j t}[1]$$

we have that the impulse response function is given by $h(t) = h_1(t) + h_2(t)$ with h_1 a continuous function of $t \in [0, \infty)$ and $e^{-at}h_1(t) \in L^p(0, \infty)$ for every $p \geq 1$. Also from the asymptotic distribution of the poles $\lambda_j \sim (j\pi)^n$ for large j we see that h_2 is integrable near $t = \infty$. The only real concern is whether the infinite sum is in L^1_{loc} at $t = 0$. It is obvious that this Dirichlet series is always singular at $t = 0$ but nevertheless we show that it is integrable on all of $(0, \infty)$. From the above definition of h_2 there is a constant $\widetilde{C}$ such that

$$|h_2(t)| \leq \widetilde{C} \sum_{j=N+1}^{\infty} |\lambda_j|^{(n-\ell-1)/n} e^{\lambda_j t}.$$

Thus for every $b > 0$, by Tonelli's Theorem,

$$\begin{aligned}
\int_0^b |h_2(t)|\, dt &\leq \widetilde{C} \sum_{j=N+1}^{\infty} |\lambda_j|^{(n-\ell-1)/n} \int_0^b e^{\lambda_j t}\, dt \\
&= \widetilde{C} \sum_{j=N+1}^{\infty} |\lambda_j|^{(n-\ell-1)/n} \frac{(1 - e^{\lambda_j b})}{\lambda_j} \\
&\leq \widetilde{C} \sum_{j=N+1}^{\infty} |\lambda_j|^{-(\ell+1)/n} < \infty
\end{aligned}$$

since by Assumption 2.1 $0 < \ell = m_1 - m_0 \leq (n-1)$ (recall that $0 \leq m_0 < m_1 \leq (n-1)$) and the final sum is finite from the asymptotic form of the open-loop poles.

Now we show that the map given by $y(t) = (h * u)(t)$ defines a bounded map on

$$U_a = \{f : (0, \infty) \to \mathbb{C} \,:\, \exp(-a\cdot)f(\cdot) \in L_2(0, \infty),\ a \in \mathbb{R}\}.$$

First note that

$$\begin{aligned}
\widehat{(e^{-a\cdot}y(\cdot))}(s) &= \widehat{y}(s + a) \\
&= \mathcal{G}_0(s + a)\widehat{u}(s + a) \\
&= \widehat{(e^{-a\cdot}h(\cdot))}(s)\widehat{(e^{-a\cdot}u(\cdot))}(s) \\
&= \widehat{(e^{-a\cdot}h(\cdot)) * (e^{-a\cdot}u(\cdot))}(s).
\end{aligned}$$

Hence

$$\begin{aligned} e^{-at}y(t) &= \left(e^{-a\cdot}h(\cdot)\right) * \left(e^{-a\cdot}u(\cdot)\right)(t) \\ &= \int_0^\infty e^{-a(t-s)}h(t-s)e^{-as}u(s)\,ds \\ &= e^{-at}\int_0^\infty h(t-s)u(s)\,ds \\ &= e^{-at}(h*u)(t) \end{aligned}$$

and we have $y(t) = (h*u)(t)$ and for $u \in U_a$

$$\begin{aligned} \int_0^\infty \left|e^{-at}y(t)\right|^2 dt &= \int_0^\infty \left|(e^{-a\cdot}h) * (e^{-a\cdot}u)\right|^2 dt \\ &\le \|e^{-a\cdot}h\|_1^2 \, \|e^{-a\cdot}u\|_2^2 < \infty. \qquad \square \end{aligned}$$

Proof of Proposition 2.6. Let $p = \max_{1\le j\le n}\{m_j\}$ and we seek a polynomial in the form

$$b(x) = \sum_{j=0}^{2p+1} a_j x^j.$$

For any given set of boundary operators $\{\mathcal{W}_j\}_{j=1}^n$ in (2.4) we can rewrite the operators in the form $(i = 1, \dots, n)$

$$\mathcal{W}_i(f) \equiv \sum_{j=0}^{p} \alpha_{ij} f^{(j)}(0) + \sum_{j=0}^{p} \beta_{ij} f^{(j)}(1)$$

by introducing zero coefficients as necessary.

Then the problem reduces to showing that given $2p+2$ pieces of data $\{\alpha_j,\ \beta_j\}_{j=0}^p$, we can find constants $\{a_j\}_{j=0}^{2p+1}$ such that the polynomial b satisfies

$$b^{(j)}(0) = \alpha_j, \quad b^{(j)}(1) = \beta_j, \qquad j = 0, 1, \dots, p.$$

Now since

$$b^{(k)}(x) = \sum_{j=k}^{2p+1} a_j[j(j-1)\cdots(j-k+1)]x^{(j-k)}, \qquad k = 0, 1, \dots, 2p+1,$$

we have

$$b^{(k)}(0) = a_k k!, \quad b^{(k)}(1) = \sum_{j=k}^{2p+1} a_j A_j^k, \qquad k = 0, 1, \dots, 2p+1,$$

where

$$A_j^k = \frac{j!}{(j-k)!} = j(j-1)\cdots(j-k+1), \qquad k \le j.$$

Hence

$$a_j = \frac{b^{(j)}(0)}{j!} = \frac{\alpha_j}{j!},$$

$$\beta_k = b^{(k)}(1) = \sum_{j=k}^{2p+1} a_j A_j^k = \sum_{j=k}^{p} a_j A_j^k + \sum_{j=p+1}^{2p+1} a_j A_j^k$$

$$= \sum_{j=k}^{p} \frac{\alpha_j}{j!} + \sum_{j=p+1}^{2p+1} a_j A_j^k,$$

which implies

$$\sum_{j=p+1}^{2p+1} a_j A_j^k = \beta_k - \sum_{j=k}^{p} \frac{\alpha_j}{(j-k)!} \equiv \gamma_k, \qquad k = 0, 1, \ldots, p.$$

Thus the remainder of the proof reduces to showing the solvability of the $(p+1)\times(p+1)$ linear system

$$Aa = \gamma,$$

where

$$A_{ij} = A_{p+j}^i, \qquad i = 0, 1, \ldots, p, \quad j = 1, 2, \ldots (p+1),$$

$$a = [a_{p+1}, a_{p+2}, \ldots, a_{2p+1}]^T, \qquad \gamma = [\gamma_0, \gamma_1, \ldots, \gamma_p]^T.$$

Thus we show that the determinant of A is not zero. To this end, we claim that for any integers ℓ and p, we have

$$I_{\ell,p} \equiv \det\left(\{A_{\ell+j}^i\}_{i=0,j=1}^{p-1,p}\right) = \prod_{k=1}^{p-1}(k!).$$

To establish this result we recall that $A_j^k = (j!/(j-k)!)$ so that the determinant $I_{\ell,p}$ can be written as

$$I_{\ell,p} = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ (\ell+1) & (\ell+2) & \cdots & (\ell+p) \\ (\ell+1)\ell & (\ell+2)(\ell+1) & \cdots & (\ell+p)(\ell+p-1) \\ \vdots & \vdots & \ddots & \vdots \\ (\ell+1)\cdots(\ell-p+4) & (\ell+2)\cdots(\ell-p+5) & \cdots & (\ell+p)\cdots(\ell+3) \\ (\ell+1)\cdots(\ell-p+3) & (\ell+2)\cdots(\ell-p+4) & \cdots & (\ell+p)\cdots(\ell+2) \end{vmatrix}.$$

But by elementary column reduction used to reduce the first row to a 1 in the first position and zeros elsewhere and then on expanding by the first row we have

$$(p-1)! \begin{vmatrix} 1 & 1 & \cdots & 1 \\ (\ell+1) & (\ell+2) & \cdots & (\ell+p-1) \\ (\ell+1)\ell & (\ell+2)(\ell+1) & \cdots & (\ell+p)(\ell+p-1) \\ \vdots & \vdots & \ddots & \vdots \\ (\ell+1)\cdots(\ell-p+5) & (\ell+2)\cdots(\ell-p+6) & \cdots & (\ell+p-1)\cdots(\ell+3) \\ (\ell+1)\cdots(\ell-p+4) & (\ell+2)\cdots(\ell-p+5) & \cdots & (\ell+p-1)\cdots(\ell+2) \end{vmatrix}$$

and hence

$$I_{\ell,p} = (p-1)!\, I_{\ell,(p-1)}.$$

Thus by induction we have

$$I_{\ell,p} = (m-1)!(m-2)!\cdots 2!\, I_{\ell 2}.$$

Now

$$I_{\ell,2} = \begin{vmatrix} 1 & 1 \\ \ell+1 & \ell+2 \end{vmatrix} = 1 = 1!$$

and we have

$$I_{\ell,p} = \prod_{k=1}^{p-1} (k!)$$

and this completes the proof.

We note that in the case where the input occurs in the highest order, i.e., $m_1 > m_i$ for $i = 2, \ldots, n$, we may choose

$$b(x) = \frac{1}{\alpha_1 m_1!} x^{m_1}(1-x)^{m_1+1},$$

which satisfies

$$\mathcal{W}_1(b) = 1, \quad \mathcal{W}_i(b) = 0, \qquad i = 2, \ldots, n.$$

This can be checked by direct verification using the fact that

$$b^{(k)}(0) = \begin{cases} 0, & k < m_1, \\ \frac{1}{\alpha_1}, & k = m_1, \end{cases} \qquad b^{(k)}(1) = 0,\ k \le m_1. \qquad \square$$

Proof of Theorem 3.4. From the asymptotic form of the open-loop poles and zeros and the Hadamard factorization theorem, there exist integers p and q and constants C_1 and C_2 such that

$$\mathcal{N}(\lambda) = \lambda^p C_1 \prod_{n=p+1}^{\infty} \left(1 - \frac{\lambda}{z_n}\right),$$

$$\mathcal{D}(\lambda) = \lambda^q C_2 \prod_{n=q+1}^{\infty} \left(1 - \frac{\lambda}{p_n}\right),$$

where $\{z_n\}$, $\{p_n\}$ are the nonzero zeros of $\mathcal{N}(\lambda)$ and $\mathcal{D}(\lambda)$, respectively, i.e., the nonzero open-loop zeros and poles ordered so that

$$|z_n| \le |z_{n+1}|,$$

$$|p_n| \le |p_{n+1}|.$$

Now by the proof of the main theorem, every open-loop pole moves continuously to an open-loop zero as the gain parameter varies from 0 to $\pm\infty$ and z_n, p_n are real and interlace for n large. Thus there is an N and $\Lambda_N > 0$ so that for $n > N$

$$z_n < p_n < -\Lambda_N$$

and

$$\{z_n\}_{n=p+1}^{N-1} \cup \{p_n\}_{n=q+1}^{N-1} \subset B(0, \Lambda_N).$$

Now for $z_N < a < p_N$ we have

$$\arg\left(\prod_{n=N}^{\infty}(1 - a/z_n)\right) = \arg\left(\prod_{n=N+1}^{\infty}(1 - a/p_n)\right) = 0.$$

Throughout the remainder of the proof it will be assumed that all equalities related to arguments of complex numbers are congruent modulo 2π. Further we note that if $\operatorname{Im}(z_0) \neq 0$ and a is real, we have

$$\arg(1 - a/z_0) + \arg(1 - a/\overline{z_0}) = 0.$$

The return difference equation can be written as

$$k\frac{\mathcal{N}(\lambda)}{\mathcal{D}(\lambda)} = -1,$$

so that a number λ being on the root locus implies

$$\arg\left(k\frac{\mathcal{N}(\lambda)}{\mathcal{D}(\lambda)}\right) = \pi.$$

Now we have

$$\begin{aligned}
\arg\left(k\frac{\mathcal{N}(a)}{\mathcal{D}(a)}\right) &= \arg\left(ka^{p-q}\frac{C_1}{C_2}\right) + \arg\left(\prod_{n=p+1}^{\infty}(1 - a/z_n)\right) - \arg\left(\prod_{n=q+1}^{\infty}(1 - a/p_n)\right) \\
&= \arg\left(ka^{p-q}\frac{C_1}{C_2}\right) + \sum_{n=p+1}^{N-1}\arg(1 - a/z_n) + \sum_{n=q+1}^{N}\arg(1 - a/p_n) \\
&= \arg\left(ka^{p-q}\frac{C_1}{C_2}\right) + \sum_{n=p+1}^{N-1}\{\arg(a - z_n) + \arg(-z_n)\} \\
&\quad + \sum_{n=q+1}^{N}\{\arg(a - p_n) + \arg(-p_n)\} \\
&= \arg\left(ka^{p-q}\frac{C_1}{C_2}\right) + (N - 1 - p)\pi + (N - q)\pi + r\pi \\
&= \arg\left(k(-1)^r a^{p-q}\frac{C_1}{C_2}\right) + (p + q + 1)\pi,
\end{aligned}$$

where r is the number of positive zeros and poles.

By the fact that

$$z_N < a < p_N,$$

we know that a is on the root locus; hence

$$\arg\left(k(-1)^r a^{p-q}\frac{C_1}{C_2}\right) = \begin{cases} 0 & \text{if } p+q \text{ even,} \\ \pi & \text{if } p+q \text{ odd.} \end{cases}$$

Therefore, in general, for $a \in \mathbb{R}$

$$\arg\left(k(-1)^r a^{p-q}\frac{C_1}{C_2}\right) = \begin{cases} 0 & \text{if } p+q \text{ even,} \\ 0 & \text{if } p+q \text{ odd,} \qquad a > 0, \\ \pi & \text{if } p+q \text{ odd,} \qquad a < 0. \end{cases}$$

With this relation we are able to finish the proof.

Let S be the number of nonzero zeros and poles that lie on the right of $a > p_N$. Then we have

$$\begin{aligned} \arg\left(k\frac{\mathcal{N}(a)}{\mathcal{D}(a)}\right) &= \arg\left(ka^{p-q}\frac{C_1}{C_2}\right) + \sum_{n=p+1}^{N-1}\{\arg(a-z_n) + \arg(-z_n)\} \\ &\quad + \sum_{n=q+1}^{N-1}\{\arg(a-p_n) + \arg(-p_n)\} \\ &= \arg\left(ka^{p-q}\frac{C_1}{C_2}\right) + S\pi + r\pi \\ &= \arg\left(k(-1)^r a^{p-q}\frac{C_1}{C_2}\right) + S\pi. \end{aligned}$$

Thus a being on the root locus implies the following.

1. If $(p+q)$ is even, we have S odd. And if $a < 0$, then the number of zeros and poles on the right of a is $(S+p-q)$ or $(S-p+q)$ which is odd; if $a > 0$, then the number of zeros and poles is S which is odd again.

2. If $(p+q)$ is odd, we have

$$S \text{ is } \begin{cases} \text{odd} & \text{when } a > 0, \\ \text{even} & \text{when } a < 0. \end{cases}$$

Since the number of zeros and poles on the right of a is S when $a > 0$ and $(S+p-q)$ or $(S-p+q)$ when $a < 0$, we know that the number is always odd. □

6. Proof of Proposition 4.2. To establish the proposition, we introduce notation similar to that found in [26]. Namely, we let

$$x_j = \exp\left(\frac{2\pi i}{n}m_j\right), \quad y_j = \exp\left(\frac{2\pi i}{n}m_{\mu+j}\right), \qquad j = 1, \ldots, \mu$$

and

$$X(t) = \operatorname{diag}(x_1^t, \ldots, x_\mu^t),$$

$$Y(t) = \operatorname{diag}(y_1^t, \ldots, y_\mu^t),$$

$$V(x) = \left[x_i^{(j-1)}\right]_{i=1,j=1}^{\mu,\mu}, \qquad V(y) = \left[y_i^{(j-1)}\right]_{i=1,j=1}^{\mu,\mu},$$

$$E(k_1, k_2, \ldots, k_\mu) = [e_{k_1}, e_{k_2}, \ldots, e_{k_\mu}],$$

$$I_1 = [e_\mu, e_{\mu-1}, \ldots, e_1],$$

where e_{k_j} is the k_jth standard unit basis vector in $\mathbb{R}^\mu$. For μ odd, let

$$E_o = E\left(\frac{\mu-1}{2}+1, \frac{\mu-1}{2}+2, \frac{\mu-1}{2}, \frac{\mu-1}{2}+3, \ldots, \mu, 1\right),$$

$$F_o = E\left(\frac{\mu-1}{2}, \frac{\mu-1}{2}+1, \frac{\mu-1}{2}-1, \frac{\mu-1}{2}+2, \ldots, 1, \mu-1, \mu\right),$$

then we have

$$[U_1, \ldots, U_\mu] = X\left(\frac{\mu}{2}\right) V(x) E_o,$$

$$[V_{\mu+1}, \ldots, V_n] = Y\left(\frac{3\mu}{2}\right) V(y) E_o I_1,$$

$$[U_1, \ldots, U_{\mu-1}, U_{\mu+1}] = X\left(\frac{\mu}{2}+1\right) V(x) F_o,$$

$$[V_\mu, V_{\mu+2}, \ldots, V_n] = Y\left(\frac{3\mu}{2}+1\right) V(y) F_o I_1.$$

For μ even, let

$$E_e = E\left(\frac{\mu}{2}, \frac{\mu}{2}+1, \frac{\mu}{2}-1, \frac{\mu}{2}+2, \ldots, 1, \mu\right),$$

$$F_e = E\left(\frac{\mu}{2}+1, \frac{\mu}{2}+2, \frac{\mu}{2}, \frac{\mu}{2}+3, \ldots, \mu, 2, 1\right),$$

and we have

$$[U_1, \ldots, U_\mu] = X\left(\frac{\mu-1}{2}\right) V(x) I_1 E_e,$$

$$[V_{\mu+1}, \ldots, V_n] = Y\left(-\frac{\mu+1}{2}\right) V(y) I_1 E_e I_1,$$

$$[U_1, \ldots, U_{\mu-1}, U_{\mu+1}] = X\left(\frac{\mu+1}{2}\right) V(x) I_1 F_e,$$

$$[V_\mu, V_{\mu+2}, \ldots, V_n] = Y\left(-\frac{\mu-1}{2}\right) V(y) I_1 F_e I_1.$$

It is easy to establish that

$$|E_o| = \prod_{j=1}^{\frac{\mu-1}{2}} (-1)^{2j} = 1, \quad |F_o| = \prod_{j=1}^{\frac{\mu-1}{2}-1} (-1)^{2j} = 1,$$

$$|E_e| = \prod_{j=1}^{\frac{\mu}{2}-1} (-1)^{2j} = 1, \quad |F_e| = (-1)^{\mu-1} \prod_{j=1}^{\frac{\mu}{2}-1} (-1)^{2j} = -1,$$

$$|I_1| = \prod_{j=1}^{\mu-1} (-1)^j = (-1)^{\mu(\mu-1)/2}$$

which now implies that

$$D_{-1} = \begin{cases} \prod_{j=1}^{\mu} (\alpha_j \beta_j) (-1)^{\mu(\mu-1)/2} \prod_{j=1}^{\mu} \left(x_j^{\mu/2} y_j^{3\mu/2}\right) |V(x)|\, |V(y)|, & \mu \text{ odd}, \\ \prod_{j=1}^{\mu} (\alpha_j \beta_j) (-1)^{\mu(\mu-1)/2} \prod_{j=1}^{\mu} \left(x_j^{(\mu-1)/2} y_j^{-(\mu+1)/2}\right) |V(x)|\, |V(y)|, & \mu \text{ even}, \end{cases}$$

$$D_1 = \begin{cases} \prod_{j=1}^{\mu} (\alpha_j \beta_j) (-1)^{\mu(\mu-1)/2} \prod_{j=1}^{\mu} \left(x_j^{(\mu/2)+1} y_j^{(3\mu/2)+1}\right) |V(x)|\, |V(y)|, & \mu \text{ odd}, \\ \prod_{j=1}^{\mu} (\alpha_j \beta_j) (-1)^{\mu(\mu-1)/2} \prod_{j=1}^{\mu} \left(x_j^{(\mu+1)/2} y_j^{-(\mu-1)/2}\right) |V(x)|\, |V(y)|, & \mu \text{ even}, \end{cases}$$

and hence

$$\frac{D_{-1}}{D_1} = \begin{cases} \prod_{j=1}^{\mu} (x_j y_j)^{-1} = \exp\left(-\frac{2m\pi i}{n}\right), & \mu \text{ odd}, \\ \prod_{j=1}^{\mu} (x_j y_j)^{-1} = \exp\left(-\frac{2m\pi i}{n}\right), & \mu \text{ even}. \end{cases}$$

Similarly, for

$$x_0 = \exp\left(\frac{2\pi i}{n} m_0\right)$$

and

$$V^0(x) = \left[x_i^{(j-1)}\right]_{i=0,2,j=1}^{\mu,\mu},$$

we have

$$N_{-1} = \begin{cases} \alpha_0 \beta_1 \prod_{j=2}^{\mu} (\alpha_j \beta_j) (-1)^{\mu(\mu-1)/2} \\ \quad \times \prod_{j=0,2}^{\mu} \left(x_j^{\mu/2}\right) \prod_{j=1}^{\mu} \left(y_j^{3\mu/2}\right) |V^0(x)|\, |V(y)|, & \mu \text{ odd}, \\ \alpha_0 \beta_1 \prod_{j=2}^{\mu} (\alpha_j \beta_j) (-1)^{\mu(\mu-1)/2} \\ \quad \times \prod_{j=0,2}^{\mu} \left(x_j^{(\mu-1)/2}\right) \prod_{j=1}^{\mu} \left(y_j^{-(\mu+1)/2}\right) |V^0(x)|\, |V(y)|, & \mu \text{ even}, \end{cases}$$

$$N_1 = \begin{cases} \alpha_0\beta_1 \prod_{j=2}^{\mu}(\alpha_j\beta_j)(-1)^{\mu(\mu-1)/2} \\ \times \prod_{j=0,2}^{\mu}\left(x_j^{(\mu/2)+1}\right)\prod_{j=1}^{\mu}\left(y_j^{(3\mu/2)+1}\right)|V^0(x)|\,|V(y)|, & \mu \text{ odd}, \\ \alpha_0\beta_1 \prod_{j=2}^{\mu}(\alpha_j\beta_j)(-1)^{\mu(\mu-1)/2} \\ \times \prod_{j=0,2}^{\mu}\left(x_j^{(\mu+1)/2}\right)\prod_{j=1}^{\mu}\left(y_j^{-(\mu-1)/2}\right)|V^0(x)|\,|V(y)|, & \mu \text{ even}, \end{cases}$$

and hence

$$\frac{N_{-1}}{N_1} = \begin{cases} \prod_{j=0,2}^{\mu}(x_j)^{-1}\prod_{j=1}^{\mu}(y_j)^{-1} \\ \prod_{j=0,2}^{\mu}(x_j)^{-1}\prod_{j=1}^{\mu}(y_j)^{-1} \end{cases} = \begin{cases} \exp\left(-\frac{2m^0\pi i}{n}\right), & \mu \text{ odd}, \\ \exp\left(-\frac{2m^0\pi i}{n}\right), & \mu \text{ even}. \end{cases}$$

Furthermore, if we define

$$\gamma = \prod_{j=2}^{\mu}\frac{(x_j - x_0)}{(x_j - x_1)}$$

then

$$\tau = \frac{N_{-1}}{D_{-1}} = \begin{cases} \frac{\alpha_0}{\alpha_1}\frac{x_0^{\mu/2}}{x_1^{\mu/2}}\gamma = \frac{\alpha_0}{\alpha_1}\exp(-\frac{\pi i\ell}{2})\gamma, & \mu \text{ odd}, \\ \frac{\alpha_0}{\alpha_1}\frac{x_0^{(\mu-1)/2}}{x_1^{(\mu-1)/2}}\gamma = \frac{\alpha_0}{\alpha_1}\exp(-\frac{(\mu-1)\pi i\ell}{n})\gamma, & \mu \text{ even}, \end{cases} \tag{6.47}$$

$$\tau_1 = \frac{N_1}{D_1} = \begin{cases} \frac{\alpha_0}{\alpha_1}\frac{x_0^{\mu/2+1}}{x_1^{\mu/2+1}}\gamma = \frac{\alpha_0}{\alpha_1}\exp(-\frac{(\mu+2)\pi i\ell}{n})\gamma, & \mu \text{ odd}, \\ \frac{\alpha_0}{\alpha_1}\frac{x_0^{(\mu+1)/2}}{x_1^{(\mu+1)/2}}\gamma = \frac{\alpha_0}{\alpha_1}\exp(-\frac{(\mu+1)\pi i\ell}{n})\gamma, & \mu \text{ even}. \end{cases} \tag{6.48}$$

Using the facts that

$$x_j^n = y_j^n = 1, \quad \overline{x_j} = x_j^{-1}, \quad \overline{y_j} = y_j^{-1}, \qquad j = 1, \ldots, n,$$

we have

$$\overline{V(x)} = X(1-\mu)V(x)I_1, \qquad \overline{X(t)} = X(-t) = X(2\mu - t).$$

Hence

$$\overline{D_{-1}} = \begin{cases} \prod_{j=1}^{\mu}(\alpha_j\beta_j)(-1)^{\mu(\mu-1)/2}\prod_{j=1}^{\mu}\left(x_j^{1+(\mu/2)}y_j^{1+(3\mu/2)}\right)|V(x)|\,|V(y)|, \\ \prod_{j=1}^{\mu}(\alpha_j\beta_j)(-1)^{\mu(\mu-1)/2}\prod_{j=1}^{\mu}\left(x_j^{(\mu+3)/2}y_j^{(3\mu+3)/2}\right)|V(x)|\,|V(y)| \end{cases}$$

$$= \begin{cases} D_{-1}\prod_{j=1}^{\mu}(x_jy_j), & \mu \text{ odd}, \\ D_{-1}\prod_{j=1}^{\mu}(x_jy_j)^2, & \mu \text{ even}, \end{cases}$$

$$\overline{D_1} = \begin{cases} \prod_{j=1}^{\mu}(\alpha_j\beta_j)(-1)^{\mu(\mu-1)/2}\prod_{j=1}^{\mu}\left(x_j^{\mu/2}y_j^{3\mu/2}\right)|V(x)|\,|V(y)|, \\ \overline{D_{-1}}\exp\left\{-\frac{2m\pi i}{n}\right\} \end{cases}$$

$$= \begin{cases} D_1\prod_{j=1}^{\mu}(x_jy_j)^{-1} = D_{-1}, & \mu \text{ odd}, \\ D_{-1}\exp\left\{\frac{2m\pi i}{n}\right\} = D_1, & \mu \text{ even}, \end{cases}$$

and a similar computation shows

$$\overline{N_{-1}} = \begin{cases} N_{-1}\prod_{j=0,2}^{\mu}(x_j)\prod_{j=1}^{\mu}(y_j), & \mu \text{ odd}, \\ N_{-1}\prod_{j=0,2}^{\mu}(x_j)^2\prod_{j=1}^{\mu}(y_j)^2, & \mu \text{ even}, \end{cases}$$

$$\overline{N_1} = \begin{cases} N_1\prod_{j=0,2}^{\mu}(x_j^{-1})\prod_{j=1}^{\mu}(y_j^{-1}) \\ N_{-1}\exp\left\{\frac{2m^0\pi i}{n}\right\} \end{cases} = \begin{cases} N_{-1}, & \mu \text{ odd}, \\ N_1, & \mu \text{ even}. \end{cases}$$

To establish 5(b), let

$$T(z) = \frac{z - x_0}{z - x_1}.$$

Then T is a Möbius transformation with

$$T(x_0) = 0, \qquad T(x_1) = \infty.$$

Hence T maps the unit circle to a straight line L passing through the origin. Now $x_j = \exp(2\pi i m_j/n)$, $j = 1, \ldots, \mu$; therefore $T(x_j)$ is on L for each j.

Since $\arg(1 - e^{it}) = (1/2)(t - \pi)$ for $0 < t < 2\pi$,

$$\arg(T(1)) = \arg(1 - x_0) - \arg(1 - x_1) = -\frac{\pi\ell}{n}.$$

Hence by the fact that $T(x_0) = 0$,

$$\arg(T(x_j)) = \begin{cases} \pi - \frac{\pi\ell}{n}, & 2 \le j \le \ell_0 - 1, \\ -\frac{\pi\ell}{n}, & \ell_0 \le j \le \mu. \end{cases}$$

And then

$$\arg(\gamma) = (\ell_0 - 2)(\pi - \pi\ell/n) + (\mu - \ell_0 + 1)(-\pi\ell/n) = (\ell_0 - 2)\pi - (\mu - 1)(\pi\ell/n)$$

so that finally

$$\arg(\tau) = \begin{cases} \arg(\alpha_0/\alpha_1) - \frac{\pi\ell}{2} + (\ell_0 - 2)\pi - (\mu - 1)\frac{\pi\ell}{n}, \\ \arg(\alpha_0/\alpha_1) - \frac{\mu-1}{n}\pi\ell + (\ell_0 - 2)\pi - (\mu - 1)\frac{\pi\ell}{n}, \end{cases}$$

$$= \begin{cases} \arg(\alpha_0/\alpha_1) + (\ell_0 - \ell - 2)\pi + \frac{\pi\ell}{n}, & \mu \text{ odd}, \\ \arg(\alpha_0/\alpha_1) + (\ell_0 - \ell - 2)\pi + \frac{2\pi\ell}{n}, & \mu \text{ even}. \end{cases} \qquad \square$$

Acknowledgments. We thank the referees and the associate editor for very helpful and constructive advice leading to a much improved manuscript.

REFERENCES

[1] G. D. BIRKHOFF, *Boundary value and expansion problems of ordinary linear differential equations*, Trans. Amer. Math. Soc., 9 (1908), pp. 373–395.
[2] ———, *Note on the expansion problems of ordinary linear differential equations*, Rend. Circ. Mat. Palermo, 36 (1913), pp. 115–126.
[3] G. D. BIRKHOFF AND R. E. LANGER, *The boundary problems and developments associated with a system of ordinary linear differential equations of first order*, Proc. Amer. Acad. Arts Sci., 58 (1923), pp. 51–128.
[4] A. E. BRYSON JR. AND B. WIE, *Modelling and Control of Flexible Space Structures*, in Proc. 3rd VPI and SU/AAIAA Symposium on the Dynamics and Control of Large Flexible Spacecraft, June 15–17, 1981, Blacksburg, VA, pp. 153–174.
[5] C. I. BYRNES AND D. S. GILLIAM, *Asymptotic properties of root locus for distributed parameter systems*, in Proc. IEEE Conf. on Decision and Control, Austin, TX, 1988.
[6] ———, *Boundary feedback stabilization of distributed parameter systems*, Signal Processing, Scattering and Operator Theory, and Numerical Methods, Proceedings of the International Symposium MTNS-89, M. Kasashoek, J. van Schuppen, and A. Rau, eds., Birkhäuser, Boston, 1990, pp. 421–428.
[7] ———, *Boundary feedback design for nonlinear distributed parameter systems*, in Proc. IEEE Conf. on Decision and Control, Britton, England 1991.
[8] C. I. BYRNES, D. S. GILLIAM, AND J. HE, *Root locus for hyperbolic distributed parameter systems*, preprint, Texas Tech University, Lubbock, TX, 1993.
[9] ———, *On the admissibility of boundary inputs for a class of distributed parameter control systems*, preprint, Texas Tech University, Lubbock, TX, 1993.
[10] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, Wiley-Interscience, New York, 1953.
[11] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sci., Vol. 8, Springer-Verlag, Berlin, 1978.
[12] R. F. CURTAIN, *Invariance Concepts and Disturbance Decoupling in Infinite Dimensions: A Survey*, in Proc. 23rd IEEE Conf. on Decision and Control, Vol. 3, 1984, pp. 1451–1456.
[13] ———, *Spectral Systems*, Internat. J. Control, 39 (1984), pp. 657-666.
[14] R. F. CURTAIN AND H. ZWART, *An Introduction to Infinite Dimensional Linear System Theory*, unpublished manuscript.
[15] R. F. CURTAIN, *Finite dimensional compensators for some parabolic systems with unbounded control and observations*, SIAM J. Control Optim., 22 (1984), pp. 255–276.
[16] R. F. CURTAIN AND D. SALAMON, *Finite dimensional compensators for infinite dimensional systems with unbounded input operators*, SIAM J. Control Optim., 24 (1986), pp. 797–816.
[17] R. F. CURTAIN, *A synthesis of time and frequency domain methods for the control of infinite-dimensional systems: a system theoretic approach*, in Control and Estimation in Distributed Parameter Systems, H. T. Banks, ed., Frontiers Appl. Math., Vol. 11, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.
[18] R. F. CURTAIN AND G. WEISS, *Well posedness of triples of operators*, International Series of Numerical Mathematics, Vol. 91, Birkhäuser, Basel, 1989.
[19] R. F. CURTAIN, *The spectrum determined growth assumption for perturbations of analytic semigroups*, Systems Control Lett., 2 (1982), pp. 106–109.
[20] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Vols. I, II, and III, Wiley Interscience, New York, 1963.
[21] H. O. FATTORINI, *Boundary Control Systems*, SIAM J. Control Optim., 6 (1968), pp. 349–385.
[22] S. HANSEN AND G. WEISS, *The operator Carleson measure criterion for admissibility of control operators on ℓ^2*, Systems Control Lett., to appear.
[23] J. HE, *Root Locus for Birkhoff Regular Distributed Parameter Control Systems*, Ph.D. thesis, Dept. of Math., Texas Tech University, Lubbock, TX, 1993.
[24] L. F. HO AND D. L. RUSSELL, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–640.
[25] T. KATO, *Perturbation theory for linear operators*, Springer-Verlag, New York, 1966.
[26] H. P. KRAMER, *Perturbation of differential operators*, Pacific J. Math., 7 (1957), pp. 1405–1435.
[27] J. L. LIONS, *Optimal Control of Systems*, Springer-Verlag, New York, 1971.
[28] G. D. MARTIN, *On the control of flexible mechanical systems*, Ph.D. thesis, Stanford University,

Dept. of Aeronautics and Astronautics, Stanford, CA, 1962, 1978.
[29] V. P. MIHAILOV, *Riesz Bases in $L_2(0,1)$*, Dokl. Akad. Nauk SSSR, 144 (1962), pp. 981–984.
[30] M. A. NAIMARK, *Linear differential operator*, I, Ungar, New York, 1967.
[31] S. A. POHJOLAINEN, *Computation of Transmission Zeros for Distributed Parameter Systems*, Internat. J. Control, 33 (1981), pp. 199–212.
[32] ———, *Robust Multivariable PI-Controller for Infinite Dimensional Systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 17–30.
[33] S. A. POHJOLAINEN AND I. LATTI, *Robust Controller for Boundary Control Systems*, Internat. J. Control, 38 (1983), pp. 1189–1197.
[34] F. RELLICH, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.
[35] D. SALAMON, *Infinite dimensional linear systems with unbounded control and observation: a functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[36] M. H. STONE, *A comparison of the series of Fourier and Birkhoff*, Trans. Amer. Math. Soc., 28 (1926), pp. 695–761.
[37] YA. D. TAMARKIN, *Some points in the theory of ordinary differential equations*, Rend. Circ. Mat. Palermo (2), 34 (1912), pp. 354–382.
[38] R. TRIGGIANI, *Boundary Feedback Stabilization of Parabolic Equations*, Appl. Math. Optim., 6 (1980), pp. 201–220.
[39] G. WEISS, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[40] ———, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.
[41] ———, *Admissibility of input elements for diagonal semigroups on ℓ^2 semigroups*, Systems Control Lett., 10 (1988), pp. 79–82.
[42] B. WIE, *On the modeling and control of flexible space structures*, Ph.D. thesis, Stanford University, Dept. of Aeronautics and Astronautics, Stanford, CA, 1981.
[43] C. E. WILDER, *Expansion problems of ordinary linear differential equations with auxillary conditions at more than two points*, Trans. Amer. Math. Soc., 27 (1916), pp. 415–442.
[44] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.
[45] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, SIAM J. Control, 8 (1970), pp. 1–18.
[46] H. J. ZWART, *Geometric theory for infinite dimensional systems*, Lecture Notes in Control and Information Sciences, Vol. 115, Springer-Verlag, New York, 1989.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1428–1446, September 1994

010

OPTIMAL CONTROLS OF NAVIER–STOKES EQUATIONS*

MIHIR DESAI† AND KAZUFUMI ITO†

Abstract. This paper studies optimal control problems of the fluid flow governed by the Navier–Stokes equations. Two control problems are formulated in the case of the driven cavity and flow through a channel with sudden expansion and solved successfully using a numerical optimization algorithm based on the augmented Lagrangian method. Existence and the first-order optimality condition of the optimal control are established. A convergence result on the augmented Lagrangian method for nonsmooth cost functional is obtained.

Key words. flow control, Navier–Stokes equation, augmented Lagrangian method

AMS subject classifications. 35Q10, 49B22, 49D29

1. Introduction.

Considerable progress has been made in mathematical analysis and computation of Navier–Stokes equations. However, very little attention has been given to the question of controlling the Navier–Stokes equations with the exception of experiments (e.g., Nagib, Reisenthal, and Koga [NRK]) and the problem of optimal shape design for drag minimization (e.g., Pironneau [Pi]). An abstract formulation of a control problem does exist in literature (e.g., [Li]). However (to our knowledge[1]) there has been very little effort toward mathematical formulation and computation in the context of control of physical flow situations. As a step in this direction we formulate and solve computationally the steady-state control problem for two flows that have been investigated extensively in numerical computations; i.e., the driven cavity and flow through a channel with sudden expansion. Finding a suitable cost functional that is relevant to the physics of the flow is a very important step in formulating control problems in both the driven cavity and channel flow. The control problems, which are formulated in §2, involve only one-dimensional control input acting through a part of the boundary as Drichlet boundary control. The influence of a one-parameter control input is limited (which can be seen in our numerical calculations) and thus our numerical calculations have been carried out for relatively low Reynolds number flow. Also, some of the hypotheses for our analysis can be verified only for relatively low Reynolds number flow. To improve the performance of our control law we must consider a full boundary control through a part of the boundary. Our analysis can be extended to treat such a problem and will be discussed in a forthcoming paper.

The paper is organized as follows. In §2 two optimal control problems are described. In §3 the basic theory of Navier–Stokes equation is given. In §4 the existence and first-order optimality condition for optimal control problems are established. In §5 a solution technique based on the augmented Lagrangian method is described. Convergence of the augmented Lagrangian method is obtained. Finally some of our numerical findings are reported in §6.

* Received by the editors January 20, 1992; accepted for publication (in revised form) April 7, 1993. This research was supported in part by Air Force Office of Scientific Research grant AFOSR-90-0091 and by National Science Foundation grant DMS-8818530.

† Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, CA 90089-1113.

[1] Since this paper was submitted, there has been increasing interest in the study of control of the Navier–Stokes equations, e.g., [AT], [FS], [GHS], [DG], and the references therein.

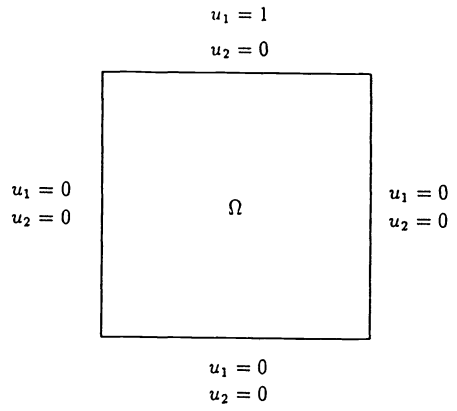


FIG. 1. *Driven cavity.*

2. Two flow problems. In this section two flow control problems will be formulated.

2.1. Driven cavity. Consider the two-dimensional motion of fluid modeled by the (stationary) Navier–Stokes equation,

$$-\nu \Delta u + (u \cdot \nabla)u + \nabla p = f \quad \text{in } \Omega \tag{2.1}$$

$$\nabla \cdot u = 0, \tag{2.2}$$

confined in a square cavity Ω, depicted in Fig. 1. Here $u = (u_1, u_2)$ is the velocity field, p the pressure, ν the kinematic viscosity of the fluid ($\nu = 1/\mathrm{Re}$, where Re is the Reynolds number), and f the density of external forces (in this example $f = 0$). The nonlinear term $(u \cdot \nabla)u$ in (2.1) (often called the convective term), is a symbolic notation for the vector

$$\left(u_1 \frac{\partial u_1}{\partial x_1} + u_2 \frac{\partial u_1}{\partial x_2}, \, u_1 \frac{\partial u_2}{\partial x_1} + u_2 \frac{\partial u_2}{\partial x_2} \right).$$

The divergence-free condition (2.2) is the equation for law of conservation of mass.

Conventionally, the problem has been treated with boundary conditions as in Fig. 1; i.e., only the top surface moving with velocity U_{top}. However we observe (numerically) that if both the top and bottom surfaces move in the same direction, the flow separates into two distinct regions as shown in Fig. 2 (where the top and bottom velocities are .5 and the viscosity $\nu = 1/50$ is used). Hence, the control problem we consider is as follows.

PROBLEM. *Given the bottom velocity U_{bot}, find the top velocity U_{top} such that the separation of flow occurs at a desired horizontal line location Γ_L.*

We cast the problem as a minimization of cost functional defined for U_{top}

$$J(U_{\text{top}}) = \int_{\Gamma_L} |u_2|^2 \, ds, \tag{2.3}$$

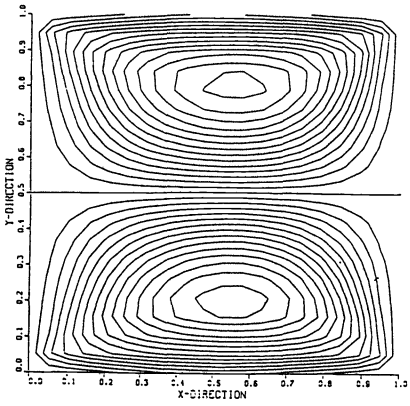


FIG. 2. *Seperation of flow in a cavity.*

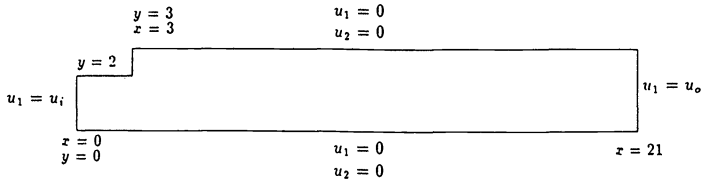


FIG. 3. *Channel with sudden expansion.*

subject to (2.1) and (2.2), where u_2 is the vertical component of the velocity field u and $u_2 = 0$ on Γ_L implies that no flow crosses the horizontal line Γ_L.

Remark 2.1. Note that the boundary condition in Fig. 1 is not in $H^{1/2}(\Gamma)$: i.e., it is not the trace of a function in $H^1(\Omega)$ on Γ. Thus, the existence theory (e.g., in [GR]; also see Theorem 3.5) cannot be applied in this case. However, we are able to show in [Di] that the Stokes equation (that is of the form (2.1), (2.2) without convective term) has a unique solution in $L^2(\Omega)^2$.

2.2. Channel flow. The second problem is a control of channel flow illustrated in Fig. 3. We assume that the inflow (at $x = 0$) and outflow (at $x = 21$) are parabolic (Poiseuille flow assumption) with $u_{\text{in}} = x_2(2 - x_2)$ and $u_{\text{out}} = cx_2(3 - x_2)$ where c is chosen such that $\int_\Gamma u \cdot n\, ds = 0$. There is a recirculation region in the corner whose size increases with the Reynolds number. Figure 4 qualitatively illustrate the flow in the corner with Re= 50. The objective is to shape the flow in the recirculation region to a desired configuration by means of controlled injection (suction) along a portion Γ of the vertical boundary facing the recirculation flow (see the shaded line in Fig. 4). The key question is then: What is a "*desirable*" *flow*? The answer to this question clearly depends on the applications in which the flow situation occurs. We consider the following two cost functionals. The first one corresponds to the total vorticity in the flow given by

$$\int_\Omega \left| \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \right|^2 dx, \tag{2.4}$$

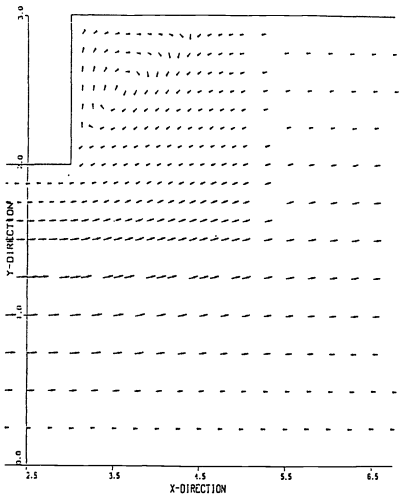


FIG. 4. *Channel flow for* Re = 20.

where the vorticity

$$\omega = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}.$$

This cost is motivated by the fact that potential flows (zero vorticity) are frictionless and incur low energy dissipation. The second one is the "back flow" cost,

$$\int_\Omega (|\min(0, u_1)|^2 + |\min(0, u_2)|^2)\, dx, \tag{2.5}$$

which is motivated by the fact that if we do not have any recirculation, then $u_1 \geq 0$ and $u_2 \geq 0$ (i.e., fluid moving upward and to the right).

3. Basic theory for the Navier–Stokes equations. In this section we summarize the basic theory of the Navier–Stokes equation that we need for our discussions. We consider the boundary-value problems (2.1), (2.2) with the Dirichlet boundary condition $u|_\Gamma = g$ where $\int_\Gamma g \cdot n\, ds = 0$ since by Green's formula

$$\int_\Omega \nabla \cdot u\, dx = \int_\Gamma g \cdot n\, ds = 0. \tag{3.1}$$

Here n is the outward unit normal vector and it is assumed that Ω is bounded open set in R^n, $n = 2, 3$ and its boundary is at least Lipschitz continuous.

Define the function spaces and notation that will be used in what follows (our treatment and notation are along the lines presented in [GR] and [Te]):

$$V = \{u \in H_0^1(\Omega)^n \text{ with } \nabla \cdot u = 0\},$$

$$H = \{u \in L^2(\Omega)^n \text{ with } \nabla \cdot u = 0 \text{ and } u \cdot n = 0 \text{ on } \Gamma\}.$$

Let $a(u,v) : H^1(\Omega)^n \times H^1(\Omega)^n \to R$ be the symmetric sesquilinear form defined by

$$a(u,v) = \nu \int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v \, dx \tag{3.2}$$

and define the bilinear form $c(u,p) : H^1(\Omega)^n \times L^2(\Omega) \to R$ by

$$c(u,p) = \int_\Omega (\nabla \cdot u) p \, dx \quad \text{for } u \in H^1(\Omega)^n \text{ and } p \in L^2(\Omega). \tag{3.3}$$

The trilinear form b on $H^1(\Omega)^n$ that corresponds to the convective term in (2.1) is defined by

$$b(u; v, w) = \int_\Omega \sum_{i,j=1}^{n} u_j \frac{\partial v_i}{\partial x_j} w_i \, dx. \tag{3.4}$$

Then the variational form for the Navier–Stokes equations (2.1), (2.2) with boundary condition $u|_\Gamma = g$ is given by

$$\begin{aligned} a(u,v) + b(u;u,v) - c(v,p) &= \langle f, v \rangle \quad \text{for all } v \in H_0^1(\Omega)^n, \\ c(u,q) &= 0 \quad \text{for all } q \in L^2(\Omega), \end{aligned} \tag{3.5}$$

where $\langle \cdot, \cdot \rangle$ denotes the dual product of $H^{-1}(\Omega)^n \times H_0^1(\Omega)^n$. Discarding the convective term in (2.1) results in the Stokes equation

$$-\nu \Delta u + \nabla p = f \quad \text{and} \quad \nabla u = 0 \tag{3.6}$$

with $u|_\Gamma = g$. Using our notation it can be written as

$$\begin{aligned} a(u,v) - c(v,p) &= \langle f, v \rangle \quad \text{for all } v \in H_0^1(\Omega)^n, \\ c(u,q) &= 0 \quad \text{for all } q \in L^2(\Omega). \end{aligned} \tag{3.7}$$

Note that (3.7) is the first-order necessary optimality condition for the following minimization problem:

$$\text{Minimize} \quad \tfrac{1}{2} a(u,u) - \langle f, u \rangle$$

$$\text{over } u \in H^1(\Omega)^n \text{ with } u|_\Gamma = g,$$

subject to $\nabla \cdot u = 0$. In fact, the Lagrangian corresponding to the constrained minimization stated above is

$$L(u,\lambda) = \tfrac{1}{2} a(u,u) - \langle f, u \rangle - c(u,\lambda),$$

where the Lagrange multiplier $\lambda \in L^2(\Omega)$ turns out to be the pressure p in (3.6). For the homogeneous boundary condition (i.e., $g = 0$) since

$$\operatorname{div} : H_0^1(\Omega) \to L_0^2(\Omega) = \left\{ \phi \in L^2(\Omega) : \int_\Omega \phi \, dx = 0 \right\}$$

is surjective, it follows from [GR] that there exists a unique solution $(u,p) \in V \times L_0^2(\Omega)$ of (3.7). Hence (3.7) is equivalently written as

$$a(u,\psi) = \langle f, \psi \rangle \quad \text{for all } \psi \in V. \tag{3.8}$$

We state the important property of the trilinear form b [GR].

LEMMA 3.1. *The trilinear form b defined by* (3.4) *is continuous on* $(H^1(\Omega)^n)^3$. *Let* $v,\ w \in H^1(\Omega)^n$ *and let* $u \in H^1(\Omega)^n$ *satisfy* $\nabla \cdot u = 0$ *and assume either* $n \cdot u|_\Gamma = 0$ *or* $w \cdot v|_\Gamma = 0$. *Then we have*

$$b(u; v, w) + b(u; w, v) = 0, \tag{3.9}$$

which implies in particular

$$b(u; v, v) = 0. \tag{3.10}$$

We now state the existence result and for the sake of completeness of our discussions we include a sketch of its proof.

THEOREM 3.2. *For* $f \in H^{-1}(\Omega)^n$ *and* $g = 0$ *there exists at least one solution* $(u, p) \in V \times L_0^2(\Omega)$ *of* (3.5).

Proof. It follows from (3.8) and [GR] that (3.5) is equivalent to an equation for $u \in V$:

$$a(u, \psi) + b(u; u, \psi) = \langle f, \psi \rangle \quad \text{for all } \psi \in V. \tag{3.11}$$

Thus, we argue the existence of solutions to (3.11). Define the map $T : V \to V$ by $z = T(u)$ where given $u \in V$, z is a unique solution to

$$a(z, \psi) + b(u; z, \psi) = \langle f, \psi \rangle \quad \text{for all } \psi \in V. \tag{3.12}$$

Existence and uniqueness of a solution to (3.12) can be shown by the Lax–Milgram theory [Yo] since the bilinear form $a(\phi, \psi) + b(u; \phi, \psi)$ defined on $V \times V$ is continuous and V-coercive by Lemma 3.1. The fixed points of T are the solutions of (3.11). Taking $\psi = z$ in (3.12) and from (3.10), we obtain $\|z\|_V \le \frac{1}{\nu} \|f\|_{V^*}$. Hence $T : C \to C$, where the set $C = \{w \in V : \|w\|_V \le \frac{1}{\nu} \|f\|_{V^*}\}$is a bounded, closed convex subset of V. Since V is a Hilbert space, every bounded set in V contains a weak convergent sequence, $u_n \to u$ in V. Since V is compactly embedded into $L^4(\Omega)^n$, the sequence converges strongly in $L^4(\Omega)^n$. However, since for $z_n = T(u_n)$ and $z = T(u)$

$$a(z_n - z, \psi) + b(u_n - u; z_n, \psi) + b(u; z_n - z, \psi) = 0 \quad \text{for all } \psi \in V,$$

and $|b(u; v, w)| \le M \|u\|_{L^4} \|v\|_{H^1} \|w\|_{H^1}$ for $u,\ v,\ w \in H^1(\Omega)^n$,

$$\nu \|z_n - z\|_V \le M \|z_n\|_V \|u_n - u\|_{L^4}.$$

Thus z_n converges strongly to z in V. Hence T is compact and therefore there exists a $u \in C$ such that $T(u) = u$ by the Shauder fixed point theorem [Is]. □

COROLLARY 3.3. *The solution* $u \in V$ *of* (3.11) *is unique provided that* $k \|f\|_{V^*} < \nu^2$ *where for a constant* $k > 0$

$$|b(u; v, w)| \le k \|u\|_V \|v\|_V \|w\|_V. \tag{3.13}$$

For the case of nonhomogeneous Dirichlet boundary conditions we need the following technical lemma due to Hopf [GR].

LEMMA 3.4. *Suppose* Ω *is a bounded open domain with Lipschtz continuous boundary. Then, given* $g \in H^{1/2}(\Gamma)^n$ *satisfying* $\int_\Gamma g \cdot n \, ds = 0$, *for any* $\varepsilon > 0$ *there exists a function* $u_\varepsilon \in H^1(\Omega)^n$ *such that* $\nabla \cdot u_\varepsilon = 0$, $u_\varepsilon|_\Gamma = g$ *and*

$$|b(\phi; u_\varepsilon, \phi)| \le \varepsilon |\phi|_V^2 \quad \textit{for all } \phi \in V. \tag{3.14}$$

Now, the nonhomogeneous boundary value problem can be transformed into the one with homogeneous boundary condition by using the transformation $u = w + u_\varepsilon$; substituting u into (3.5) we obtain the equation for $w \in V$:

$$a(w,\psi) + b(w;u_\varepsilon,\psi) + b(u_\varepsilon;w,\psi) + b(w;w,\psi) = \langle \bar{f},\ \psi\rangle \quad \text{for all } \psi \in V, \tag{3.15}$$

where $\bar{f} \in V^*$ is given by

$$\langle \bar{f},\ \psi\rangle = \langle f,\ \psi\rangle - a(u_\varepsilon,\psi) - b(u_\varepsilon;u_\varepsilon,\psi).$$

Define the sesquilinear form $\hat{a}$ on $V \times V$ by

$$\hat{a}(\phi,\psi) = a(\phi,\psi) + b(\phi;u_\varepsilon,\psi) + b(u_\varepsilon;\phi,\psi) \quad \text{for } \phi,\ \psi \in V.$$

Then, if $\varepsilon < \nu$ then $\hat{a}$ is V-coercive from (3.10) and (3.14). Thus, applying the Shauder fixed point theorem as in the proof of Theorem 3.2 to (3.15) we obtain the following result.

THEOREM 3.5. *Given $f \in H^{-1}(\Omega)$ and $g \in H^{1/2}(\Gamma)$ with $\int_\Gamma g \cdot n\, ds = 0$ there exists at least one pair $(u,p) \in H^1(\Omega)^n \times L_0^2(\Omega)$ satisfying* (3.5).

Remark. $u_\varepsilon \in H^1(\Omega)$ appearing in the Hopf's lemma is used to show the existence of a solution to the Navier–Stokes equations. However, it is not feasible to be used in computations. In our numerical calculations we use $\bar{u} \in H^1(\Omega)^n$ that satisfies the Stokes equation

$$a(\bar{u},v) - c(v,p) = 0 \quad \text{and} \quad c(\bar{u},q) = 0 \tag{3.16}$$

for all $(v,q) \in H_0^1(\Omega) \times L^2(\Omega)$ with boundary condition $\bar{u} = g$ on Γ. Note that $\bar{u}$ is unique [GR] but with $\bar{u}$, condition (3.14) is not necessarily satisfied for arbitrary $\varepsilon > 0$.

4. Existence and first-order necessary condition of optimal solutions. Two control problems described in §2 can be formulated as a constrained minimization in a Hilbert sapce using the notation of §3:

$$\begin{aligned} &\text{Minimize} && J(u) \\ &\text{subject to} && a(u,\psi) + b(u;u,\psi) = 0 \quad \text{for all } \psi \in V \\ &&& \nabla \cdot u = 0 \\ &&& u = g_0 + \sum_{i=1}^{m} f_i\, \chi_i \quad \text{on } \Gamma \\ &&& f \in U, \end{aligned} \tag{4.1}$$

where $g_0,\ \chi \in H^{1/2}(\Gamma)$ with $\int_\Gamma g \cdot n\, ds = \int_\Gamma \chi_i \cdot n\, ds = 0$ and U is a closed bounded set in R^m. We discuss the Dirichlet boundary control problem and thus the body force is discarded. The function $f \cdot \chi = \sum_{i=1}^m f_i\, \chi_i,\ f \in U$ is the control input and influences the equation only through a part of boundary Γ, and the functions χ_i represent

distribution functions of control input at Γ. The specific form of g_0 and χ $(m = 1)$ for the cavity and channel problems is as follows. For the cavity $g_0 = (U_{\text{bot}}, 0)$ at the bottom surface, zero otherwise and $\chi = (1, 0)$ at the top surface, zero otherwise. For the channel

$$g_0 = \begin{cases} x_2(2 - x_2) & \text{at } x_1 = 0, \\ c\,x_2(3 - x_2) & \text{at } x_1 = 21, \\ \text{zero} & \text{otherwise}, \end{cases}$$

and

$$\chi = \begin{cases} (x_2 - 2.375)(2.625 - x_2) & \text{at } x_1 = 3, \\ \hat{c}\,x_2(3 - x_2) & \text{at } x_1 = 21, \\ \text{zero} & \text{otherwise}, \end{cases}$$

where c, $\hat{c}$ are chosen such that $\int_\Gamma g_0 \cdot n\,ds = \int_\Gamma \chi \cdot n\,ds = 0$. As pointed out in Remark 2.1, for the driven cavity problem g_0, χ are not in $H^{1/2}(\Gamma)$. Hence in our discussion we consider the problem in which g_0 and χ are replaced by $C^\infty(\Gamma)$ function that approximates g_0 and χ in $L^\infty-$ norm, respectively.

Let $u = w + \bar{u}^{(0)} + \sum_{i=1}^m f_i\,\bar{u}^{(i)}$ with $w \in V$ where $\bar{u}^{(0)}$ and $\bar{u}^{(i)}$, $1 \le i \le m$ are the solution of the Stokes equation (3.16) with boundary condition g_0 and χ_i, $1 \le i \le m$, respectively. Let $\bar{u} = \text{col}(\bar{u}^{(1)}, \cdots, \bar{u}^{(m)})$. Then the problem (1.4) can be equivalently written as

$$\begin{aligned} &\text{Minimize} \quad J(u) \\ &\text{subject to} \quad a(w, \psi) + b(u; u, \psi) = 0 \quad \text{for all } \psi \in V, \end{aligned} \tag{4.2}$$

where $u = w + \bar{u}^{(0)} + \sum_{i=1}^m f_i\,\bar{u}^{(i)}$ and the cost functional J is minimized over $(w, f) \in V \times U$. Here not only is the boundary control problem transformed into the distributed control problem but also the control f appears directly in the cost functional J.

In our example without loss of generality we assume that $U = [-1, 1]$.

THEOREM 4.1. *The set S of solutions defined by*

$$S = \{u \in H^1(\Omega)^n : u = w + \bar{u}^{(0)} + \textstyle\sum_{i=1}^m f_i\,\bar{u}^i,\ w \in V\ f \in U \text{ and}$$

$$u \text{ satisfies } a(u, \psi) + b(u; u, \psi) = 0 \text{ for all } \psi \in V\ \},$$

is bounded in $H^1(\Omega)^n$.

Proof. Let $u^{(0)}$ and $u^{(i)}$, $1 \le i \le m$ are the Hopf's function (see Lemma 3.4) corresponding to g_0 and χ_i, $1 \le i \le m$, respectively. Then for any $f \in U$ we can write $u = w + u^{(0)} + \sum_{i=1}^m f_i\,u^{(i)}$ where $w \in V$ satisfies

$$a(w, \psi) + b(z; w, \psi) + b(w; z, \psi) + b(w; w, \psi) = -a(z, \psi) - b(z; z, \psi) \quad \text{for all } \psi \in V, \tag{4.3}$$

where $z = u^{(0)} + \sum_{i=1}^m f_i\,u^{(i)}$. It follows from Theorems 3.2 and 3.5 that there exists a solution $w \in V$ of (4.3) and taking $\psi = w$ in (4.3) and using Lemma 3.1, we obtain

$$(\nu - (m+1)\varepsilon)\|w\|_V \le \alpha\,\|z\|_{H^1} \quad \text{for some } \alpha > 0.$$

Thus, there exists a constant $\gamma > 0$ such that $\|u\|_{H^1} \leq \gamma$. □

THEOREM 4.2. *Suppose the cost functional J is weakly, sequentially lower semicontinuous. Then, the control problem* (4.2) *has at least one solution. In particular, each control problem described in* §2 *has at least one solution.*

Proof. Let (w_n, f_n) be a minimizing sequence. Since U is compact and the solution set S is bounded in $H^1(\Omega)^n$ there exists a subsequence (w_n, f_n) such that $f_n \to f^*$ in U and $w_n \to w^*$ weakly in V. Note that $(w_n, f_n) \in V \times U$ satisfies

$$a(u_n, \psi) + b(u_n; u_n, \psi) = 0 \quad \text{for all } \psi \in V$$

$$u_n = w_n + \bar{u}^{(0)} + f_n \cdot \bar{u}$$

and that $|b(w; u, \psi)| \leq M \|w\|_{L^4} \|u\|_{H^1} \|\psi\|_{H^1}$ for w, u, $\psi \in H^1(\Omega)^n$. Since $H^1(\Omega)$ is compactly embedded int $L^4(\Omega)$, $\|u_n - u^*\|_{L^4} \to 0$ and therefore it follows from (3.4) and Lemma 3.1 that $(w^*, f^*) \in S$. Hence if the cost functional $J : H^1(\Omega)^n \to R$ is weakly, sequentially lower semicontinuos, then (w^*, f^*) minimizes J. It is not difficult to show that each cost functional J is weakly, sequentially lower semicontinuous. In fact, for (2.3) the claim follows from the fact that the trace operator of $H^1(\Omega)$ on Γ_L is compact in $L^2(\Gamma_L)$. The cost functional (2.4) is the square of a norm on $H^1(\Omega)^2$. For (2.5) note that the cost functional is continuous on $L^2(\Omega)^2$. Hence each control problem has at least one solution. □

Remark. It is not difficult to extend Theorems 4.1 and 4.2 to the case when the control input g belongs to a compact subset of $H^{1/2}(\Gamma)^n$.

Next we discuss the first-order necessary optimality condition. Assume that $u^* = (w^*, f^*) \in V \times U$ is a local solution of (4.2) and that $f^* \in \text{int}(U)$. Let $G : V \times U \to V^*$ be defined by

$$\langle G(w, f), \psi \rangle = a(w, \psi) + b(u; u, \psi) \quad \text{for } \psi \in V$$

where $u = w + \bar{u}^{(0)} + \sum_{i=1}^m f_i \bar{u}^{(i)}$. In what follows we identify u with the pair (w, f) whenever $u = w + \bar{u}^{(0)} + \sum_{i=1}^m f_i \bar{u}^{(i)}$. It follows from [MZ] that if the Fréchet derivative of G at (w^*, f^*) is surjective, then the regular point condition is satisfied and hence there exists a Lagrange multiplier $\lambda \in V$ such that

$$J'(u^*)(v + h \cdot \bar{u}) + \langle G'(u^*)(v, h), \lambda \rangle = 0 \quad \text{for all } (v, h) \in V \times R^m.$$

LEMMA 4.3. *$G'(u^*)$ is given by*

$$\begin{aligned} \langle G'(u^*)(v, h), \psi \rangle = {} & a(v, \psi) + b(v; u^*, \psi) + b(u^*; v, \psi) \\ & + h \cdot (b(\bar{u}; u^*, \psi) + b(u^*, \bar{u}, \psi)) \quad \textit{for } \psi \in V \end{aligned} \tag{4.6}$$

and $G'(u^*)$ is surjective if and only if the equation for $\psi \in V$

$$a(v, \psi) + b(v, u^*, \psi) + b(u^*; v, \psi) = 0 \quad \textit{for all } v \in V \tag{4.7}$$

$$b(\bar{u}^{(i)}; u^*, \psi) + b(u^*; \bar{u}^{(i)}, \psi) = 0 \quad \textit{for all } 1 \leq i \leq m \tag{4.8}$$

implies $\psi = 0$.

Proof. It is easy to show that the Fréchet derivative G at (w^*, f^*) is given by (4.6). It thus follows from (4.6) that (4.7), (4.8) are equivalent to the fact that

$\ker(G'(u^*)^*) = \{0\}$. Note that $G'(u^*)$ is surjective if and only if there exists a $(v, h) \in V \times R^m$ such that

$$v + Cv + Bh = T_0 f \quad \text{for arbitrary } f \in V^*, \tag{4.9}$$

where $T_0 f = (-\nu \Delta_S)^{-1} f$, $f \in V^*$ is the unique solution to the Stokes equation (3.8), the linear operator $C : V \to V$, defined by $Cv = T_0((v \cdot \nabla)u^* + (u^* \cdot \nabla)v)$, and the linear operator $B : R^m \to V$ is defined by $Bh = h \cdot T_0((\bar{u} \cdot \nabla)u^* + (u^* \cdot \nabla)\bar{u})$. Then since $H^1(\Omega)$ is embedded compactly into $L^4(\Omega)$, C is compact. Since B is of finite rank, V admits the orthogonal decomposition $V = \text{range}(B) \oplus \ker(B^*)$. Let Q be the orthogonal projection onto $\ker(B^*)$. Then (4.9) is equivalent to

$$v + QCv = QT_0 f \quad v \in \ker(B^*).$$

Since QC is compact it follows from the Riesz–Schauder theory [Yo] that $\text{range}(I + QC)$ on $\ker(B^*)$ is closed, which implies that $\text{range}(T_0 G'(u^*))$ is closed. Since $T_0 : V^* \to V$ is isometric isomorphism it follows from the Banach closed range theory [Yo] that $G'(u^*)$ is surjective if and only if $\ker(G'(u^*)^*) = \{0\}$. □

Remark. Lemma 4.3 implies that if the only solution of $v + Cv = 0$ is the zero solution (i.e., -1 is not an eigenvalue of C) then the surjectivity of $G'(u^*)$ is satisfied. Such a case occurs when u^* satisfies

$$|b(\phi; u^*, \phi)| < \nu \|\phi\|_V^2. \tag{4.10}$$

This inequality is, for example, satisfied when ν is sufficiently large since from Theorem 4.1 $\|u^*\|_{H^1(\Omega)^n}$ is uniformly bounded in $\nu \geq \nu_0 > 0$. Moreover, if C has the eigenvalue -1 with multiplicity less than $m + 1$, then the assumption of Lemma 4.3 still holds when (4.8) has only trivial solution on the eigenmanifold corresponding to the eigenvalue -1 of C.

Under the condition in Lemma 4.3 we obtain the first-order necessary condition for optimality:

$$a(u^*, \psi) + b(u^*; u^*, \psi) = 0 \quad \text{for all } \psi \in V, \tag{4.11a}$$

$$a(\lambda, \phi) + b(u^*; \lambda, \phi) + b(\lambda; u^*, \phi) + J'(u^*)(\phi) = 0 \quad \text{for all } \phi \in V, \tag{4.11b}$$

$$b(u^*; \lambda, \bar{u}^{(i)}) + b(\lambda; u^*, \bar{u}^{(i)}) + J'(u^*)(\bar{u}^{(i)}) = 0 \quad \text{for } 1 \leq i \leq m. \tag{4.11c}$$

5. Augmented Lagrangian method and convergence analysis. We solve the constrained minimization problem (4.2) (or equivalently (4.1)) using the augmented Lagragian method [He], [Po]. In our approach the divergence free constraint is imposed explicitly (without augmentation) but the Navier–Stokes constraint $G(w, f) = 0$ in V^* will be treated by the augmented Lagrangian method. We consider the augmented Lagrange functional

$$L_c(u, \lambda) = J(u) + \langle \lambda, G(w, f) \rangle_{V, V^*} + \frac{c}{2} a(z, z) \tag{5.1}$$

where $z \in V$ satisfies

$$a(z, \psi) = a(u, \psi) + b(u; u, \psi) \quad \text{for all } \psi \in V. \tag{5.2}$$

Note that $a(z, z)$ represents the square of the norm $\|G(w, f)\|_{V^*}$. Then, the augmented Lagrangian method applied to (4.2) is the following iterative scheme.

AUGMENTED LAGRANGIAN METHOD

Step 1. Choose the starting $\lambda_0 \in V$, a nondecreasing sequence of positive numbers c_k and set $k = 0$.

Step 2. Given λ_k, c_k find $u_k = (w_k, f_k) \in V \times U$ by

$$L_{c_k}(u_k, \lambda_k) = \min\ L_{c_k}(u, \lambda_k) \quad \text{over } (w, f) \in V \times U.$$

Step 3. Update λ_k by $\lambda_{k+1} = \lambda_k + c_k z_k$ where $z_k \in V$ satisfies

$$a(z_k, \psi) = \langle G(w_k, f_k), \psi \rangle \quad \text{for all } \psi \in V. \tag{5.3}$$

Step 4. If convergence criterion is not satisfied then $k = k + 1$ and go to Step 2.

Let $u^* = (w^*, f^*)$ be a local solution of (4.2). Assume that the assumption in Lemma 4.3 is satisfied. Then there exists a Lagrange multiplier $\lambda^* \in V$ such that (4.5) holds, where J is assumed to be continuously Fréchet differentiable in a neighborhood of u^*. Augmentabilty defined as below is of central importance in showing the convergence of the augmented Lagrangian method [He], [IK1].

DEFINITION. *The problem* (4.2) *is augmentable at* u^* *if there exist a neighborhood* $\tilde{U}(u^*)$ *of* (w^*, f^*) *in* $V \times U$ *and positive constants* $\bar{\sigma}$, $\bar{c}$ *such that*

$$L_c(u, \lambda^*) - L_c(u^*, \lambda^*) \geq \bar{\sigma}\,(\|w - w^*\|_V^2 + |f - f^*|^2) \tag{5.4}$$

for all $(w, f) \in \tilde{U}$ *and* $c \geq \bar{c}$.

Then the following theorem follows from [IK1], [IK2].

THEOREM 5.1. *Assume that the augmentability* (5.4) *holds. Given* $\lambda_0 \in V$ *for each* k *assume that* (w_k, f_k) *in the neighborhood* $\tilde{U}$ *satisfies*

$$L_c(u_k, \lambda_k) \leq L_c(u^*, \lambda_k) = J(u^*) \tag{5.5}$$

and update $\lambda_{k+1} = \lambda_k + (c_k - \bar{c})\, z_k$ *with* z_k *satisfying* (5.3). *Then we have*

(1) $$\bar{\sigma}\,(\|w_k - w^*\|_V^2 + |f_k - f^*|^2) + (c_k - \bar{c})\,\|\lambda_k - \lambda^*\|_V^2 \leq (c_k - \bar{c})\,\|\lambda_{k-1} - \lambda^*\|_V^2,$$

(2) $$\bar{\sigma} \sum_{k-1}^{\infty} (\|w_k - w^*\|_V^2 + |f_k - f^*|^2) \leq \frac{1}{c_0 - \bar{c}}\,\|\lambda_0 - \lambda^*\|_V^2.$$

Remark. (i) The condition (w_k, f_k) being in the neighborhood of u^* can be satisfied either by taking c_0 sufficiently large or λ_0 sufficiently close to λ^*.

(ii) The statement (2) implies the strong convergence of u_k to u^* in $H^1(\Omega)^n$.

(ii) The condition (5.5) means the sufficient reduction of successive cost functional.

Next we discuss a sufficient condition for the augmentability of (4.2). Let L be the Lagrangian corresponding to (4.2) defined by

$$L(u, \lambda) = J(u) + \langle \lambda, G(w, f) \rangle.$$

Suppose $J : H^1(\Omega) \to R$ is twice continuously differentiable in a neighborhood of u^*. The second derivative of $L(u, \lambda^*)$ at $u^* = (w^*, f^*)$ is given by

$$L''(u^*, \lambda^*)((v, h), (v, h)) = J''(u^*)(\xi, \xi) + b(\xi; \xi, \lambda^*), \tag{5.6}$$

where $\xi = v + h \cdot \bar{u}$. Then it follows from [IK1] that the augmentability (5.4) is achieved if the following second-order sufficient optimality condition is satisfied.

$$
\begin{aligned}
&L''(u^*, \lambda^*)((v,h),(v,h)) \geq \sigma\,(\|v\|_V^2 + |h|^2) \\
&\text{for all } (v,h) \in V \times R^m \text{ satisfying } G'(u^*)(v,h) = 0,
\end{aligned}
\tag{5.7}
$$

where $\sigma > 0$. Moreover, it follows from [PT], [IK1] that $(u_k, \lambda_k) \in H^1(\Omega)^n \times V$ converges q-linearly to (u^*, λ^*); i.e.,

$$
\|\lambda_n - \lambda^*\|_V \leq \frac{K^n}{c_0 \cdots c_{n-1}} \|\lambda_0 - \lambda\|_V,
$$

$$
\|u_n - u^*\|_{H^1(\Omega)^n} \leq \frac{K^n}{c_0 \cdots c_{n-1}} \|\lambda_0 - \lambda\|_V
$$

for $n \geq 1$ and some constant K, provided that c_0 is sufficiently large or λ_0 sufficiently close to λ^*.

The following lemma gives an algebraic characterization of the second order sufficient optimality (5.7).

LEMMA 5.2. *Assume that for each i equation for V*

$$
a(v,\psi)+b(v;u^*,\psi)+b(u^*;v,\psi) = -b(u^*;\bar{u}^{(i)},\psi)-b(\bar{u}^{(i)};u^*,\psi)) \quad \text{for all } \psi \in V \tag{5.8}
$$

has a unique solution $v_i \in V$ Then the condition (5.7) *is satisfied if and only if the matrix M on R^m, defined by*

$$
M_{i,j} = J''(u^*)(\phi_i, \phi_j) + b(\phi_i; \phi_j, \lambda^*) \quad \text{with } \phi_i = v_i + \bar{u}^{(i)} \tag{5.9}
$$

is positive definite; i.e., $h^t M h \geq \alpha\,|h|^2$ for all $h \in R^m$ and some $\alpha > 0$.

Proof. Suppose $(v,h) \in V \times R^m$ satisfies $G'(u^*)(v,h) = 0$. Since for each i (5.8) has a unique solution v_i it thus follows from (4.6) and (5.8) that $v = \sum_{i=1}^m h_i v_i$ and therefore $\xi = \sum_{i=1}^m h_i(v_i + \bar{u}^{(i)})$ in (5.6). Then from (5.6) and (5.9) we have

$$
L''(u^*)((v,h),(v,h)) = h^t M h \quad \text{for } (v,h) \in \ker\,(G'(u^*)).
$$

Since $\|v\|_V \leq |h| \sqrt{\sum_{i=0}^m \|v_i\|_V^2}$ the positivity of the matrix M defined by (5.9) is equivalent to (5.7). □

COROLLARY 5.3. *Suppose the matrix $\bar{M}$ on R^m, defined by $\bar{M}_{i,j} = J''(\bar{u}^{(i)}, \bar{u}^{(j)})$, is positive definite. Then, if ν is sufficiently large then* (5.8) *has a unique solution and* (5.9) *holds.*

Proof. It follows from (3.16) and Theorem 4.1 that $\|u^*\|_{H^1(\Omega)^n}$, $\|\bar{u}^{(i)}\|_{H^1(\Omega)^n}$, $1 \leq i \leq m$ are uniformly bounded. Thus, if ν is sufficiently large then (4.10) holds and hence (5.8) has a unique solution $v_i \in V$ for each i. It follows from (4.11b) and (5.8) that

$$
\|v_i\|_V \leq \frac{M_1}{\nu} \|u^*\| \|\bar{u}^{(i)}\| \quad \text{and} \quad \|\lambda^*\|_V \leq \frac{M_2}{\nu} \|J'(u^*)\|_{V^*}
$$

for some constants M_1, M_2. Thus, from Lemma 3.1 and (5.9) if the matrix $\bar{M}$ is positive definite then the matrix M is positive definite provided that ν is sufficiently large. Hence, the corollary follows from Lemma 5.2. □

Note that the cost functional (2.5) is not twice Fréchet differentiable. Motivated by this example we consider the following minimization problem in a Hilbert space X:

$$\text{Minimize } f(x) \text{ subject to } g(x) = 0,$$

where $g : X \to Y$ and y is a Hilbert space. Assume the following standard hypotheses:

(H1) $x^* \in X$ is a local solution;

(H2) g is twice continuously F-differentiable and f is (only) continuously $F-$ differentiable in a convex neighborhood of x^*;

(H3) $g'(x^*) : X \to Y$ is surjective.

Then there exists a unique Lagrange multiplier $\lambda^* \in Y^*$ such that

$$f'(x^*)(h) + \langle \lambda^*, g'(x^*)(h) \rangle = 0 \quad \text{for all } h \in X. \tag{5.10}$$

Define the functional F_t defined on $X \times X$ for $t > 0$ by

$$F_t(x)(h,h) = \frac{1}{t^2}\,(f(x + t\,h) - f(x) - t\,f'(x)h). \tag{5.11}$$

We make the following hypotheses on F_t:

(H4) There exists $\delta > 0$ and $0 < c_1 < 1 < c_2$ such that for $h = h_1 + h_2,\ h_1,\ h_2 \in X$ and $0 < t < \delta$

$$F_t(x^*)(h,h) \geq c_1\, F_t(x^*)(h_1,h_1) - c_2\, \|h_2\|_X^2.$$

(H5) There exists a $\sigma > 0$ such that

$$c_1\, F_t(x^*)(h,h) + \tfrac{1}{2}\,\langle \lambda^*, g''(h,h) \rangle \geq \sigma\, \|h\|_X^2$$

for all $h \in X$ satisfying $g'(x^*)h = 0$ and t sufficiently small.

Remark. If is f is twice differentiable the hypothesis (H5) reduces to the second sufficient optimality condition (5.7). Moreover, assuming $f''(x^*)$ is nonnegative definite then (H4) holds with $c_1 = 1/2$ and $c_2 = \|f''(x^*)\|$.

We have the following result.

THEOREM 5.4. *Assuming* (H1)–(H5), *we have the augmentability*; *i.e., there exists a neighborhood* $\tilde{U}(x^*)$ *and* $\bar{\sigma},\ \bar{c} > 0$ *such that*

$$L_c(x, \lambda^*) - L_c(x^*, \lambda^*) \geq \bar{\sigma}\, \|x - x^*\|_X^2$$

for all $x \in \tilde{U}(x^*)$ *and* $c \geq \bar{c}$ *where for* $c \geq 0$ *and* $\lambda \in Y^*$

$$L_c(x, \lambda) = f(x) + \langle \lambda,\, g(x) \rangle_{Y^*,Y} + \tfrac{1}{2}\, \|g(x)\|_Y^2.$$

Proof. For each $h \in X$ with $\|h\| = 1$ we have

$$\begin{aligned} L_c(x + t\,h, \lambda^*) - L_c(x^*, \lambda^*) &= t\,(f'(x^*)h + \langle \lambda^*,\, g'(x^*)h\rangle) \\ &\quad + t^2\,(F_t(x^*)(h,h) + \frac{1}{2}\,\langle \lambda^*,\, g''(x^*)(h,h)\rangle) \\ &\quad + \frac{1}{2}t^2\,\langle \lambda^*,\, g''(\zeta(t))(h,h) - g''(x^*)(h,h)\rangle) \\ &\quad + \frac{c}{2}t^2\,\|g'(x^*)h + \frac{t}{2}\,g''(\zeta(t))(h,h)\|_Y^2, \end{aligned}$$

where $\zeta(t) = x^* + \hat{t}\,h$, $\hat{t} \in (0,t)$ (which depends upon $(t,h) \in R \times X$). Since $g'(x^*) : X \to Y$ is surjective there exists a constant $\beta > 0$ such that

$$\|g'(x^*)h\|_Y^2 \geq \beta\,\|h\|_X^2 \tag{5.12}$$

for $h \in \text{range}(g'(x^*)^*)$ where by the closed range theory [Yo] we have

$$X = \text{range}(g'(x^*)^*) \oplus \ker(g'(x^*)).$$

Thus, for each $h \in X$ we have a unique representation $h = h_1 + h_2$, $h_1 \in \ker(g'(x^*))$ and $h_2 \in \text{range}(g'(x^*)^*)$. It then follows from (5.10) – (5.12) and (H4) that for $x = x^* + t\,h$

$$\begin{aligned} I &= \frac{1}{t^2}\,(L_c(x, \lambda^*) - L_c(x^*, \lambda^*)) \\ &\geq c_1\,F_t(x^*)(h_1,h_1) + \frac{1}{2}\,\langle \lambda^*, g''(x^*)(h_1,h_1)\rangle + \left(\frac{c\beta}{4} - c_2\right)\beta\,\|h_2\|^2 \\ &\quad + \langle \lambda^*, g''(x^*)(h_1,h_2)\rangle + \frac{1}{2}\,\langle \lambda^*, g''(x^*)(h_2,h_2)\rangle - Mt(1+ct) \end{aligned}$$

for some $M > 0$, where we used the fact that g'' is Lipschitz continuous in a neighborhood of x^*. It then follows from (H5) that

$$\begin{aligned} I &\geq \sigma\,\|h_1\|^2 + \left(\frac{c\beta}{4} - c_2\right)\,\|h_2\|^2 - Mt(1+ct) - \frac{\varepsilon}{2}\,\|h_1\|^2 \\ &\quad - \frac{1}{2}\left(1 + \frac{\|\lambda^*\|\,\|g''(x^*)\|}{\varepsilon}\right)\,\|\lambda^*\|\,\|g''(x^*)\|\,\|h_2\|^2, \end{aligned}$$

for all $\varepsilon > 0$. Thus, for $t > 0$ sufficiently small (say, $0 < t < \delta$) we can choose $\bar{c} > 0$ such that

$$I \geq \bar{\sigma}\,\|h\|_X^2 \quad \text{with } \bar{\sigma} = \frac{\sigma}{2},$$

for all $c \geq \bar{c}$ and $h \in X$ satisfying $\|h\| = 1$ which completes the proof. □

Now we apply Theorems 5.1 and 5.4 to the control problem (2.5). An elementary calculation in [Di] shows that the function defined by

$$f_t(x,k) = \tfrac{1}{2}\,(|\min(0, x + tk)|^2 - |\min(0,x)|^2) - tk\,\min(0,x)$$

for $x,\ k \in R$, satisfies

$$f_t(x, k_1 + k_2) \geq \tfrac{1}{2} f_t(x, k_1) - t^2 |k_2|^2 \tag{5.14}$$

for $k_1,\ k_2 \in R$. Note that for the control problem (2.5), F_t defined by (5.11) is given by

$$F_t(u^*)(h,h) = \frac{1}{t^2} \int_\Omega (f_t(u_1^*(x), h_1(x)) + f_t(u_2^*(x), h_2(x)))\, dx, \tag{5.15}$$

where $u^* = (u_1^*, u_2^*)$ and $h = (h_1, h_2) \in L^2(\Omega)^2$. It thus follows from (5.14) and the dominated convergence theorem that

$$F_t(u^*)(\xi,\xi) \geq \tfrac{1}{2} F_t(\xi^{(1)}, \xi^{(1)}) - \|\xi^{(2)}\|_{L^2}^2, \tag{5.16}$$

where $h = v + k\bar{u}^{(1)}$, $(v,k) \in V \times R$ and $\xi = \xi^{(1)} + \xi^{(2)} \in H^1(\Omega)^2$. Assume henceforth that (5.8) has a unique solution $v_1 \in V$. Then, as argued in the proof of Lemma 5.2, the hypothesis (H5) is equivalent to the following condition:

$$F_t(u^*)(\phi,\phi) + b(\phi;\phi,\lambda^*) > 0,$$

where $\phi = v_1 + \bar{u}^{(1)}$, for the control (2.5). Let $\Omega^* = \{x \in \Omega : u_1^*(x) \leq 0$ and $u_2^*(x) \leq 0\}$. Then it follows from (5.14), (5.15) that

$$F_t(u^*)(\phi,\phi) \geq \frac{1}{t^2} \int_{\Omega^*} (f_t(u_1^*(x), h_1(x)) + f_t(u_2^*(x), h_2(x)))\, dx, \tag{5.17}$$

since the right-hand side of (5.17) is monotonically nonincreasing in t. Thus, assuming

$$\frac{1}{t^2} \int_{\Omega^*} (f_t(u_1^*(x), h_1(x)) + f_t(u_2^*(x), h_2(x)))\, dx$$

is positive for some $t > 0$, it can be shown as in the proof of Corollary 5.3 that for ν sufficiently large (5.16) holds. Hence Theorems 5.1 and 5.4 apply to the control problem (2.5).

6. Numerical results. In this section we discuss numerical solution of the optimal control problems formulated in §2. The solution to (4.2) (equivalently (4.1)) is determined using the augmented Lagrangian method described in §5. The method involves the successive minimization of the cost functional of form

$$L_c(u,\lambda) = J(u) + a(u,\lambda) + b(u;u,\lambda) + \frac{c}{2} a(z,z) \tag{6.1}$$

over $(w,f) \in V \times U$ where $c > 0$ and $\lambda \in V$ are given, $u = w + \bar{u}^{(0)} + h \cdot \bar{u}$ and $z \in v$ satisfies (5.2). We use the projected conjugate gradient method (e.g., see [Gl]) to solve the constrained (i.e., $w \in V$ involves the divergence free condition $\nabla \cdot w = 0$) minimization problem (6.1).

PROJECTED CONJUGATE GRADIENT METHOD

Step 1. Choose the start-up $(w_0, f_0) \in V \times R^m$ and set $k = 0$.

Step 2. Compute the gradient $(g_k, r_k) \in V^* \times R^m$ by

$$\langle g_k, \psi \rangle_{V^*,V} = J'(u_k)(\psi) + a(\psi, \tilde{\lambda}) + b(\psi; u_k, \tilde{\lambda}) + b(u_k; \psi, \tilde{\lambda}) \quad \text{for all } \psi \in V$$

$$r_k = b(u_k; \bar{u}, \tilde{\lambda}) + b(u_k; \bar{u}, \tilde{\lambda}),$$

where $\tilde{\lambda} = \lambda + c\, z_k$ and $z_k \in V$ satisfies (5.3).

Step 3. The gradient $g_k \in V^*$ is projected onto V by the Stokes projection [Gl]; i.e., the projection $h_k \in V$ of g_k is given by

$$a(h_k, \psi) = \langle g_k, \psi \rangle \quad \text{for all } \psi \in V. \tag{6.2}$$

Set $d_k = (h_k, r_k) \in V \times R^m$ as the search direction.

Step 4. Set $\eta_k = h_k + r_k \cdot \bar{u}$. Compute $a_k = \text{Argmin } L_c(u_k - \alpha\, \eta_k, \lambda)$ over $\alpha > 0$ (line search) and set

$$w_{k+1} = w_k - \alpha_k\, h_k,$$

$$f_{k+1} = f_k - \alpha_k\, r_k.$$

Step 5. Find gradients $(h_{k+1}, r_{k+1}) \in V \times R^m$ as in Step 2, 3 and compute

$$\beta_k = \frac{a(h_k, h_k) + |r_k|^2}{a(h_{k+1}, h_{k+1}) + |r_{k+1}|^2}$$

and set the search direction as

$$d_{k+1} = \begin{pmatrix} h_{k+1} \\ r_{k+1} \end{pmatrix} + \beta_k\, d_k.$$

Step 6. If the convergence criterion is not satisfied, then set $k = k + 1$ and go to Step 4.

Remark. Note that for $\alpha > 0$ if $z(\alpha) \in V$ satisfies

$$a(z(\alpha), \psi) = a(u(\alpha), \psi) + b(u(\alpha); u(\alpha), \psi) \quad \text{for all } \psi \in V$$

with $u(\alpha) = u_k - \alpha \eta_k$, then $z(\alpha)$ can be written as

$$z(\alpha) = z_k + \alpha\, z_k^{(1)} + \alpha^2\, z_k^{(2)},$$

where $z_k^{(1)}$, $z_k^{(2)} \in V$ satisfy

$$a(z_k^{(1)}, \psi) = -a(\eta_k, \psi) - b(\eta_k; u_k, \psi) - b(u_k; \eta_k, \psi) \tag{6.3}$$

$$a(z_k^{(2)}, \psi) = b(\eta_k; \eta_k, \psi) \tag{6.4}$$

for all $\psi \in V$. Thus, $L_c(u_k - \alpha\, \eta_k, \lambda)$ is the polynomial of degree four in α and one can carry out the line search in Step 4 exactly provided that J is quadratic in u. Moreover, once the value of α_k in Step 4 is determined, then $z_{k+1} \in V$ is given by

$$z_{k+1} = z_k + \alpha_k\, z_k^{(1)} + \alpha_k^2\, z_k^{(2)}.$$

Hence our algorithm is reduced to solving a series of Stokes problem. Each inner iteration of the augmented Lagrangian algorithm (Step 2) requires solution of three Stokes problems; two (i.e., (6.3), (6,4)) in the line search and one ((6.2)) for the projection of the gradient onto V.

To carry out the computation we discretized the problem using the mixed finite element method [GR], [Te]. In our calculations the quadrilateral element for the

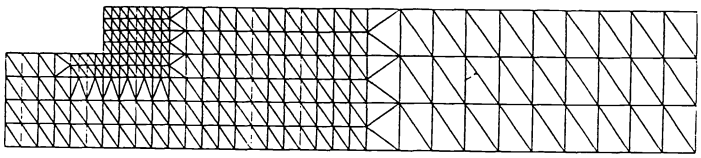


FIG. 5. *The triangulation for the channel*

velocity and the bilinear element for the pressure defined on the uniform rectangular grid with meshsize $h = .1$, are used in the cavity problem. The piecewise quadratic element for the velocity and the linear element for the pressure over the triangular grid shown Fig. 5 are used for the channel flow.

Let $\{\phi_h^i\}$, $\{\chi_h^i\}$ be the linearly independent basis functions of $H_h \subset H^1(\Omega)^2$ and $Q_h \subset L^2(\Omega)$ for velocity and pressure, respectively. Then, for example, the discretization of the Stokes equation (3.6) is given by

$$\begin{aligned} &a(u_h, \psi_h) - c(\psi_h, p_h) = \langle f, \psi_h \rangle \quad \text{for all } \psi_h \in H_h^0 \\ &c(u_h, \chi_h) = 0 \quad \text{for all } \chi_h \in Q_h \end{aligned} \tag{6.5}$$

where $u_h = \sum u_h^i \phi_h^i \in H_h$, $p_h = \sum p_h^i \chi_h^i \in Q_h$ and H_h^0 is the subspace of H_h that consists of all functions ψ_h in H_h satisfying $\psi_h \in H_0^1(\Omega)^2$. The solution $(u_h, p_h) \in H_h \times Q_h$ of (6.5) can be obtained by solving the following system of linear equations:

$$\begin{aligned} &A_h x + B_h^t y = f_h, \\ &B_h x = c_h \end{aligned} \tag{6.6}$$

where the (i,j)th element of the square matrix A_h is given by $a(\psi_h^i, \psi_h^j)$, ψ_h^i, $\psi_h^j \in H_h^0$ and the one of B_h is given by $c(\psi_h^i, \chi_h^j)$, $\chi_h^j \in Q_h$. We refer to [Di] for the detailed discussion of the discretization procedure and solution techniques for (6.6).

We now present numerical results for the control problems. For the problem (2.3) we take the Reynolds number to be 50 ($\nu = 1/50$) and Γ_L to be the horizontal line 0.4 units from the bottom. Given the bottom velocity as 0.5 we obtain the top velocity $U_{\text{top}}^{\text{opt}} = 1.16$ after four iterations of the augmented Lagrangian method with the value of c's equal to 20. The resulting flow field is shown in Fig. 6.

For the channel flow problem with Reynolds number 20 ($\nu = 1/20$) we obtain the following results. For the vorticity cost (2.4), using $c_k = 20$ in Step 2, the optimal control $f^{\text{opt}} = -.77$ (suction) and the optimal cost functional 23.13 are attained after five updates of the Lagrange multiplier. The resulting flow field is shown in Fig. 7. For the back flow cost (2.5), using $c_k = 0.05$ the optimal control $f^{\text{opt}} = 0.11$ (injection) is obtained and the optimal field flow is shown in Fig. 8. It must be noted, however, that a larger injection than f^{opt} decreases the size of the "bubble," but the strength (velocity of recirculation) is higher.

We also simulate the flow corresponding to the optimal control input using the ADI scheme (e.g., see [Gl]) with a finer discretization. Such simulations are in good agreement with the resulting flows we obtain in all three problems.

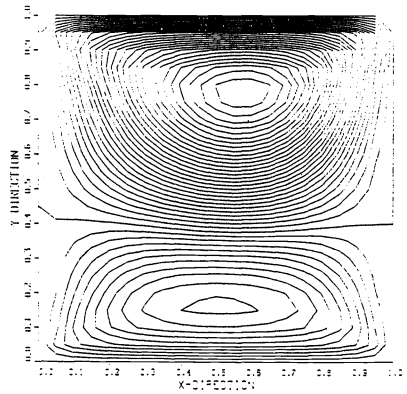


FIG. 6. *Locating seperation line at* $x_2 = 0.4$.

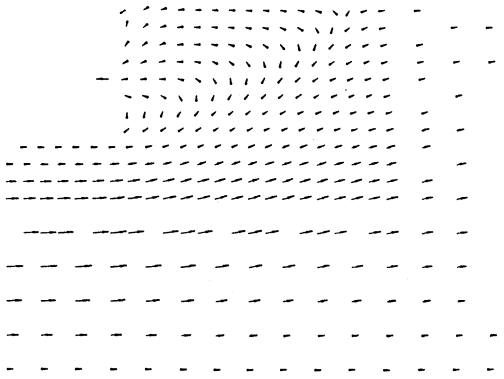


FIG. 7. *Control problem for channel with vorticity cost function* (Re = 20).

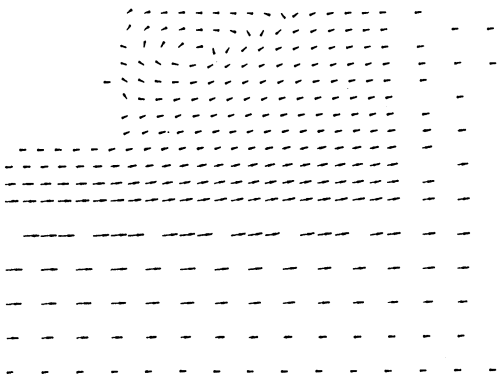


FIG. 8. *Control problem for channel with back flow cost function* (Re = 20).

REFERENCES

[AT] F. ABERGEL AND R. TEMAM, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Mech., 1 (1990), pp. 303–325.

[Di] M. C. DESAI, *Computations in Optimal Control of Navier–Stokes Equations*, Ph.D. Thesis, Division of Applied Mathematics, Brown University, May, 1990.

[DG] E. DEAN AND P. GUBERNATIS, *Pointwise control of Burgers' equation -a numerical approach*, Computer and Mathematics, 22 (1991), pp. 93–100.

[FS] H. O. FATTORINI AND S. S. SRITHARAN, *Necessary and sufficient conditions for optimal controls in viscous flow problems*, Proc. Roy. Soc. of Edinburgh Sect. A, to appear.

[Gl] R. GLOWINSKI, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, Berlin, 1984.

[GHS] M. GUNZBURGER, L. HOU, AND T. SVOBODNY, *Boundary velocity control of incompressible flow with an application to drag reduction*, SIAM J. Control Optim., 30 (1992), pp. 167–181.

[GR] V. GIRAULT AND P. A. RAVIART, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1984.

[He] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1968), pp. 303–320.

[Is] V. ISTRATESCU, *Fixed Point Theory*, D. Riedel Publishing Company, Holland, 1981.

[IK1] K. ITO AND K. KUNISCH, *The augmented Lagrangian method for equality and inequality constraints in Hilbert spaces*, Math. Programming, 46 (1990), pp. 341–360.

[IK2] ———, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113–136.

[Li] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[MZ] H. MAURER AND J. ZOWE, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math Programming, 16 (1979), pp. 98–110.

[NRK] H. M. NAGIB, P. H. REISENTHAL, AND D. J. KOGA, *Control of separated flows using forced unsteadiness*, AIAA Shear Flow Conference, Boulder, CO, 1985.

[Pi] O. PIRONNEAU, *On optimal design in fluid mechanics*, J. Fluid Mech., 64 (1974), pp. 97–110.

[Po] M. J. D. POWELL, *A method for nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1969.

[PT] V. T. POLYAK AND N. Y. TRET'YAKOV, *The method of penalty estimates for conditional extremum problems*, Z. Vychisl. Mat. i Mat. Fiz., 13 (1973), pp. 34–46.

[Te] R. TEMAM, *Navier-Stokes Equations: Theory and Numerical Analysis*, North–Holland, Amsterdam, 1971.

[Yo] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1447–1475, September 1994

011

NECESSARY CONDITIONS FOR OPTIMAL CONTROL OF STOCHASTIC SYSTEMS WITH RANDOM JUMPS*

SHANJIAN TANG† AND XUNJING LI†

Abstract. A maximum principle is proved for optimal controls of stochastic systems with random jumps. The control is allowed to enter into both diffusion and jump terms. The form of the maximum principle turns out to be quite different from the one corresponding to the pure diffusion system (the word "pure" here means the absence of the jump term). In calculating the first-order coefficient for the cost variation, only a property for Lebesgue integrals of scalar-valued functions in the real number space $\mathcal{R}$ is used. This shows that there is no essential difference between deterministic and stochastic systems as far as the derivation of maximum principles is concerned.

Key words. maximum principle, optimal stochastic control, Poisson point process, Lebesgue integral

AMS subject classifications. 49K45, 93E20

1. Introduction.

1.1. Basic notations. We write $\mathcal{R}^k$ for k-dimensional Euclidean space, $\mathcal{R}^{k\times l}$ for the space of matrices with order $k \times l$,and $\mathcal{Z}$ for some nonempty subset of $\mathcal{R}^l$. We denote by $L(\mathcal{R}^d, \mathcal{R}^{n\times n})$ the space of linear continuous operators from $\mathcal{R}^d$ to $\mathcal{R}^{n\times n}$. The element $B \in L(\mathcal{R}^d, \mathcal{R}^{n\times n})$ is represented as $B = \{B^i\}_1^d =: (B^1, B^2, \dots, B^d)$, with $B^i \in \mathcal{R}^{n\times n}$ $(i = 1, 2, \dots, d)$.

Let $(\Omega, \mathcal{F}, \nu)$ be a complete probability space with a ν-completed right-continuous filtration $\mathcal{F}_t$, let $w(\cdot)$ $(= (w^1(\cdot), w^2(\cdot), \dots, w^d(\cdot)))$ be an $\mathcal{R}^d$-valued standard Wiener process, and let $k(\cdot)$ be a stationary $(\mathcal{F}_t)$-Poisson point process on $\mathcal{Z}$ with the characteristic measure $\pi(dz)$ [16]. We denote by $N_k(dzdt)$ the counting measure induced by $k(\cdot)$ and set $\widetilde{N}_k(dzdt) = N_k(dzdt) - \pi(dz)dt$. We assume that

$$\mathcal{F}_t = \sigma\left[\iint_{A\times(0,s]} N_k(dzd\tau); s \le t, A \in \mathcal{B}(\mathcal{Z})\right] \bigvee \sigma[w(s); s \le t] \bigvee \mathcal{N},$$

where $\mathcal{N}$ denotes the totality of ν-null sets and $\sigma_1 \vee \sigma_2$ denotes the σ-field generated by $\sigma_1 \cup \sigma_2$.

Let $\mathcal{H}$ be a finite-dimensional vector space. We also denote by $L^2_{\mathcal{F}}[[0,1];\mathcal{H}]$ the space of $\mathcal{H}$-valued square integrable $(\mathcal{F}_t)$-adapted process, by $L^2_{\mathcal{F},p}[[0,1];\mathcal{H}]$ the space of $(\mathcal{F}_t)$-predictable versions of equivalent classes in $L^2_{\mathcal{F}}[[0,1];\mathcal{H}]$, by $F^2_\pi[\mathcal{Z};\mathcal{H}]$ the Hilbert space of the square integrable functions $\bar{f}(\cdot) : \mathcal{Z} \to \mathcal{H}$, and by $F^2_p[[0,1];\mathcal{H}]$ the space of $\mathcal{H}$-valued $(\mathcal{F}_t)$-predictable (see [16]) vector processes $\hat{f}(\cdot,\cdot,\cdot)$ defined on $\mathcal{Z} \times [0,1] \times \Omega$ such that

$$\int\int_{\mathcal{Z}\times[0,1]} E|\hat{f}(z,t,\cdot)|^2 \pi(dz)dt < \infty.$$

* Received by the editors July 9, 1992; accepted for publication (in revised form) April 27, 1993. This work was partially supported by National Science Foundation of China grants 19131050 and 19301012 and by Chinese State Education Commission Science Foundation grant 9224609.

† Institute of Mathematics, Fudan University, Shanghai 200433, People's Republic of China.

For $\hat{f} \in F_p^2[[0,1];\mathcal{H}]$, we define the following stochastic integral

$$\int\int_{\mathcal{Z}\times(0,t]} \hat{f}(z,\tau,\cdot)\widetilde{N}_k(dzd\tau)$$

as in [16], and set

$$\int\int_{\mathcal{Z}\times(s,t]} \hat{f}(z,\tau,\omega)\widetilde{N}_k(dzd\tau) =: \int\int_{\mathcal{Z}\times(0,t]} \hat{f}(z,\tau,\omega)\widetilde{N}_k(dzd\tau) - \int\int_{\mathcal{Z}\times(0,s]} \hat{f}(z,\tau,\omega)\widetilde{N}_k(dzd\tau) \quad \text{for } s \le t.$$

Throughout this paper we adopt Einstein's notation on summation. That is, we use repeated scripts to stand for the summation over these scripts.

1.2. Formulation of the optimal control problem and basic assumptions. Consider the following stochastic control system:

$$\begin{aligned} x(t) = x_0 &+ \int_{(0,t]} a(x(s),v(s))ds + \int_{(0,t]} b^i(x(s),v(s))dw^i(s) \\ &+ \int\int_{\mathcal{Z}\times(0,t]} c(x(s-),v(s),z)\widetilde{N}_k(dzds). \end{aligned} \tag{1.1}$$

An admissible control $v(\cdot)$ is defined as a U-valued $(\mathcal{F}_t)$-predictable process such that

$$||v(\cdot)|| =: \sup_{0\le t\le 1} [E|v(t)|^8]^{1/8} < \infty, \tag{1.2}$$

with U being a nonempty subset of $\mathcal{R}^m$. The set of admissible controls $v(\cdot)$ is denoted by U_{ad}. When $U = \mathcal{R}^m$, we write $L^{\infty,8}_{\mathcal{F},p}[[0,1];\mathcal{R}^m]$ for U_{ad}. The terminal constraint is

$$Ef(x_0,x(1)) \in Q \subset \mathcal{R}^k. \tag{1.3}$$

The cost functional is

$$J(v(\cdot),x_0) = E\left[\int_0^1 g(x(s),v(s))ds + h(x_0,x(1))\right]. \tag{1.4}$$

In the above statement, $a(\cdot,\cdot) : \mathcal{R}^n \times U \to \mathcal{R}^n; b^i(\cdot,\cdot) : \mathcal{R}^n \times U \to \mathcal{R}^n, i = 1,\dots,d$; $c(\cdot,\cdot,\cdot) : \mathcal{R}^n \times U \times \mathcal{Z} \to \mathcal{R}^n; f(\cdot,\cdot) =: \{f^i(\cdot,\cdot)\}_1^k : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}^k; g(\cdot,\cdot) : \mathcal{R}^n \times U \to \mathcal{R}$; and $h(\cdot,\cdot) : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$. The optimal control problem is to find a pair $(y_0,u(\cdot)) \in \mathcal{R}^n \times U_{\text{ad}}$ such that (1.1) and (1.3) are satisfied and (1.4) is minimized.

Throughout the paper we make the following assumptions.

Assumption 1. The vector functions $a(x,v), b(x,v) =: \{b^i(x,v)\}_{i=1}^d, c(x,v,z)$, $f(y,x), g(x,v)$, and $h(y,x)$ are twice differentiable in x, and $f(y,x), h(y,x)$ are differentiable in y. They and their derivatives in x or y are continuous in (x,v) or (y,x). The vector functions $a(x,v), b(x,v)$, $f_{y^i}(y,x)$, $f_{x^i}(y,x)$, $g_{x^i}(x,v), h_{y^i}(y,x)$, $h_{x^i}(y,x)$, and

$$\left[\int_{\mathcal{Z}} |c(x,v,z)|^{2k}\pi(dz)\right]^{1/2k}, \qquad k = 1,2$$

$(i = 1,\dots,n)$ are bounded by $(1+|x|+|y|+|v|)$. The vector functions $f(y,x)$, $g(x,v)$, $h(y,x)$ are bounded by $(1+|x|^2+|y|^2+|v|^2)$, $a_{x^i}(x,v), a_{x^ix^j}(x,v), b_{x^i}(x,v), b_{x^ix^j}(x,v)$,

$f_{x^ix^j}(y,x), g_{x^ix^j}(x,v))$, $h_{x^ix^j}(y,x)$, and

$$\int_{\mathcal{Z}} |c_{x^i}(x,v,z)|^{2k}\pi(dz), \quad k=1,2, \quad \int_{\mathcal{Z}} |c_{x^ix^j}(x,v,z)|^2\pi(dz)$$

$(i,j=1,\dots,n)$ are bounded. Here x^i, y^i $(i=1,\dots,n)$ stand for the ith coordinates of x and y, respectively.

Assumption 2. The set Q is closed and convex.

For the given $(x_0, v(\cdot)) \in \mathcal{R}^n \times U_{\text{ad}}$, an $\mathcal{R}^n$-valued process $x(\cdot)$ is called a solution of (1.1) if it is an $(\mathcal{F}_t)$-adapted cadlag (i.e., right-continuous with left-hand limits) process such that (1.1) holds. Under Assumption 1, (1.1) admits a unique solution for the given $x_0 \in \mathcal{R}^n$ and $v(\cdot) \in U_{\text{ad}}$ (cf.[16]), and the above formulation of optimal control problem is well defined.

1.3. Developments of optimal stochastic control and contributions of the paper. Since [11], a number of results have been obtained on optimal stochastic control problems, (cf., for example, [1], [2], [4], [12]–[15], [21], [23]). Two major advances have been made in the last two decades. One is the definition of the adjoint processes and its characterization by Itô-type equations. This was done by Kushner [17] and Bismut [4], and summarized by Bensoussan [2], [3] via functional analysis methods. The other advance, which is well-worth mentioning, was marked by the idea of second-order variation in calculating the variation of the cost functional caused by the spike variation of the given optimal control. This was motivated by the study of the nonconvex optimal stochastic control of diffusion processes with the control entering into the diffusion term, and was developed by Peng [21]. On nonconvex controls of diffusion processes, we refer the reader to Kushner [17], Haussmann [12], Bensoussan [2], Hu [14], Peng [21], and Zhou [24], [25]. In this paper we apply the idea of second-order variation to study the general optimal stochastic control with random jumps. Here, by the word "general" we mean the allowance of the control into both diffusion and jump terms and the nonconvexity property of the set U. We mention that optimal control of jump processes was first considered by Boel [5], Boel and Varaiya [6], Rishel [22], Davis and Elliott [7], and Situ [23]. In this paper, however, we need not assume $c(x,v) \equiv c(x)$. Our result (see Theorem 2.1, below) shows that the maximum principle with random jumps is different from the pure diffusion version.

It is well known that vector-valued measure theory (see [18, Cor. 1]) has been playing a fundamental role in studying maximum principles for deterministic optimal control problems (cf. [18], [19]). For the application of the theory to optimal stochastic control problems, the reader is referred to [14]. However, the context of [14] seems difficult for adaptation even to the case considered in [21]. We also provide in this paper an application of the vector-valued measure theory to general stochastic control problems. It is this application that enables us to obtain the stochastic maximum principle with random jumps (more precisely, this application will solve the "differentiability" problem in an elegant way). In fact, we use only the one-dimentional case of vector-valued measure theory. That is to say, we use only a property for Lebesgue integrals of scalar-valued functions in the real number space $\mathcal{R}$. We remark here that our situation does not use the finite dimensionality of the state space and therefore can be adapted without any essential difficulty to the infinite-dimensional case, which will be the subject of a forthcoming paper. All these results show that the proof of maximum principle is almost identical and can be completed by means of very elementary tools whether the control system is finite, infinite dimensional, deterministic, or stochastic.

In this paper we also consider the optimal stochastic control with a general terminal constraint.

The rest of the paper is organized as follows. Section 2 contains some preliminary lemmas and the main result, §3 contains the main proof, and §4 contains a conclusion.

2. Preliminary lemmas and the main result.

Notation. Let $(y_0, y(\cdot), u(\cdot))$ be an optimal triplet. For the given $(x_0, v(\cdot)) \in \mathcal{R}^n \times U_{\text{ad}}$, write $y(\cdot; v(\cdot), x_0)$ for the solution of (1.1). For $v, v_1, v_2 \in U$, denote

$$
\begin{aligned}
\Delta m(s; v_2, v_1) &=: m(y(s-), v_2) - m(y(s-), v_1),\\
\Delta m(s; v) &=: m(y(s-), v) - m(y(s-), u(s)),\\
m(s; v_1) &=: m(y(s), v_1),\\
m(s) &=: m(y(s), u(s)),\\
\Delta n(s, z; v_2, v_1) &=: n(y(s-), v_2, z) - n(y(s-), v_1, z),\\
\Delta n(s, z; v) &=: n(y(s-), v, z) - n(y(s-), u(s), z),\\
n(s, z; v_1) &=: n(y(s-), v_1, z),\\
n(s, z) &=: n(y(s-), u(s), z),
\end{aligned}
\tag{2.1}
$$

with m standing for a, b, g and all their (up to second-) derivatives in x, and n for c and its (up to second-) derivatives in x.

For $I_0 \subset [0,1]$, let $|I_0|$ denote the Lebesgue measure of the set I_0. Let $v(\cdot)$, $v_1(\cdot)$, $v_2(\cdot) \in U_{\text{ad}}$. Define

$$
\hat{d}(v_1(\cdot), v_2(\cdot)) =: |\{t \in [0,1]; E|v_1(t) - v_2(t)|^2 > 0\}|. \tag{2.2}
$$

For $\rho \in (0,1]$, $I_\rho \subset [0,1]$, and $v(\cdot) \in U_{\text{ad}}$, define

$$
\begin{aligned}
u^\rho(s) &=: u(s)\chi_{[0,1]\setminus I_\rho}(s) + v(s)\chi_{I_\rho}(s), \qquad s \in [0,1],\\
y_0^\rho &=: y_0 + |I_\rho|\eta, \qquad \eta \in \mathcal{R}^n,\\
y^\rho(\cdot) &= y(\cdot; u^\rho(\cdot), y_0^\rho),
\end{aligned}
\tag{2.3}
$$

with $\chi_A(\cdot)$ denoting the indicator function of some set A. Obviously, we have

$$
\hat{d}(u_\rho(\cdot), u(\cdot)) = |I_\rho|. \tag{2.4}
$$

We can prove that $u^\rho(\cdot) \in U_{\text{ad}}$.

Let the process $y_1(t; v_2(\cdot), v_1(\cdot))$ be the solution of

$$
\begin{aligned}
y_1(t) = &\int_{(0,t]} a_x(s; v_1(s)) y_1(s) ds\\
&+ \int_{(0,t]} [b_x^k(s; v_1(s)) y_1(s) + \Delta b^k(s; v_2(s), v_1(s))] dw^k(s)\\
&+ \int\int_{\mathcal{Z}\times(0,t]} [c_x(s, z; v_1(s)) y_1(s-) + \Delta c(s, z; v_2(s), v_1(s))] \widetilde{N}_k(dzds),
\end{aligned}
\tag{2.5}
$$

and the process $y_2(t; v_2(\cdot), v_1(\cdot))$ be the solution of

$$
\begin{aligned}
y_2(t) = {} & \hat{d}(v_2(\cdot), v_1(\cdot))\eta \\
& + \int_{(0,t]} \Big[a_x(s; v_1(s)) y_2(s) + \Delta a(s; v_2(s), v_1(s)) \\
& \qquad + \frac{1}{2} a_{x^i x^j}(s; v_1(s)) y_1^i(s) y_1^j(s) \Big] ds \\
& + \int_{(0,t]} \Big[b_x^k(s; v_1(s)) y_2(s) + \Delta b_x^k(s; v_2(s), v_1(s)) y_1(s) \\
& \qquad + \frac{1}{2} b_{x^i x^j}^k(s; v_1(s)) y_1^i(s) y_1^j(s) \Big] dw^k(s) \\
& + \int\int_{\mathcal{Z} \times (0,t]} \Big[c_x(s, z; v_1(s)) y_2(s-) + \Delta c_x(s, z; v_2(s), v_1(s)) y_1(s-) \\
& \qquad + \frac{1}{2} c_{x^i x^j}(s, z; v_1(s)) y_1^i(s-) y_1^j(s-) \Big] \tilde{N}_k(dz ds),
\end{aligned} \tag{2.6}
$$

where $y_1^i(\cdot)$ $(i = 1, \ldots, n)$ stand for the ith coordinate of vector $y_1(\cdot)(=: y_1(\cdot; v_1(\cdot), v_2(\cdot)))$. Write $y_i(t; v(\cdot))$ for $y_i(t; v(\cdot), u(\cdot))$, $i = 1, 2$.

Remark 2.1. The definitions of $y_1(t; v(\cdot))$ and $y_2(t; v(\cdot))$ are different from those in [21]. These changes have the following advantage: $y_1(t; v(\cdot))$ represents the half-order component and $y_2(t; v(\cdot))$ represents the first-order component of the variation for the state; when the control appears in neither diffusion nor jump terms, $y_1(t; v(\cdot))$ vanishes in an automatic way and $y_2(t; v(\cdot))$ is the usual first-order component of the variation.

We have the following estimates, which play a crucial role in calculating the variation of the cost.

LEMMA 2.1. *Assume that Assumption* 1 *is satisfied. Then for* $v(\cdot), v_i(\cdot) \in U_{\text{ad}}, i = 1, 2, x_0 \in \mathcal{R}^n$, *we have*

$$
\begin{aligned}
& \sup_{0 \le t \le 1} E|y(t; v(\cdot), x_0)|^8 = O((1 + ||v(\cdot)||)^8), \\
& \sup_{0 \le t \le 1} E|y(t; v_1(\cdot), x_0) - y(t; v_2(\cdot), x_0)|^4 = O(\hat{d}^2(v_2(\cdot), v_1(\cdot))(1 + ||v_1(\cdot)|| + ||v_2(\cdot)||)^4), \\
& \sup_{0 \le t \le 1} E|y_1(t; v_2(\cdot), v_1(\cdot))|^8 = O(\hat{d}^4(v_2(\cdot), v_1(\cdot))(1 + ||v_1(\cdot)|| + ||v_2(\cdot)||)^8), \\
& \sup_{0 \le t \le 1} E|y_2(t; v_2(\cdot), v_1(\cdot))|^4 = O(\hat{d}^4(v_2(\cdot), v_1(\cdot))(1 + ||v_1(\cdot)|| + ||v_2(\cdot)||)^8), \\
& \sup_{0 \le t \le 1} E|y(t; v_2, y_0 + \hat{d}(v_i, v_1)\eta) - y(t; v_1, y_0) - y_1(t; v_2, v_1) - y_2(t; v_2, v_1)|^2 \\
& \qquad = o(\hat{d}^2(v_2(\cdot), v_1(\cdot))(1 + ||v_1(\cdot)|| + ||v_2(\cdot)||)^8), \quad \text{as } \hat{d}(v_2(\cdot), v_1(\cdot)) \to 0.
\end{aligned} \tag{2.7}
$$

Remark 2.2. Assume that Assumption 1 is satisfied. Then for fixed $v(\cdot) \in U_{\text{ad}}$,

we have

$$\begin{aligned} &\sup_{0\le t\le 1} E|y_1(t)|^8 = O(|I_\rho|^4), \\ &\sup_{0\le t\le 1} E|y_2(t)|^4 = O(|I_\rho|^4), \\ &\sup_{0\le t\le 1} E|y^\rho(t) - y(t) - y_1(t) - y_2(t)|^2 = o(|I_\rho|^2), \quad \text{as } |I_\rho| \to 0. \end{aligned} \tag{2.8}$$

Here $y_i(t) =: y_i(t; u^\rho(\cdot)), i = 1, 2$.

Remark 2.3. When $I_\rho = [t, t+\rho] \subset [0,1]$ and $c \equiv 0$, Remark 2.2 is Lemma 1 of [21]. We point out that a misprint was made in the statement of Lemma 1 in [21]; however, our Remark 2.2 for the above-mentioned case is what was actually proved in [21].

Remark 2.4. Note that the first two estimates (2.7) are concerned with the boundedness and the continuity of the system state $y(\cdot; v(\cdot), x_0)$, as a functional (x_0 is fixed) on the metric space $(U_{\text{ad}}, \hat{d})$, and they depend on the upper bounds of $||v_1(\cdot)||+||v_2(\cdot)||$. When U_{ad} is bounded in $L^{\infty,8}_{\mathcal{F},p}[[0,1]; \mathcal{R}^m]$, they imply the boundedness and the continuity in $(U_{\text{ad}}, \hat{d})$ of $y(\cdot; v(\cdot), x_0)$ with respect to $v(\cdot)$; then the same boundedness and continuity is possessed by the cost functional since the cost can be viewed as a component of the state in the augmented system.

Proof. Without loss of generality, we assume that

$$\eta = 0, \quad v_1(\cdot) = u(\cdot), \quad v_2(\cdot) = u^\rho(\cdot). \tag{2.9}$$

Define

$$\begin{aligned} \int_{I_\rho} f_0(s)\, dw(s) &=: \int_0^1 \chi_{I_\rho}(s) f_0(s)\, dw(s), \\ \iint_{\mathcal{Z}\times I_\rho} g_0(s,z)\, \widetilde{N}_k(dz\, ds) &=: \iint_{\mathcal{Z}\times(0,1]} \chi_{I_\rho}(s) g_0(s,z)\, \widetilde{N}_k(dz\, ds). \end{aligned}$$

We have the following inequalities:

$$\begin{aligned} E\left|\int_{I_\rho} f_0(s)\, ds\right|^p &\le C_p |I_\rho|^{p-1} E \int_{I_\rho} |f_0(s)|^p\, ds, \\ E\left|\int_{I_\rho} f_0(s)\, dw(s)\right|^{2p} &\le C_p |I_\rho|^{p-1} E \int_{I_\rho} |f_0(s)|^{2p}\, ds, \\ E\left|\iint_{\mathcal{Z}\times I_\rho} g_0(s,z)\, \widetilde{N}_k(dz\, ds)\right|^{2p} &\le C_p |I_\rho|^{p-1} E \int_{I_\rho} \left|\int_{\mathcal{Z}} |g_0(s,z)|^2\, \pi(dz)\right|^p ds, \end{aligned} \tag{2.10}$$

with $p > 1$.

By virtue of the Assumption 1, we have

$$
\begin{aligned}
&\sup_{0\le t\le 1} E|y(t)|^8 = O((1+||v(\cdot)||+||u(\cdot)||)^8),\\
&\sup_{0\le t\le 1} E|\Delta a(t;u^\rho(s))|^4 = O((1+||v(\cdot)||+||u(\cdot)||)^4),\\
&\sup_{0\le t\le 1} E|\Delta b(t;u^\rho(s))|^8 = O((1+||v(\cdot)||+||u(\cdot)||)^8),\\
&\sup_{0\le t\le 1} E\left|\int_{\mathcal{Z}} |\Delta c(t,z;u^\rho(s))|^2\,\pi(dz)\right|^4 = O((1+||v(\cdot)||+||u(\cdot)||)^8).
\end{aligned}
\tag{2.11}
$$

Then we can obtain the following inequalities by using (2.10):

$$
\begin{aligned}
&E\left|\int_0^1 \Delta a(s;u^\rho(s))\,ds\right|^4 = O(|I_\rho|^4(1+||v(\cdot)||+||u(\cdot)||)^4),\\
&E\left|\int_0^1 \Delta b(s;u^\rho(s))\,dw(s)\right|^8 = O(|I_\rho|^4(1+||v(\cdot)||+||u(\cdot)||)^8),\\
&E\left|\iint_{\mathcal{Z}\times(0,1]} \Delta c(s,z;u^\rho(s))\,\tilde N_k(dz\,ds)\right|^8 = O(|I_\rho|^4(1+||v(\cdot)||+||u(\cdot)||)^8).
\end{aligned}
\tag{2.12}
$$

Then the first four estimates of (2.7) are easily proved by using the familiar elementary inequalities

$$(m_1+m_2+m_3)^i \le C(|m_1|^i+|m_2|^i+|m_3|^i), \qquad i=4,8,$$

and the well-known Gronwall's inequality.

The proof for the last estimate follows. Set $y_3 = y_1 + y_2$. We have

$$
\begin{aligned}
&\int_0^t a(y+y_3,u^\rho)ds + \int_0^t b^k(y+y_3,u^\rho)dw^k(s)\\
&\quad + \iint_{\mathcal{Z}\times(0,t]} c(y+y_3,u^\rho,z)\tilde N_k(dzds)\\
&= \int_0^t \left[a(y,u^\rho) + a_x(y,u^\rho)y_3 + \int_0^1\int_0^1 \lambda a_{x^ix^j}(y+\lambda\mu y_3,u^\rho)d\lambda\,d\mu y_3^i y_3^j\right] ds\\
&\quad + \int_0^t \Bigg[b^k(y,u^\rho) + b^k_x(y,u^\rho)y_3\\
&\qquad + \int_0^1\int_0^1 \lambda b^k_{x^ix^j}(y+\lambda\mu y_3,u^\rho)d\lambda\,d\mu y_3^i y_3^j\Bigg] dw^k(s)\\
&\quad + \iint_{\mathcal{Z}\times(0,t]} \Bigg[c(y,u^\rho,z) + c_x(y,u^\rho,z)y_3\\
&\qquad + \int_0^1\int_0^1 \lambda c_{x^ix^j}(y+\lambda\mu y_3,u^\rho,z)d\lambda\,d\mu y_3^i y_3^j\Bigg] \tilde N_k(dz\,ds)\\
&= y(t)+y_3(t)-y_0 + \int_0^t A(s;\rho)\,ds + \int_0^t B^k(s;\rho)\,dw^k(s)\\
&\quad + \iint_{\mathcal{Z}\times(0,t]} C(s,z;\rho)\,\tilde N_k(dz\,ds),
\end{aligned}
$$

where

$$\begin{aligned}A(s;\rho) =&\frac{1}{2}a_{x^ix^j}(s)(y_2^i(s)y_2^j(s)+2y_1^i(s)y_2^j(s))\\&+\Delta a_x(s;u^\rho(s))y_3(s)\\&+\int_0^1\int_0^1\lambda[a_{x^ix^j}(y+\lambda\mu y_3,u^\rho)-a_{x^ix^j}(s)]d\lambda\,d\mu\,y_3^i(s)y_3^j(s),\\B(s;\rho) =&\frac{1}{2}b_{x^ix^j}(s)(y_2^i(s)y_2^j(s)+2y_1^i(s)y_2^j(s))\\&+\Delta b_x(s;u^\rho(s))y_2(s)\\&+\int_0^1\int_0^1\lambda[b_{x^ix^j}(y+\lambda\mu y_3,u^\rho)-b_{x^ix^j}(s)]d\lambda\,d\mu\,y_3^i(s)y_3^j(s),\\C(s,z;\rho) =&\frac{1}{2}c_{x^ix^j}(s,z)(y_2^i(s)y_2^j(s)+2y_1^i(s)y_2^j(s))\\&+\Delta c_x(s,z;u^\rho(s))y_2(s)\\&+\int_0^1\int_0^1\lambda[c_{x^ix^j}(y+\lambda\mu y_3,u^\rho,z)-c_{x^ix^j}(s,z)]d\lambda\,d\mu\,y_3^i(s)y_3^j(s).\end{aligned}$$

We can derive that

$$\begin{aligned}(y^\rho-y-y_3)(t)=&\int_0^t[\widetilde{A}(s;\rho)(y^\rho-y-y_3)(s)+A(s;\rho)]\,ds\\&+\int_0^t[\widetilde{B}^k(s;\rho)(y^\rho-y-y_3)(s)+B^k(s;\rho)]\,dw^k(s)\\&+\iint_{\mathcal{Z}\times(0,t]}[\widetilde{C}(s,z;\rho)(y^\rho-y-y_3)(s)+C(s,z;\rho)]\,\widetilde{N}_k(dz\,ds),\end{aligned}$$

$$|\widetilde{A}(s;\rho)|+|\widetilde{B}^k(s;\rho)|+|\widetilde{C}(s,\cdot;\rho)|\le C,$$

$$\begin{aligned}&\sup_{0\le t\le 1}E\left(\left|\int_0^t A(s;\rho)ds\right|^2+\left|\int_0^t B^k(s;\rho)dw^k(s)\right|^2+\left|\iint_{\mathcal{Z}\times(0,t]}C(s,z;\rho)\widetilde{N}_k(dzds)\right|^2\right)\\&\quad=o(|I_\rho|^2(1+||v(\cdot)||+||u(\cdot)||^8).\end{aligned}$$

From these we can use Ito's formula and Gronwall's inequality to obtain the fifth estimate (2.7). □

LEMMA 2.2. *Assume that $l(\cdot)$ is a scalar-valued Lebesgue integrable function defined on $[0,1]$. Then for $\rho\in(0,1]$, there exists a measurable subset $I_\rho\subset[0,1]$ such that*

$$\begin{aligned}&|I_\rho|=\rho,\\&\int_{I_\rho}l(s)ds=\rho\int_{[0,1]}l(s)ds+o(\rho)\quad as\ \rho\to 0.\end{aligned}\tag{2.13}$$

The proof is quite elementary and the reader is referred to [18]. We would like to mention that a stronger statement (more precisely, this lemma with the term $o(\rho)$ vanishing) is available via an application of the well-known Liapunov convexity theorem to the $\mathcal{R}^2$-valued integrable vector function $(l(\cdot),1)$. However, this lemma is enough for our argument below.

LEMMA 2.3 (Martingale representation). *Let $\mathcal{H}$ be a finite-dimensional space and let $m(t)$ be an $\mathcal{H}$-valued $(\mathcal{F}_t)$-adapted square-integrable Martingale. Then there exist $q^i(\cdot) \in L^2_{\mathcal{F},p}[(0,1);\mathcal{H}]$ $(i = 1, \dots, d)$ and $r(\cdot,\cdot) \in F^2_p[(0,1);\mathcal{H}]$ such that*

$$m(t) = m(0) + \int_{(0,t]} q^i(s)dw^i(s) + \int\int_{\mathcal{Z}\times(0,t]} r(s,z)\widetilde{N}_k(dzds).$$

The lemma is a combination of [16, Chap. 2, Thms. 6.6–6.7], and its proof is given in the Appendix.

Consider the following Itô-type stochastic integral equation with terminal conditions

$$\begin{aligned} p(t) = p_1 &+ \int_{(t,1]} \hat{a}(s,\cdot,p(s),q(s),r_s(\cdot))ds - \int_{(t,1]} q(s)dw(s) \\ &- \int\int_{\mathcal{Z}\times(t,1]} r(s,z)\,\widetilde{N}_k(dz\,ds), \qquad 0 \le t \le 1. \end{aligned} \tag{2.14}$$

Here p_1 is an $\mathcal{H}$-valued square integrable random variable on $(\Omega,\mathcal{F}_1)$, and $\hat{a}(\cdot,\cdot,\cdot,\cdot,\cdot) : [0,1]\times\Omega\times\mathcal{H}\times L(\mathcal{R}^d,\mathcal{H})\times F^2_\pi[\mathcal{Z};\mathcal{H}] \to \mathcal{H}$ whose $a(t,\cdot,p,q,r)$ is $\mathcal{F}_t$-measurable for given $(t,p,q,r) \in [0,1]\times\mathcal{H}\times L(\mathcal{R}^d,\mathcal{H})\times F^2_\pi[\mathcal{Z};\mathcal{H}]$ and which satisfies the following:

$$\begin{aligned} &|\hat{a}(t,\omega,0,0,0)| \le C, \\ &|\hat{a}(t,\omega,p_1,q_1,r_1) - \hat{a}(t,\omega,p_2,q_2,r_2)| \le C(|p_1-p_2| + |q_1-q_2| + |r_1-r_2|), \end{aligned} \tag{2.15}$$

with the constant C being independent of $(t,\omega) \in [0,1]\times\Omega$. Based on Lemma 2.3, we have the following lemma.

LEMMA 2.4. *Let (2.15) hold. Then there exists unique*

$$(p(\cdot),q(\cdot),r(\cdot,\cdot)) \in L^2_{\mathcal{F}}[[0,1];\mathcal{H}]\times L^2_{\mathcal{F},p}[[0,1];L(\mathcal{R}^d,\mathcal{H})]\times F^2_p[[0,1];\mathcal{H}],$$

with $p(\cdot)$ being a cadlag process, which solves (2.14).

Proof. For each

$$(\bar{p}(\cdot),\bar{q}(\cdot),\bar{r}(\cdot,\cdot)) \in L^2_{\mathcal{F}}[[0,1];\mathcal{H}]\times L^2_{\mathcal{F},p}[[0,1];L(\mathcal{R}^d,\mathcal{H})]\times F^2_p[[0,1];\ \mathcal{H}],$$

we know from Lemma 2.3 that there exist $q(\cdot) \in L^2_{\mathcal{F},p}[[0,1];L(\mathcal{R}^d,\mathcal{H})]$, $r(\cdot,\cdot)) \in F^2_p[[0,1];\mathcal{H}]$, such that

$$\begin{aligned} &E^{\mathcal{F}_t}[p_1 + \int_{(0,1]} \hat{a}(s,\cdot,\bar{p}(s),\bar{q}(s),\bar{r}(s,\cdot))ds] \\ &\quad = X + \int_{(0,t]} q(s)dw(s) + \int\int_{\mathcal{Z}\times(0,t]} r(s,z)\,\widetilde{N}_k(dzds). \end{aligned}$$

This implies

$$\begin{aligned} X = p_1 &+ \int_{(0,1]} \hat{a}(s,\cdot,\bar{p}(s),\bar{q}(s),\bar{r}(s,\cdot))ds - \int_{(0,1]} q(s)dw(s) \\ &- \int\int_{\mathcal{Z}\times(0,1]} r(s,z)\widetilde{N}_k(dzds). \end{aligned}$$

Set

$$p(t) =: E^{\mathcal{F}_t}\left[p_1 + \int_{(t,1]} \hat{a}(s,\cdot,\bar{p}(s),\bar{q}(s),\bar{r}(s,\cdot))ds\right].$$

We verify that for given triplet $(\bar{p}(\cdot), \bar{q}(\cdot), \bar{r}(\cdot,\cdot))$, the corresponding triplet $(p(\cdot), q(\cdot), r(\cdot,\cdot))$ is characterized by the equation

$$\begin{aligned} p(t) = & p_1 + \int_{(t,1]} \hat{a}(s,\cdot,\bar{p}(s),\bar{q}(s),\bar{r}(s,\cdot))ds \\ & - \int_{(t,1]} q(s)dw(s) - \int\int_{\mathcal{Z}\times(t,1]} r(s,z)q(dzds). \end{aligned} \tag{2.16}$$

This implies

$$\begin{aligned} p(t) = & p(0) - \int_{(0,t]} \hat{a}(s,\cdot,\bar{p}(s),\bar{q}(s),\bar{r}(s,\cdot))ds \\ & + \int_{(0,t]} q(s)dw(s) + \int\int_{\mathcal{Z}\times(0,t]} r(s,z)\widetilde{N}_k(dzds). \end{aligned} \tag{2.17}$$

Equation (2.16) defines a map $\Lambda : (\bar{p}(\cdot), \bar{q}(\cdot), \bar{r}(\cdot,\cdot)) \to (p(\cdot), q(\cdot), r(\cdot,\cdot))$. Let $D^2_{\mathcal{F}}(0,1;\mathcal{H})$ be the Banach space of $\mathcal{H}$-valued $(\mathcal{F}_t)$-adapted cadlag processes $\hat{p}(t)$ such that

$$\sup_{0\le t\le 1} E|\hat{p}(t)|^2 < \infty.$$

We introduce, for $k =: (p(\cdot), q(\cdot), r(\cdot,\cdot)) \in D^2_{\mathcal{F}}(0,1;\mathcal{H}) \times L^2_{\mathcal{F},p}[[0,1]; L(\mathcal{R}^d,\mathcal{H})] \times F^2_p[[0,1];\mathcal{H}]$, the norm defined by

$$\begin{aligned} ||k|| =: & \sup_{0\le t\le 1} e^{bt} E|p(t)|^2 \\ & + \sup_{0\le t\le 1} e^{bt} \left[\int_t^1 E|q(s)|^2 ds + \int\int_{\mathcal{Z}\times(t,1]} E|r(s,z)|^2 \pi(dz)ds \right], \end{aligned} \tag{2.18}$$

with $b > 0$ to be determined later. To complete the proof, it is enough to show that Λ maps $D^2_{\mathcal{F}}(0,1;\mathcal{H}) \times L^2_{\mathcal{F},p}[[0,1]; L(\mathcal{R}^d,\mathcal{H})] \times F^2_p[[0,1];\mathcal{H}]$ into itself and is a contraction under the norm (2.18). For this purpose, let $(\bar{p}_i(\cdot), \bar{q}_i(\cdot), \bar{r}_i(\cdot,\cdot)) \in D^2_{\mathcal{F}}(0,1;\mathcal{H}) \times L^2_{\mathcal{F},p}[[0,1]; L(\mathcal{R}^d,\mathcal{H})] \times F^2_p[[0,1];\mathcal{H}]$ and $(p_i(\cdot), q_i(\cdot), r_i(\cdot,\cdot)) =: \Lambda(\bar{p}_i(\cdot), \bar{q}_i(\cdot), \bar{r}_i(\cdot,\cdot))$ for $i = 1, 2$. Then, using Itô's formula [10], we have from (2.14)–(2.16) that

(2.19)

$$\begin{aligned} & E|p_1(t) - p_2(t)|^2 + E\int_t^1 \sum_{i=1}^d |q_1^i(s) - q_2^i(s)|^2 ds \\ & \qquad + E\int\int_{\mathcal{Z}\times(t,1]} |r_1(s,z) - r_2(s,z)|^2 \pi(dz)ds \\ & \le \hat{\gamma} C^2 E \int_{(t,1]} |p_1(s) - p_2(s)|^2 ds + \frac{1}{\hat{\gamma}} \Bigg[E\int_{(t,1]} |\bar{p}_1(s) - \bar{p}_2(s)|^2 ds \\ & \qquad\qquad + E\int_{(t,1]} \sum_{i=1}^d |\bar{q}_1^i(s) - \bar{q}_2^i(s)|^2 ds \\ & \qquad\qquad + E\int\int_{\mathcal{Z}\times(t,1]} |\bar{r}_1(s,z) - \bar{r}_2(s,z)|^2 \pi(dz)ds \Bigg]. \end{aligned}$$

This implies from Gronwall's inequality that

$$
\begin{aligned}
&E|p_1(t)-p_2(t)|^2+E\int_t^1\sum_{i=1}^d|q_1^i(s)-q_2^i(s)|^2ds\\
&\qquad+E\int\int_{\mathcal{Z}\times(t,1]}|r_1(s,z)-r_2(s,z)|^2\pi(dz)ds\\
&\le\frac{1}{\hat\gamma}\Bigg[E\int_{(t,1]}|\bar p_1(s)-\bar p_2(s)|^2ds+E\int_{(t,1]}\sum_{i=1}^d|\bar q_1^i(s)-\bar q_2^i(s)|^2ds\\
&\qquad+E\int\int_{\mathcal{Z}\times(t,1]}|\bar r_1(s,z)-\bar r_2(s,z)|^2\pi(dz)ds\Bigg]\\
&\quad+C^2\int_t^1e^{\hat\gamma C^2(s-t)}\Bigg[E\int_{(s,1]}|\bar p_1(\tau)-\bar p_2(\tau)|^2d\tau\\
&\qquad+E\int_{(s,1]}\sum_{i=1}^d|\bar q_1^i(\tau)-\bar q_2^i(\tau)|^2d\tau\\
&\qquad+E\int\int_{\mathcal{Z}\times(s,1]}|\bar r_1(\tau,z)-\bar r_2(\tau,z)|^2\pi(dz)d\tau\Bigg]ds,
\end{aligned}
\tag{2.20}
$$

with $\hat\gamma$ being any positive real number. Noting (2.18), we conclude that

$$
\begin{aligned}
&||(p_1-p_2,q_1-q_2,r_1-r_2)||\\
&\quad\le\max\left\{\frac{2}{b\hat\gamma},\frac{2}{\hat\gamma},\frac{2C^2}{b(b-\hat\gamma C^2)},\frac{2C^2}{b-\hat\gamma C^2}\right\}||(\bar p_1-\bar p_2,\bar q_1-\bar q_2,\bar r_1-\bar r_2)||,
\end{aligned}
\tag{2.21}
$$

which completes the proof by choosing appropriate $\hat\gamma$ and b. □

Remark 2.5. We note that a special case of Lemma 2.4 is obtained in [20].

The Hamiltonian is defined as

$$
\begin{aligned}
H(x,v,\lambda,p,\{q^i\}_1^d,r(\cdot))=&\lambda g(x,v)+\langle p,a(x,v)\rangle+\langle q^i,b^i(x,v)\rangle\\
&+\int_{\mathcal{Z}}\langle r(z),c(x,v,z)\rangle\pi(dz);
\end{aligned}
$$

this is a map from $\mathcal{R}^n\times U\times\mathcal{R}\times\mathcal{R}^n\times\mathcal{R}^{n\times d}\times F_\pi^2[\mathcal{Z};\mathcal{R}^n]$ into $\mathcal{R}$. Here we have used $\langle,\rangle$ for the scalar product of Euclidean spaces.

From Lemma 2.3 and Assumption 1 we see for the given $p(1)\in L^2(\Omega,\mathcal{F}_1;\mathcal{R}^n)$, $P(1)\in L^2(\Omega,\mathcal{F};\mathcal{R}^{n\times n})$ that the Itô-type adjoint equations

$$
\begin{aligned}
p(t)=&p(1)+\int_{(t,1]}H_x(y(s),u(s),\lambda,p(s),\{q^i(s)\}_1^d,r(s,\cdot))\,ds\\
&-\int_{(t,1]}q^i(s)dw^i(s)-\int\int_{\mathcal{Z}\times(t,1]}r(s,z)\,\tilde N_k(dz\,ds)
\end{aligned}
\tag{2.22}
$$

and

$$
\begin{aligned}
P(t) = P(1) + \int_{(t,1]} \Big\{ & a_x^*(s)P(s) + P(s)a_x(s) \\
& + b_x^{i*}(s)P(s)b_x^i(s) + b_x^{i*}(s)Q^i(s) + Q^i(s)b_x^i(s) \\
& + \int_{\mathcal{Z}} c_x^*(s,z)P(s)c_x(s,z)\,\pi(dz) \\
& + \int_{\mathcal{Z}} [c_x^*(s,z)R(s,z)c_x(s,z) \\
& \qquad + c_x^*(s,z)R(s,z) + R(s,z)c_x(s,z)]\pi(dz) \\
& + H_{xx}(y(s),u(s),\lambda,p(s),\{q^i(s)\}_1^d,r(s,\cdot))\Big\}\,ds \\
& - \int_{(t,1]} Q^i(s)dw^i(s) - \int\int_{\mathcal{Z}\times(t,1]} R(s,z)\,\widetilde{N}_k(dz\,ds)
\end{aligned} \tag{2.23}
$$

admit unique solutions $(p(\cdot), \{q^i(\cdot)\}_{i=1}^d, r(\cdot,\cdot))$ and $(P(\cdot), \{Q^i(\cdot)\}_{i=1}^d, R(\cdot,\cdot))$, with $p(\cdot)$ and $P(\cdot)$ being cadlag processes.

Define the following function:

$$
\Phi(s,z;\varepsilon) =: \inf_{(\bar{z},t)\in Q\times(-\infty,J(u(\cdot),y_0)-\varepsilon]} \sqrt{|\bar{z}-z|^2 + |s-t|^2}. \tag{2.24}
$$

LEMMA 2.5. *For given $\varepsilon > 0$, the function $\Phi(s,z;\varepsilon)$ is continuously differentiable on the open set $\hat{Q} =: \{(s,z) : \Phi(s,z;\varepsilon) > 0\}$. Moreover, when $\Phi(s,z;\varepsilon) > 0$, we have*

$$
\begin{aligned}
\langle \Phi_z(s,z;\varepsilon), \hat{z} - z\rangle &\le 0, \qquad \forall\, \hat{z} \in Q, \\
\Phi_s(s,z;\varepsilon) &\ge 0, \\
|\Phi_s(s,z;\varepsilon)|^2 + |\Phi_z(s,z;\varepsilon)|^2 &= 1.
\end{aligned} \tag{2.25}
$$

Proof. We easily see that the distance function $\Phi(\cdot,\cdot;\varepsilon)$ is Lipschitz with Lipschitz constant 1 and we derive from Lemma 3.4 and Corollary 3.5 of [19] that $\Phi(s,z;\varepsilon)$ is continuously differentiable at (s,z), and further, that

$$
\partial\Phi(s,z;\varepsilon) = \{D\Phi(s,z;\varepsilon)\}, \tag{2.26}
$$

$$
|D\Phi(s,z;\varepsilon)|^2_{\mathcal{R}\times\mathcal{R}^k} = 1, \tag{2.27}
$$

whenever $\Phi(s,z;\varepsilon) > 0$. From the definition of $\partial\Phi(s,z;\varepsilon)$ (see [19]) and (2.24), we have

$$
\langle D\Phi(s,z;\varepsilon), (\hat{s},\hat{z}) - (s,z)\rangle \le 0, \qquad \forall (\hat{s},\hat{z}) \in (-\infty, J(u(\cdot),y_0) - \varepsilon] \times Q. \tag{2.28}
$$

This implies

$$
\langle \Phi_s(s,z;\varepsilon), \hat{s} - s\rangle \le 0, \qquad \forall \hat{s} \in (-\infty, J(u(\cdot),y_0) - \varepsilon] \tag{2.29}
$$

and

$$
\langle \Phi_z(s,z;\varepsilon), \hat{z} - z\rangle \le 0, \qquad \forall \hat{z} \in Q. \tag{2.30}
$$

The last two relations (2.25) follow from (2.29) and (2.27), respectively. □

We introduce the smooth function $\alpha(\cdot)$ defined by

$$\alpha(t,z) =: \begin{cases} C\exp(t^2+|z|^2-1)^{-1}, & t^2+|z|^2<1, \\ 0, & t^2+|z|^2 \geq 1. \end{cases}$$

Choose the constant C such that

$$\int_{\mathcal{R}\times\mathcal{R}^k} \alpha(t,z)dtdz = 1.$$

Set

$$\alpha_\delta(t,z) = \delta^{-(k+1)}\alpha\left(\frac{t}{\delta},\frac{z}{\delta}\right).$$

We define the smooth approximations $\Psi(\cdot,\cdot;\varepsilon,\delta)$ of $\Phi(\cdot,\cdot;\varepsilon)$ as follows:

$$\Psi(s,z;\varepsilon,\delta) =: \int_{\mathcal{R}\times\mathcal{R}^k} \Phi(s-\bar{s},z-\bar{z};\varepsilon)\alpha_\delta(\bar{s},\bar{z})d\bar{s}d\bar{z}. \tag{2.31}$$

Then we easily have

$$0 \leq \Psi(J(u(\cdot),y_0),Ef(y_0,y(1));\varepsilon,\delta) \leq \varepsilon+\sqrt{2}\delta. \tag{2.32}$$

Moreover, we have the following lemma.

LEMMA 2.6. *For $\hat{Q}$ defined in Lemma 2.5, we have for $(s,z) \in \hat{Q}$,*

$$\begin{aligned} &\lim_{\delta\to 0+} \Psi_s(s,z;\varepsilon,\delta) = \Phi_s(s,z;\varepsilon), \\ &\lim_{\delta\to 0+} \Psi_z(s,z;\varepsilon,\delta) = \Phi_z(s,z;\varepsilon). \end{aligned} \tag{2.33}$$

Our main result is the following theorem.

THEOREM 2.1. *Assume Assumptions 1 and 2 hold. Let $(y_0,y(\cdot),u(\cdot))$ be an optimal triplet. Then there exist $0 \leq \lambda \in \mathcal{R}, \mu =: \{\mu^i\}_1^k \in \mathcal{R}^k, p(\cdot) \in L^2_{\mathcal{F}}[[0,1];\mathcal{R}^n], \{q^i(\cdot)\}_1^d \in L^2_{\mathcal{F},p}[[0,1];\mathcal{R}^{n\times d}]$, $r(\cdot,\cdot) \in F^2_p[[0,1];\mathcal{R}^n]$, and $P(\cdot) \in L^2_{\mathcal{F}}[[0,1];\mathcal{R}^{n\times n}], \{Q^i(\cdot)\}_1^d \in L^2_{\mathcal{F},p}[[0,1]; L(\mathcal{R}^d,\mathcal{R}^{n\times n})]$, $R(\cdot,\cdot) \in F^2_p[[0,1];\mathcal{R}^{n\times n}]$ such that we have the following.*

1) *The nontrivial condition*

$$|\lambda|^2+|\mu|^2 = 1 \tag{2.34}$$

is satisfied.

2) *The Itô-type adjoint equations* (2.22), (2.23), *as well as*

$$\begin{aligned} p(1) =& \lambda h_x(y_0,y(1)) + \mu^j f^j_x(y_0,y(1)), \\ p(0) =& -\lambda Eh_y(y_0,y(1)) - \mu^j Ef^j_y(y_0,y(1)), \end{aligned} \tag{2.35}$$

and

$$P(1) = \lambda h_{xx}(y_0,y(1)) + \mu^j f^j_{xx}(y_0,y(1)), \tag{2.36}$$

are satisfied, with $p(\cdot)$ and $P(\cdot)$ being cadlag processes.

3) *The following maximum condition holds:*

(2.37)
$$H(y(s-),v,\lambda,p(s-),\{q^i(s)\}_1^d,r(s,\cdot))-H(y(s-),u(s),\lambda,p(s-),\{q^i(s)\}_1^d,r(s,\cdot))$$
$$+\frac{1}{2}\text{trace } P(s-)\left[\Delta b^i(s;v)\Delta b^{i*}(s;v)+\int_{\mathcal{Z}}\Delta c(s,z;v)\Delta c^*(s,z;v)\,\pi(dz)\right]$$
$$+\frac{1}{2}\text{trace}\int_{\mathcal{Z}}R(s,z)[\Delta c(s,z;v)\Delta c^*(s,z;v)]\,\pi(dz)\geq 0,\qquad \forall\, v\in U,\ a.e.\ a.s..$$

4) *The following transversality condition holds:*

$$\langle\mu,z-Ef(y_0,y(1))\rangle\leq 0,\qquad \forall\, z\in Q. \tag{2.38}$$

Remark 2.6. When $c(x,v,z)\equiv c(x,z), f(y,x)\equiv y, Q\equiv\{x_0\}$, Theorem 2.1 can be stated as follows. Let $(y(\cdot),u(\cdot))$ be an optimal pair; then there exist $p(\cdot)\in L^2_{\mathcal{F}}[[0,1];\mathcal{R}^n]$, $\{q^i(\cdot)\}_1^d\in L^2_{\mathcal{F},p}[[0,1];\mathcal{R}^{n\times d}]$, $r(\cdot,\cdot)\in F_p^2[[0,1];\mathcal{R}^n]$, and $P(\cdot)\in L^2_{\mathcal{F}}[[0,1];\mathcal{R}^{n\times n}]$, $\{Q^i(\cdot)\}_1^d\in L^2_{\mathcal{F},p}[[0,1];L(\mathcal{R}^d,\mathcal{R}^{n\times n})]$, $R(\cdot,\cdot)\in F_p^2[[0,1];\mathcal{R}^{n\times n}]$ such that

1) the Itô-type adjoint equations (2.22), (2.23) with

$$p(1)=h_x(y_0,y(1)),\qquad P(1)=h_{xx}(y_0,y(1))$$

are satisfied, with $p(\cdot)$ and $P(\cdot)$ being cadlag processes;

2) the following maximum condition holds:

$$H(y(s-),v,1,p(s-),\{q^i(s)\}_1^d,0)-H(y(s-),u(s),1,p(s-),\{q^i(s)\}_1^d,0)$$
$$+\frac{1}{2}\text{trace } P(s-)[\Delta b^i(s;v)\Delta b^{i*}(s;v)]\geq 0,\qquad \forall\, v\in U,\text{a.e. a.s..}$$

This is the result of Situ [23].

Remark 2.7. In [7], Davis and Elliott considered an optimal control problem for an elementary jump process $k(\cdot)$. Let the random jump time and the random jump position of $k(\cdot)$ be denoted by $T(\omega)$ and $Z(\omega)$, respectively. The problem of the case $T(\omega)\leq 1$ can be stated, under our framework, as follows. Minimize the following cost functional:

$$J(v)=E[x^1(1)x^2(1)] \tag{2.39}$$

subject to the system

$$\begin{cases} x^1(t)=1+\displaystyle\int_{\mathcal{Z}\times(0,t]}x^1(s-)[\alpha(s,v(s))\beta(s,v(s),z)-1]\,\widetilde{N}_k(dzds),\\ x^2(t)=\displaystyle\int_{\mathcal{Z}\times(0,t]}f(s,z,v(s))\alpha(s,v(s))\beta(s,v(s)z)\,\widehat{N}_k(dzds). \end{cases} \tag{2.40}$$

Here $\widehat{N}_k(dzds)$ denotes the predictable dual projection of $N_k(dzds)$. On one hand, (2.40) does not contain the diffusion term $b(x,v)$ and is linear in the state $x=(x^1,x^2)^*$, and also the cost is of a special quadratic form. On the other hand, however, the measure $\widehat{N}(dzds)=\lambda(dz,s)d\Lambda(s)$ in [7] is allowed to have atoms with respect to $s\in[0,1]$. In this paper, we have assumed $\widehat{N}_k(dzds)=\pi(dz)ds$; actually, however, we only need to assume that $\widehat{N}(dzds)$ is nonatomic with respect to $s\in[0,1]$. In the case of nonatomic $\widehat{N}(dzds)$, Theorem 2.1 holds for the above stated optimal control

problem. Moreover, we can check that

$$p(t) = (p^1(t), y^1(t))^*, \quad y^1(t) > 0, \quad q(t) = 0,$$
$$r(t,z) = (r^1(t,z), y^1(t-)[\alpha(t,u(t))\beta(t,u(t),z) - 1])^*,$$
$$P(t) = \begin{pmatrix} 0 & P^{12}(t) \\ P^{21}(t) & 0 \end{pmatrix}, \quad Q(t) = 0, \quad R(t,z) = \begin{pmatrix} 0 & R^{12}(t,z) \\ R^{21}(t,z) & 0 \end{pmatrix}.$$

Then the following minimum principle is true. Let $(y^1(\cdot), y^2(\cdot), u(\cdot))$ be optimal; then there exist $p^1(\cdot), r^1(\cdot,\cdot)$ satisfying the equation

$$\begin{aligned} p^1(t) = y^2(1) &+ \iint_{\mathcal{Z}\times(t,1]} r^1(s,z)[\alpha(s,u(s))\beta(s,u(s),z) - 1]\widehat{N}_k(dzds) \\ &- \iint_{\mathcal{Z}\times(t,1]} r^1(s,z)\widetilde{N}_k(dzds), \end{aligned} \tag{2.41}$$

such that $u(t)$ almost everywhere, almost surely, minimizes

$$\alpha(t,v)\int_{\mathcal{Z}} [f(t,z,v) + r^1(t,z)]\beta(t,v)\lambda(t,dz). \tag{2.42}$$

Next, we explain how to derive the related result of Davis and Elliott [7]. Equation (2.41) can be rewritten as

$$p_1(t) = y^2(1) - \iint_{\mathcal{Z}\times(t,1]} r^1(s,z)\widetilde{N}_k^{u(\cdot)}(dzds), \tag{2.43}$$

where $\widetilde{N}_k^{u(\cdot)}(dzdt)$ is the Martingale part of $N_k(dzdt)$ under the measure $\nu^{u(\cdot)}$ determined by the Lévy system $(\beta(t,u(t),z)\lambda(t,dz), \alpha(t,u(t))\Lambda(t))$. We have

$$J(u(\cdot)) = E^{u(\cdot)}[f(T,Z)], \tag{2.44}$$

where $E^{u(\cdot)}$ denotes the expectation relative to the measure $\nu^{u(\cdot)}$. According to the Martingale representation theorem, $E^{u(\cdot)}[f(T,Z)|\,\mathcal{F}_t]$ can be written as the following integral:

$$E^{u(\cdot)}[f(T,Z)|\,\mathcal{F}_t] = J(u(\cdot)) + \iint_{\mathcal{Z}\times(0,t]} g(s,z)\widetilde{N}_k^{u(\cdot)}(dzds) \tag{2.45}$$

for some g. Then from (2.43)–(2.45), we can show that

$$r^1(t,z) = g(t,z) - f(t,z,u(t)). \tag{2.46}$$

Putting (2.46) into (2.42), we obtain the necessary part of Theorem 5.1 of Davis and Elliott [7].

3. The proof of Theorem 2.1.

Step 1. *Applying Ekeland's variational principle.* We first consider the case that the set U_{ad} is bounded in $L^{\infty,8}_{\mathcal{F},p}[[0,1];\mathcal{R}^m]$; the unbounded case can be reduced to the bounded case (see step 5 below). Assume that

$$U_{\text{ad}} \text{ is bounded in } L^{\infty,8}_{\mathcal{F},p}[[0,1];\mathcal{R}^m].$$

An application of Ekeland's variational principle will lead to the reduction of a general end-constraint problem to a family of free end-constraint problems.

Define the following auxiliary function:

$$J(v(\cdot),x_0;\varepsilon,\delta)=\Psi(J(v(\cdot),x_0),Ef(x_0,x(1));\varepsilon,\delta), \tag{3.1}$$

with $\Psi(\cdot,\cdot;\varepsilon,\delta)$ being defined as in §2. Then consider the metric space $(\mathcal{R}^n\times U_{\rm ad},d)$, with the distance d defined by

$$d((x_1,v_1(\cdot)),(x_2,v_2(\cdot)))=\sqrt{|x_1-x_2|^2+\hat{d}^2(v_1(\cdot),v_2(\cdot))}. \tag{3.2}$$

We can verify that $(\mathcal{R}^n\times U_{\rm ad},d)$ is complete and $J(v(\cdot),x_0;\varepsilon,\delta)$ is continuous and bounded. In fact, since $U_{\rm ad}$ is bounded under the norm $||\cdot||$, the system state of (1.1), as a functional defined on $(\mathcal{R}^n\times U_{\rm ad},d)$, is bounded and continuous (note Remark 2.4); from this, we can check that $J(v(\cdot),x_0;\varepsilon,\delta)$ is bounded and continuous. To prove the completeness of $(\mathcal{R}^n\times U_{\rm ad},d)$, we only need to show the completeness of $(U_{\rm ad},\hat{d})$. For this purpose, we suppose that $\{v_i(\cdot)\}_{i=1}^{\infty}$ is a Cauchy sequence in $(U_{\rm ad},\hat{d})$. Then we can find a subsequence $\{v_{i_K}(\cdot)\}_{K=1}^{\infty}$ such that

$$\hat{d}(v_{i_{K+1}}(\cdot),v_{i_K}(\cdot))<\frac{1}{2^K}.$$

Set

$$I^j=:\bigcup_{K=j}^{\infty}\{t\in[0,1]|\,\hat{d}(v_{i_{K+1}}(\cdot),v_{i_K}(\cdot))>0\},\qquad j=1,2,\ldots.$$

Choose

$$v(t)=:\begin{cases}v_1(t), & \text{as } t\in[0,1]\setminus I^1,\\ v_j(t), & \text{as } t\in I^{j-1}\setminus I^j,\qquad j=2,3,\ldots.\end{cases}$$

Then we have

$$v(\cdot)\in U_{\rm ad},\quad \hat{d}(v_{i_K}(\cdot),v(\cdot))\to 0\quad\text{as } K\to\infty.$$

Since $\{v_i(\cdot)\}_{i=1}^{\infty}$ is a Cauchy sequence in $(U_{\rm ad},\hat{d})$, we also have

$$v_i(\cdot)\to v(\cdot)\,(\in U_{\rm ad})\quad\text{as } i\to\infty.$$

The completeness of $(U_{\rm ad},\hat{d})$ then follows.

Also, we have for any given $\varepsilon>0$,

$$\begin{aligned}&\Phi(J(v(\cdot),x_0),Ef(x_0,x(1));\varepsilon)>0,\qquad \forall\,(x_0,v(\cdot))\in\mathcal{R}^n\times U_{\rm ad};\\ &\Phi(J(u(\cdot),y_0),Ef(y_0,y(1));\varepsilon)=\varepsilon;\\ &J(v(\cdot),x_0;\varepsilon,\delta)>0,\qquad \forall\,(x_0,v(\cdot))\in\mathcal{R}^n\times U_{\rm ad}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\text{for sufficiently small }\delta>0;\\ &J(u(\cdot),y_0;\varepsilon,\delta)\le\varepsilon+2\delta+\inf_{(x_0,v(\cdot))\in\mathcal{R}^n\times U_{\rm ad}}J(v(\cdot),x_0;\varepsilon,\delta).\end{aligned} \tag{3.3}$$

Therefore we can apply Ekeland's variational principle (cf. [8], [9]) and conclude that there exist $u^{\varepsilon\delta}(\cdot)\in U_{\rm ad}$ and $y_0^{\varepsilon\delta}\in\mathcal{R}^n$ such that

1) $J(u^{\varepsilon\delta}(\cdot),y_0^{\varepsilon\delta};\varepsilon,\delta)\le\varepsilon+2\delta$;

2) $d((y_0^{\varepsilon\delta},u^{\varepsilon\delta}(\cdot)),(y_0,u(\cdot)))\le\sqrt{\varepsilon+2\delta}$;

3) $\bar{J}(v(\cdot),x_0;\varepsilon,\delta)=:J(v(\cdot),x_0;\varepsilon,\delta)+\sqrt{\varepsilon+2\delta}d((x_0,v(\cdot)),(y_0^{\varepsilon\delta},u^{\varepsilon\delta}(\cdot)))$
$\ge J(u^{\varepsilon\delta}(\cdot),y_0^{\varepsilon\delta}),\ \ \forall\,(x_0,v(\cdot))\in\mathcal{R}^n\times U_{\rm ad}$.

Set

$$
\begin{aligned}
\lambda^{\varepsilon\delta} &=: \Psi_s(J(u^{\varepsilon\delta}(\cdot), y_0^{\varepsilon\delta}), Ef(y_0^{\varepsilon\delta}, y^{\varepsilon\delta}(1)); \varepsilon, \delta),\\
\mu^{\varepsilon\delta} &=: \Psi_z(J(u^{\varepsilon\delta}(\cdot), y_0^{\varepsilon\delta}), Ef(y_0^{\varepsilon\delta}, y^{\varepsilon\delta}(1)); \varepsilon, \delta),\\
\hat{\lambda}^{\varepsilon\delta} &=: \Phi_s(J(u^{\varepsilon\delta}(\cdot), y_0^{\varepsilon\delta}), Ef(y_0^{\varepsilon\delta}, y^{\varepsilon\delta}(1)); \varepsilon),\\
\hat{\mu}^{\varepsilon\delta} &=: \Phi_z(J(u^{\varepsilon\delta}(\cdot), y_0^{\varepsilon\delta}), Ef(y_0^{\varepsilon\delta}, y^{\varepsilon\delta}(1)); \varepsilon).
\end{aligned} \tag{3.4}
$$

From the first relation (3.3) and Lemmas (2.5) and (2.6), we have for each sufficiently small $\varepsilon > 0$,

$$
\lim_{\delta\to 0+}(\lambda^{\varepsilon\delta} - \hat{\lambda}^{\varepsilon\delta}) = 0,
$$

$$
\lim_{\delta\to 0+}(\mu^{\varepsilon\delta} - \hat{\mu}^{\varepsilon\delta}) = 0,
$$

$$
|\hat{\lambda}^{\varepsilon\delta}|^2 + |\hat{\mu}^{\varepsilon\delta}|^2 = 1.
$$

Therefore, for each sufficiently small $\varepsilon > 0$, we can choose $\delta(\varepsilon) > 0$ such that

$$
\begin{aligned}
&J(u^{\varepsilon\delta(\varepsilon)}(\cdot), y_0^{\varepsilon\delta(\varepsilon)}; \varepsilon, \delta(\varepsilon)) > 0,\\
&\delta(\varepsilon) \le \varepsilon,\\
&|\mu^{\varepsilon\delta(\varepsilon)}|^2 + |\mu^{\varepsilon\delta(\varepsilon)}|^2 = 1 + o(1), \quad \text{as } \varepsilon \to 0.
\end{aligned} \tag{3.5}
$$

Set

$$
\begin{aligned}
\lambda^{\varepsilon} &=: \lambda^{\varepsilon\delta(\varepsilon)}, \qquad \mu^{\varepsilon} =: \mu^{\varepsilon\delta(\varepsilon)},\\
y_0^{\varepsilon} &=: y_0^{\varepsilon\delta(\varepsilon)}, \qquad u^{\varepsilon}(\cdot) =: u^{\varepsilon\delta(\varepsilon)}(\cdot).
\end{aligned}
$$

Then we see (noting Lemma 2.5) that $\lambda^{\varepsilon} \ge 0$ and $\mu^{\varepsilon} \in \mathcal{R}^k$ satisfy the following:

$$
\begin{aligned}
&\lim_{\varepsilon\to 0+}(|\lambda^{\varepsilon}|^2 + |\mu^{\varepsilon}|^2) = 1,\\
&\langle \mu^{\varepsilon}, z - Ef(y_0^{\varepsilon}, y^{\varepsilon}(1))\rangle \le \delta(\varepsilon) \le \varepsilon.
\end{aligned} \tag{3.6}
$$

Step 2. *Computing the first-order component of the cost variation.* In this and the next steps,we look for the necessary conditions for the minimization of $\bar{J}(v(\cdot), x_0; \varepsilon, \delta)$ at $(y_0^{\varepsilon}, u^{\varepsilon}(\cdot))$.

For given $(\eta, v(\cdot)) \in \mathcal{R}^n \times U_{\text{ad}}$, set

$$
\begin{aligned}
u^{\varepsilon\rho}(t) &= u^{\varepsilon}(t)\chi_{[0,1]\setminus I_\rho}(t) + v(t)\chi_{I\rho}(t),\\
y_0^{\varepsilon\rho} &= y_0^{\varepsilon} + |I_\rho|\eta,\\
y^{\varepsilon\rho}(\cdot) &= y(\cdot; u^{\varepsilon\rho}(\cdot), y_0^{\varepsilon\rho}).
\end{aligned} \tag{3.7}
$$

We introduce, as in §2, the following simplified notations:

$$
\begin{aligned}
\Delta m^{\varepsilon}(s; v) &=: m(y^{\varepsilon}(s), v) - m(y^{\varepsilon}(s), u^{\varepsilon}(s)),\\
m^{\varepsilon}(s) &=: m(y^{\varepsilon}(s), u^{\varepsilon}(s)),\\
\Delta n^{\varepsilon}(s, z; v) &=: n(y^{\varepsilon}(s-), v, z) - n(y^{\varepsilon}(s-), u^{\varepsilon}(s), z),\\
n^{\varepsilon}(s, z) &=: n(y^{\varepsilon}(s-), u^{\varepsilon}(s), z),
\end{aligned} \tag{3.8}
$$

with m standing for a, b, g and all their (up to second-) derivatives in x, and n for c and its (up to second-) derivatives in x.

Let $y^{\varepsilon\rho}(\cdot)$ be the solution of (1.1) corresponding to $(y_0^{\varepsilon\rho}, u^{\varepsilon\rho}(\cdot))$. We define, as in §2, the half- and first-order processes $y_1^\varepsilon(\cdot), y_2^\varepsilon(\cdot)$, respectively, by

$$
\begin{aligned}
y_1^\varepsilon(t) = &\int_{(0,t]} a_x^\varepsilon(s) y_1^\varepsilon(s) ds + \int_{(0,t]} [b_x^{\varepsilon k}(s) y_1^\varepsilon(s) + \Delta b^{\varepsilon k}(s; u^{\varepsilon\rho}(s))] dw^k(s) \\
&+ \int\int_{\mathcal{Z}\times(0,t]} [c_x^\varepsilon(s,z) y_1^\varepsilon(s-) + \Delta c^\varepsilon(s,z; u^{\varepsilon\rho}(s))] \widetilde{N}_k(dzds)
\end{aligned} \tag{3.9}
$$

and

(3.10)

$$
\begin{aligned}
y_2^\varepsilon(t) = &|I_\rho|\eta + \int_{(0,t]} \left[a_x^\varepsilon(s) y_2^\varepsilon(s) + \Delta a^\varepsilon(s; u^{\varepsilon\rho}(s)) + \frac{1}{2} a_{x^i x^j}^\varepsilon(s) y_1^{\varepsilon i}(s) y_1^{\varepsilon j}(s) \right] ds \\
&+ \int_{(0,t]} \left[b_x^{\varepsilon k}(s) y_2^\varepsilon(s) + \Delta b_x^{\varepsilon k}(s; u^{\varepsilon\rho}(s)) y_1^\varepsilon(s) + \frac{1}{2} b_{x^i x^j}^{\varepsilon k}(s) y_1^{\varepsilon i}(s) y_1^{\varepsilon j}(s) \right] dw^k(s) \\
&+ \int\int_{\mathcal{Z}\times(0,t]} \Big[c_x^\varepsilon(s,z) y_2^\varepsilon(s-) + \Delta c_x^\varepsilon(s,z; u^{\varepsilon\rho}(s)) y_1^\varepsilon(s-) \\
&\qquad + \frac{1}{2} c_{x^i x^j}^\varepsilon(s,z) y_1^{\varepsilon i}(s-) y_1^{\varepsilon j}(s-) \Big] \widetilde{N}_k(dzds).
\end{aligned}
$$

From Remark 2.2, we see

$$
\begin{aligned}
&\sup_{0\le t\le 1} E|y_1^\varepsilon(t)|^8 = O(|I_\rho|^4), \\
&\sup_{0\le t\le 1} E|y_2^\varepsilon(t)|^4 = O(|I_\rho|^4), \\
&\sup_{0\le t\le 1} E|y^{\varepsilon\rho}(t) - y^\varepsilon(t) - y_1^\varepsilon(t) - y_2^\varepsilon(t)|^2 = o(|I_\rho|^2), \text{ as } |I_\rho| \to 0.
\end{aligned} \tag{3.11}
$$

In this step, we are to calculate the first-order component of the cost variation. From 3) of step 1,we have

$$
\begin{aligned}
-|I_\rho|\sqrt{\varepsilon + 2\delta}\sqrt{1+|\eta|^2} &\le J(u^{\varepsilon\rho}(\cdot), y^{\varepsilon\rho}(0); \varepsilon) - J(u^\varepsilon(\cdot), y_0^\varepsilon; \varepsilon) \\
&\le \lambda^\varepsilon [J(u^{\varepsilon\rho}(\cdot), y_0^\varepsilon + |I_\rho|\eta) - J(u^\varepsilon(\cdot), y_0^\varepsilon)] \\
&\quad + \mu^{\varepsilon j} [Ef^j(y_0^\varepsilon + |I_\rho|\eta, y^{\varepsilon\rho}(1)) - Ef^j(y_0^\varepsilon, y^\varepsilon(1))] \\
&\quad + O(|J(u^{\varepsilon\rho}(\cdot), y_0^\varepsilon + |I_\rho|\eta) - J(u^\varepsilon(\cdot), y_0^\varepsilon)|^2) \\
&\quad + O(|Ef(y_0^\varepsilon + |I_\rho|\eta, y^{\varepsilon\rho}(1)) - Ef(y_0^\varepsilon, y^\varepsilon(1))|^2).
\end{aligned} \tag{3.12}
$$

Using (3.11), we have

(3.13)

$$
\begin{aligned}
&J(u^{\varepsilon\rho}(\cdot), y_0^\varepsilon + |I_\rho|\eta) - J(u^\varepsilon(\cdot), y_0^\varepsilon) \\
&= E\int_0^1 [g(y^{\varepsilon\rho}(t), u^{\varepsilon\rho}(t)) - g(y^\varepsilon(t), u^\varepsilon(t))] dt + E[h(y_0^{\varepsilon\rho}, y^{\varepsilon\rho}(1)) - h(y_0^\varepsilon, y^\varepsilon(1))] \\
&= E\int_0^1 [g(y^\varepsilon(t) + y_1^\varepsilon(t) + y_2^\varepsilon(t), u^{\varepsilon\rho}(t)) - g(y^\varepsilon(t), u^\varepsilon(t))] dt \\
&\quad + E[h(y_0^{\varepsilon\rho}, y^\varepsilon(1) + y_1^\varepsilon(1) + y_2^\varepsilon(1)) - h(y_0^\varepsilon, y^\varepsilon(1))] + o(|I_\rho|) \\
&= E\int_0^1 [g(y^\varepsilon(t) + y_1^\varepsilon(t) + y_2^\varepsilon(t), u^\varepsilon(t)) - g(y^\varepsilon(t), u^\varepsilon(t))] dt
\end{aligned}
$$

$$
\begin{aligned}
&+ E[h(y_0^{\varepsilon\rho}, y^\varepsilon(1) + y_1^\varepsilon(1) + y_2^\varepsilon(1)) - h(y_0^\varepsilon, y^\varepsilon(1))] \\
&+ E\int_0^1 [g(y^\varepsilon(t) + y_1^\varepsilon(t) + y_2^\varepsilon(t), u^{\varepsilon\rho}(t)) - g(y^\varepsilon(t) + y_1^\varepsilon(t) + y_2^\varepsilon(t), u^\varepsilon(t))] + o(|I_\rho|) \\
&= |I_\rho|\langle Eh_y(y_0^\varepsilon, y^\varepsilon(1)), \eta\rangle + E\langle h_x(y_0^\varepsilon, y^\varepsilon(1)), y_1^\varepsilon(1) + y_2^\varepsilon(1)\rangle \\
&+ \frac{1}{2}Ey_1^{\varepsilon *}(1)h_{xx}(y_0^\varepsilon, y^\varepsilon(1))y_1^\varepsilon(1) + \\
&+ E\int_0^1 g_x^\varepsilon(s)[y_1^\varepsilon(s) + y_2^\varepsilon(s)]ds + \frac{1}{2}E\int_0^1 y_1^{\varepsilon *}(s)g_{xx}^\varepsilon(s)y_1^\varepsilon(s)ds \\
&+ E\int_0^1 \Delta g^\varepsilon(s; u^{\varepsilon\rho}(s))ds + o(|I_\rho|)
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
Ef^j(y_0^\varepsilon + |I_\rho|\eta, y^{\varepsilon\rho}(1)) - Ef^j(y_0^\varepsilon, y^\varepsilon(1)) &= |I_\rho|\langle Ef_y^j(y_0^\varepsilon, y^\varepsilon(1)), \eta\rangle_n \\
&+ E\langle f_x^j(y_0^\varepsilon, y^\varepsilon(1)), y_1^\varepsilon(1) + y_2^\varepsilon(1)\rangle \\
&+ \frac{1}{2}Ey_1^{\varepsilon *}(1)f_{xx}^j(y_0^\varepsilon, y^\varepsilon(1))y_1^\varepsilon(1) + o(|I_\rho|).
\end{aligned} \tag{3.14}
$$

From Lemma 2.4, we see that

$$
\begin{cases}
p^\varepsilon(t) = p^\varepsilon(1) + \displaystyle\int_{(t,1]} H_x(y^\varepsilon(s), u^\varepsilon(s), \lambda^\varepsilon, p^\varepsilon(s), \{q^{\varepsilon i}(s)\}_1^d, r^\varepsilon(s,\cdot))ds \\
\qquad\qquad - \displaystyle\int_{(t,1]} q^{\varepsilon i}(s)dw^i(s) - \int\int_{\mathcal{Z}\times(t,1]} r^\varepsilon(s,z)\widetilde{N}_k(dzds), \\
p^\varepsilon(1) = \lambda^\varepsilon h_x(y_0^\varepsilon, y^\varepsilon(1)) + \mu^{\varepsilon j} f_x^j(y_0^\varepsilon, y^\varepsilon(1))
\end{cases} \tag{3.15}
$$

and

$$
\begin{cases}
P^\varepsilon(t) = P^\varepsilon(1) \\
\qquad + \displaystyle\int_{(t,1]} \Big\{ a_x^{\varepsilon *}(s)P^\varepsilon(s) + P^\varepsilon(s)a_x^\varepsilon(s) \\
\qquad\qquad + b_x^{\varepsilon i *}(s)P^\varepsilon(s)b_x^{\varepsilon i}(s) + (b_x^{\varepsilon i *}(s)Q^{\varepsilon i}(s) + Q^{\varepsilon i}(s)b_x^{\varepsilon i}(s) \\
\qquad\qquad + \displaystyle\int_{\mathcal{Z}} c_x^{\varepsilon *}(s,z)P(s)c_x^\varepsilon(s,z)\pi(dz) \\
\qquad\qquad + \displaystyle\int_{\mathcal{Z}} [c_x^{\varepsilon *}(s,z)R^\varepsilon(s,z)c_x^\varepsilon(s,z) + c_x^{\varepsilon *}(s,z)R^\varepsilon(s,z) \\
\qquad\qquad\qquad\qquad + R^\varepsilon(s,z)c_x^\varepsilon(s,z)]\pi(dz) \\
\qquad\qquad + H_{xx}(y^\varepsilon(s), u^\varepsilon(s), \lambda^\varepsilon, p^\varepsilon(s), \{q^{\varepsilon i}(s)\}_1^d, r^\varepsilon(s,\cdot)) \Big\} ds \\
\qquad - \displaystyle\int_{(t,1]} Q^{\varepsilon i}(s)dw^i(s) - \int\int_{\mathcal{Z}\times(t,1]} R^\varepsilon(s,z)\widetilde{N}_k(dzds), \\
P^\varepsilon(1) = \lambda^\varepsilon h_{xx}(y_0^\varepsilon, y^\varepsilon(1)) + \mu^{\varepsilon j} f_{xx}^j(y_0^\varepsilon, y^\varepsilon(1))
\end{cases} \tag{3.16}
$$

have unique solutions $(p^\varepsilon(\cdot), \{q^{\varepsilon i}(\cdot)\}_1^d, r^\varepsilon(\cdot,\cdot))$ and $(P^\varepsilon(\cdot), \{Q^{\varepsilon i}(\cdot)\}_1^d, R^\varepsilon(\cdot,\cdot))$, respectively, with $p^\varepsilon(\cdot)$ and $P^\varepsilon(\cdot)$ being cadlag processes.

Using Itô's formula, we have from (3.9), (3.15) and (3.16), that

$$
E\langle \lambda^\varepsilon h_x(y_0^\varepsilon, y^\varepsilon(1)) + \mu^{\varepsilon j} f_x^j(y_0^\varepsilon, y^\varepsilon(1)), y_1^\varepsilon(1) + y_2^\varepsilon(1)\rangle = E\langle p^\varepsilon(1), y_1^\varepsilon(1) + y_2^\varepsilon(1)\rangle
$$

$$
\begin{aligned}
&= \langle p^\varepsilon(0), \eta\rangle |I_\rho| + E\int_0^1 [H(y^\varepsilon(s), u^{\varepsilon\rho}(s), 0, p^\varepsilon(s), \{q^{\varepsilon i}(s)\}_1^d, r_s^\varepsilon(\cdot)) \\
&\qquad - H(y^\varepsilon(s), u^\varepsilon(s), 0, p^\varepsilon(s), \{q^{\varepsilon i}(s)\}_1^d, r_s^\varepsilon(\cdot))]ds \\
&\quad + \frac{1}{2}E\int_0^1 y_1^{\varepsilon *}(s) H_{xx}(y^\varepsilon(s), u^\varepsilon(s), 0, p^\varepsilon(s), \{q^{\varepsilon i}(s)\}_1^d, r_s^\varepsilon(\cdot)) y_1^\varepsilon(s) ds \qquad (3.17)\\
&\quad + E\int_0^1 \langle q^{\varepsilon i}, \Delta b_x^{\varepsilon i}(s; u^{\varepsilon\rho}(s)) y_1^\varepsilon(s)\rangle ds \\
&\quad + E\int\int_{\mathcal{Z}\times[0,1]} \langle r_s^\varepsilon(z), \Delta c_x^\varepsilon(s; u^{\varepsilon\rho}(s)) y_1^\varepsilon(s)\rangle \pi(dz)ds,
\end{aligned}
$$

$$
\begin{aligned}
y_1^\varepsilon(t) y_1^{\varepsilon *}(t) = &\int_{(0,t]} [a_x^\varepsilon(s) y_1^\varepsilon(s) y_1^{\varepsilon *}(s) + y_1^\varepsilon(s) y_1^{\varepsilon *}(s) a_x^{\varepsilon *}(s) \\
&+ b_x^{\varepsilon k}(s) y_1^\varepsilon(s) y_1^{\varepsilon *}(s) b_x^{\varepsilon k *}(s) + \int_{\mathcal{Z}} c_x^\varepsilon(s,z) y_1^\varepsilon(s) y_1^{\varepsilon *}(s) c_x^{\varepsilon *}(s,z)\pi(dz) \\
&+ b_x^{\varepsilon k}(s) y_1^\varepsilon(s) \Delta b^{\varepsilon k *}(s; u^{\varepsilon\rho}(s)) + \Delta b^{\varepsilon k}(s; u^{\varepsilon\rho}(s)) y_1^{\varepsilon *}(s) b_x^{\varepsilon k *}(s) \\
&+ \Delta b^{\varepsilon k}(s; u^{\varepsilon\rho}(s)) \Delta b^{\varepsilon k *}(s; u^{\varepsilon\rho}(s)) \\
&+ \int_{\mathcal{Z}} [c_x^\varepsilon(s,z) y_1^\varepsilon(s) \Delta c^{\varepsilon *}(s,z; u^{\varepsilon\rho}(s)) \\
&\qquad + \Delta c^\varepsilon(s,z; u^{\varepsilon\rho}(s)) y_1^{\varepsilon *}(s) c_x^{\varepsilon *}(s,z) \\
&\qquad + \Delta c^\varepsilon(s,z; u^{\varepsilon\rho}(s)) \Delta c^{\varepsilon *}(s,z; u^{\varepsilon\rho}(s))]\pi(dz)]ds \qquad (3.18)\\
&+ \int_{(0,t]} [b_x^{\varepsilon k}(s) y_1^\varepsilon(s) y_1^{\varepsilon *}(s) + y_1^\varepsilon(s) y_1^{\varepsilon *}(s) b_x^{\varepsilon k *}(s) \\
&\qquad + y_1^\varepsilon(s) \Delta b^{\varepsilon k *}(s; u^{\varepsilon\rho}(s)) + \Delta b^{\varepsilon k}(s; u^{\varepsilon\rho}(s)) y_1^{\varepsilon *}(s)] dw^k(s) \\
&+ \int\int_{\mathcal{Z}\times(0,t]} [c_x^\varepsilon(s,z) y_1^\varepsilon(s-) y_1^{\varepsilon *}(s-) + y_1^\varepsilon(s-) y_1^{\varepsilon *}(s-) c_x^{\varepsilon *}(s,z) \\
&\qquad + c_x^\varepsilon(s,z) y_1^\varepsilon(s-) y_1^{\varepsilon *}(s-) c_x^{\varepsilon *}(s,z) \\
&\qquad + c_x^\varepsilon(s,z) y_1^\varepsilon(s-) \Delta c^{\varepsilon *}(s,z; u^{\varepsilon\rho}(s)) \\
&\qquad + \Delta c^\varepsilon(s,z; u^{\varepsilon\rho}(s)) y_1^{\varepsilon *}(s-) c_x^{\varepsilon *}(s,z) \\
&\qquad + \Delta c^\varepsilon(s,z; u^{\varepsilon\rho}(s)) \Delta c^{\varepsilon *}(s,z; u^{\varepsilon\rho}(s)) \\
&\qquad + y_1^\varepsilon(s-) \Delta c^{\varepsilon *}(s,z; u^{\varepsilon\rho}(s)) \\
&\qquad + \Delta c^\varepsilon(s,z; u^{\varepsilon\rho}(s)) y_1^{\varepsilon *}(s-)] \widetilde{N}_k(dzds),
\end{aligned}
$$

and

$$
\begin{aligned}
&\lambda^\varepsilon E y_1^{\varepsilon *}(1) h_{xx}(y_0^\varepsilon, y^\varepsilon(1)) y_1^\varepsilon(1) + \mu^{\varepsilon j} E y_1^{\varepsilon *}(1) f_{xx}^j(y_0^\varepsilon, y^\varepsilon(1)) y_1^\varepsilon(1) \\
&\quad = \text{trace } E[P^\varepsilon(1) y_1^\varepsilon(1) y_1^{\varepsilon *}(1)] \\
&\quad = E\int_0^1 \Big\{ y_1^{\varepsilon *}(s) H_{xx}(y^\varepsilon(s), u^\varepsilon(s), \lambda^\varepsilon, p^\varepsilon(s), \{q^{\varepsilon i}(s)\}_1^d, r_s^\varepsilon(\cdot)) y_1^\varepsilon(s) ds \\
&\qquad + \text{trace } P^\varepsilon(s) \Big[\Delta b^{\varepsilon i}(s; u^{\varepsilon\rho}(s)) \Delta b^{\varepsilon i *}(s; u^{\varepsilon\rho}(s)) \\
&\qquad\qquad + \int_{\mathcal{Z}} \Delta c^\varepsilon(s,z; u^{\varepsilon\rho}(s)) \Delta c^{\varepsilon *}(s,z; u^{\varepsilon\rho}(s)) \pi(dz) \Big]
\end{aligned}
$$

(3.19)
$$
\begin{aligned}
&+\int_{\mathcal{Z}} \operatorname{trace}\,[R_s^\varepsilon(z)\Delta c^\varepsilon(s,z;u^{\varepsilon\rho}(s))\Delta c^{\varepsilon *}(s,z;u^{\varepsilon\rho}(s))]\pi(dz)\Big\}ds\\
&+2E\int_0^1 \Big\{\operatorname{trace} P^\varepsilon(s)\Big[b_x^{\varepsilon k}(s)y_1^\varepsilon(s)\Delta b^{\varepsilon k *}(s;u^{\varepsilon\rho}(s))\\
&\qquad\qquad +\int_{\mathcal{Z}} c_x^\varepsilon(s,z)y_1^\varepsilon(s)\Delta c^{\varepsilon *}(s,z;u^{\varepsilon\rho}(s))\pi(dz)\Big]\\
&\qquad +\operatorname{trace}[Q^\varepsilon(s)y_1^\varepsilon(s)\Delta b^{\varepsilon k *}(s;u^{\varepsilon\rho}(s))]\\
&\qquad +\int_{\mathcal{Z}} \operatorname{trace} R_s^\varepsilon(z)[c_x^\varepsilon(s,z)y_1^\varepsilon(s)\Delta c^{\varepsilon *}(s,z;u^{\varepsilon\rho}(s))\\
&\qquad\qquad +y_1^\varepsilon(s)\Delta c^{\varepsilon *}(s,z;u^{\varepsilon\rho}(s))]\pi(dz)\Big\}ds.
\end{aligned}
$$

Noting (3.11), we conclude from (3.12)–(3.14) and (3.17)–(3.19) that

$$
\begin{aligned}
&\langle \lambda^\varepsilon Eh_y(y_0^\varepsilon, y^\varepsilon(1)) + \mu^{\varepsilon j} Ef_y^j(y_0^\varepsilon, y^\varepsilon(1)) + p^\varepsilon(0), \eta\rangle |I_\rho|\\
&\quad +\int_0^1 l^\varepsilon(s;u^{\varepsilon\rho}(s))ds + o(|I_\rho|) \geq -|I_\rho|\sqrt{\varepsilon+2\delta(\varepsilon)}\sqrt{1+|\eta|^2},
\end{aligned}
\tag{3.20}
$$

where $l^\varepsilon(\cdot;v)$ is defined by

$$
\begin{aligned}
l^\varepsilon(s;v) =:\ & E\{H(y^\varepsilon(s),v,\lambda^\varepsilon,p^\varepsilon(s),\{q^{\varepsilon i}(s)\}_1^d,r_s^\varepsilon(\cdot))\\
&-H(y^\varepsilon(s),u^\varepsilon(s),\lambda^\varepsilon,p^\varepsilon(s),\{q^{\varepsilon i}(s)\}_1^d,r_s^\varepsilon(\cdot))\\
&+\frac{1}{2}\operatorname{trace} P^\varepsilon(s)\Big[\Delta b^{\varepsilon i}(s;v)\Delta b^{\varepsilon i *}(s;v)\\
&+\int_{\mathcal{Z}} \Delta c^\varepsilon(s,z;v)\Delta c^{\varepsilon *}(s,z;v)\pi(dz)\Big]\\
&+\frac{1}{2}\operatorname{trace}\int_{\mathcal{Z}} R_s(z)\Delta c^\varepsilon(s,z;v)\Delta c^{\varepsilon *}(s,z;v)\pi(dz)\}.
\end{aligned}
\tag{3.21}
$$

Step 3. *Differentiability.* For given $v(\cdot)\in U_{\text{ad}}$, applying Lemma 2.2 to the real valued Lebesgue integrable function, we know that there exists $I_\rho\subset[0,1]$ such that

$$
\begin{aligned}
&|I_\rho| = \rho,\\
&\int_{I_\rho} l^\varepsilon(s;v(s))ds = \rho\int_0^1 l^\varepsilon(s;v(s))ds + o(\rho), \quad \text{as } \rho\to 0.
\end{aligned}
\tag{3.22}
$$

Next choose the above I_ρ in (3.7), and we have

$$
\int_{I_\rho} l^\varepsilon(s;v(s))ds = \int_0^1 l^\varepsilon(s;u^{\varepsilon\rho}(s))ds. \tag{3.23}
$$

From (3.20)–(3.23), we conclude for given $v(\cdot)\in U_{\text{ad}}$ that

$$
\begin{aligned}
&\langle \lambda^\varepsilon Eh_y(y_0^\varepsilon, y^\varepsilon(1)) + \mu^{\varepsilon j} Ef_y^j(y_0^\varepsilon, y^\varepsilon(1)) + p^\varepsilon(0), \eta\rangle\rho + \rho\int_0^1 l^\varepsilon(s;v(s))ds\\
&\quad \geq -\rho\sqrt{\varepsilon+2\delta(\varepsilon)}\sqrt{1+|\eta|^2} + o(\rho), \quad \text{as } \rho\to 0.
\end{aligned}
\tag{3.24}
$$

Hence

$$\begin{aligned}(3.25)\qquad &\langle \lambda^\varepsilon E h_y(y_0^\varepsilon, y^\varepsilon(1)) + \mu^{\varepsilon j} E f_y^j(y_0^\varepsilon, y^\varepsilon(1)) + p^\varepsilon(0), \eta\rangle + \int_0^1 l^\varepsilon(s; v(s))ds \\ &\quad \geq -\sqrt{\varepsilon + 2\delta(\varepsilon)}\sqrt{1+|\eta|^2}, \quad \forall\, \eta \in \mathcal{R}^n \quad \text{and} \quad \forall\, v(\cdot) \in U_{\text{ad}}.\end{aligned}$$

This implies that

$$\begin{aligned}(3.26)\qquad &|p^\varepsilon(0) + \lambda^\varepsilon E h_y(y_0^\varepsilon, y^\varepsilon(1)) + \mu^{\varepsilon j} E f_y^j(y_0^\varepsilon, y^\varepsilon(1))| \leq C\sqrt{3\varepsilon}, \\ &\int_0^1 l^\varepsilon(s; v(s))ds \geq -\sqrt{\varepsilon + 2\delta(\varepsilon)}, \quad \forall\, v(\cdot) \in U_{\text{ad}}.\end{aligned}$$

Step 4. *Passing to the limit.* Without loss of generality, we assume that $\lambda^\varepsilon \to \lambda$, $\mu^\varepsilon \to \mu$, as $\varepsilon \to 0+$.

Let $\varepsilon \to 0+$. Equation $(3.6)_2$ gives the following:

$$\begin{aligned}(3.27)\qquad &E\int_0^1 \Big\{ H(y(s), v(s), \lambda, p(s), \{q^i(s)\}_1^d, r(s,\cdot)) - H(y(s), u(s), \lambda, p(s), \{q^i(s)\}_1^d, r(s,\cdot)) \\ &+ \frac{1}{2}\text{trace}\, P(s)\Big[\Delta b^i(s; v(s))\Delta b^{i*}(s; v(s)) + \int_{\mathcal{Z}} \Delta c(s,z; v(s))\Delta c^*(s,z; v(s))\pi(dz)\Big] \\ &+ \frac{1}{2}\text{trace}\int_{\mathcal{Z}} R(s,z)[\Delta c(s,z; v(s))\Delta c^*(s,z; v(s))]\,\pi(dz)\Big\} ds \geq 0, \qquad \forall v(\cdot) \in U_{\text{ad}};\end{aligned}$$

this implies (2.37). Furthermore, (2.34) is obtained from $(3.6)_1$, $(2.35)_2$ is obtained from $(3.26)_1$, and the rest of Theorem 2.1 is checked from (3.15)–(3.16).

Step 5. *The unbounded case of* U_{ad} *in* $L^{\infty,8}_{\mathcal{F},p}[[0,1];\mathcal{R}^m]$. This case can be treated via the bounded case. The details are as follows.

Set

$$(3.28)\qquad U_{\text{ad}}^K =: \{v(\cdot) \in U_{\text{ad}} |\, ||v(\cdot)|| \leq ||u(\cdot)|| + K\}, \quad K = 1, 2, \dots .$$

Obviously, we have

$$\begin{aligned}(3.29)\qquad &u(\cdot) \in U_{\text{ad}}^K \subset U_{\text{ad}}^{K+1}, \qquad K = 1, 2, \dots ; \\ &U_{\text{ad}} = \bigcup_{K=1}^{\infty} U_{\text{ad}}^K.\end{aligned}$$

Moreover, U_{ad}^K $(K = 1, 2, \dots)$ are bounded in $L^{\infty,8}_{\mathcal{F},p}[[0,1];\mathcal{R}^m]$, and the triplet $(y_0, y(\cdot), u(\cdot))$ is still optimal when the original admissible control set U_{ad} is replaced with U_{ad}^K $(K = 1, 2, \dots)$. Then, for each $K = 1, 2, \dots$, by steps 1–4, there exists

$$\{\lambda_K, \mu_K; p_K, \{q_K^i\}_{i=1}^d, r_K; P_K, \{Q_K^i\}_{i=1}^d, R_K\}$$

satisfying all the conditions of Theorem 2.1 except that the maximum condition (2.37) is replaced with the following (cf. (3.27)):

(3.30)
$$
\begin{aligned}
&E\int_0^1 \Big\{ H(y(s), v(s), \lambda_K, p_K(s), \{q_K^i(s)\}_1^d, r_K(s,\cdot)) \\
&- H(y(s), u(s), \lambda_K, p_K(s), \{q_K^i(s)\}_1^d, r_K(s,\cdot)) \\
&+ \frac{1}{2}\text{trace } P_K(s)\Big[\Delta b^i(s; v(s))\Delta b^{i*}(s; v(s)) + \int_{\mathcal{Z}} \Delta c(s,z; v(s))\Delta c^*(s,z; v(s))\pi(dz)\Big] \\
&+ \frac{1}{2}\text{trace}\int_{\mathcal{Z}} R_K(s,z)[\Delta c(s,z; v(s))\Delta c^*(s,z; v(s))]\,\pi(dz)\Big\}\,ds \geq 0, \qquad \forall v(\cdot) \in U_{\text{ad}}^K.
\end{aligned}
$$

Without loss of generality, we assume

$$\lambda_K \to \lambda, \quad \mu_K \to \mu \quad \text{as } K \to \infty.$$

Then we see that $\{p_K, \{q_K^i\}_{i=1}^d, r_K; P_K, \{Q_K^i\}_{i=1}^d, R_K\}_{K=1}^\infty$ is a Cauchy sequence. Let its limit be denoted by $\{p, \{q^i\}_{i=1}^d, r; P, \{Q^i\}_{i=1}^d, R\}$. By letting $K \to \infty$, we can obtain conditions 1), 2), and 4) of Theorem 2.1. The remainder is to prove the maximum condition 3).

For fixed $v(\cdot) \in U_{\text{ad}}$, we see from (3.29) and (3.30) that there exists K_0 such that

(3.31)
$$
\begin{aligned}
&E\int_0^1 \Big\{ H(y(s), v(s), \lambda_K, p_K(s), \{q_K^i(s)\}_1^d, r_K(s,\cdot)) \\
&- H(y(s), u(s), \lambda_K, p_K(s), \{q_K^i(s)\}_1^d, r_K(s,\cdot)) \\
&+ \frac{1}{2}\text{trace } P_K(s)\Big[\Delta b^i(s; v(s))\Delta b^{i*}(s; v(s)) + \int_{\mathcal{Z}} \Delta c(s,z; v(s))\Delta c^*(s,z; v(s))\pi(dz)\Big] \\
&+ \frac{1}{2}\text{trace}\int_{\mathcal{Z}} R_K(s,z)[\Delta c(s,z; v(s))\Delta c^*(s,z; v(s))]\,\pi(dz)\Big\}\,ds \geq 0, \qquad \text{as } K > K_0.
\end{aligned}
$$

Hence

(3.32)
$$
\begin{aligned}
&E\int_0^1 \Big\{ H(y(s), v(s), \lambda, p(s), \{q^i(s)\}_1^d, r(s,\cdot)) - H(y(s), u(s), \lambda, p(s), \{q^i(s)\}_1^d, r(s,\cdot)) \\
&+ \frac{1}{2}\text{trace } P(s)\Big[\Delta b^i(s; v(s))\Delta b^{i*}(s; v(s)) + \int_{\mathcal{Z}} \Delta c(s,z; v(s))\Delta c^*(s,z; v(s))\pi(dz)\Big] \\
&+ \frac{1}{2}\text{trace}\int_{\mathcal{Z}} R(s,z)[\Delta c(s,z; v(s))\Delta c^*(s,z; v(s))]\,\pi(dz)\Big\}\,ds \geq 0.
\end{aligned}
$$

Note that (3.31) holds for all $v(\cdot) \in U_{\text{ad}}$, and therefore the maximum condition 3) follows.

The proof of Theorem 2.1 is complete. □

Remark 3.1. In the proof of Theorem 2.1, I_ρ is chosen so that $\int_{I_\rho} l^\varepsilon(s; u^{\varepsilon\rho}(s))\,ds$ is differentiable with respect to $|I_\rho|$, and the derivative is $\int_0^1 l^\varepsilon(s; v(\cdot))\,ds$. When choosing $I_\rho = [t, t+\rho]$,$u^\rho(\cdot)$, defined by (2.1) is the so-called "spike-variation" control of the optimal control $u(\cdot)$. This is the approach in [17], [21] for the case $c \equiv 0$. When the control enters into neither diffusion nor jump terms, the spike-variation method is enough for the "differentiating" argument (step 3) to be rigorous; for this justification,

the reader is referred to [1], [10], [13], [17] for the case $c \equiv 0$. When the control appears in either diffusion or jump terms, however, the classical spike-variation method is not rigorous in the "differentiating" argument (because, even for the case of $c \equiv 0$, we are not sure that the stochastic process $q(\cdot)$ is continuous), and the I_ρ has to be carefully chosen as in step 3 so that the "differentiating" argument is rigorous.

4. Conclusion. The stochastic maximum principle with random jumps is quite different from that corresponding to the pure diffusion system. In calculating the first-order coefficient of the cost variation, we use only a property for Lebesgue integrals of scalar-valued functions in the real number space $\mathcal{R}$. This application can overcome the difficulty of "differentiation" caused by the appearence of the control in either diffusion or jump terms.

Appendix. The Proof of Lemma 2.3. In the Appendix, we provide the detailed proof of Lemma 2.3 concerning the Martingale representation result. The proof is an adaptation of that for Theorem 6.6 (see [16, p. 80]). Before starting the proof, we first give some lemmas.

Define

$$\begin{aligned} \mathcal{F}_t^w &=: \sigma[w(s); s \le t] \bigvee \mathcal{N}, \\ \mathcal{F}_t^k &=: \sigma[N_k(A, (0, s]); s \le t, A \in \mathcal{B}(\mathcal{Z})] \bigvee \mathcal{N}, \\ \mathcal{F}_t^* &=: \mathcal{F}_t^w \bigvee \mathcal{F}_t^k. \end{aligned} \tag{A1}$$

LEMMA A1. $\mathcal{F}_{t+0}^* = \mathcal{F}_t^*$.

Proof. We can check that $w(\cdot)$ is an $\mathcal{F}_t^*$-adapted Wiener process, and $k(\cdot)$ is an $\mathcal{F}_t^*$-adapted Poisson point process.

Let $0 \le t_1 < t_2 < \cdots < t_n$ and $f_1, f_2, \ldots, f_n \in C_0(\mathcal{R}^d)$,[1] $g_{ij} (i = 1, \ldots, n, j = 1, \ldots, m)$ be bounded functions from the positive integer set $\mathbb{N}$ to $\mathcal{R}$, and $U_1, \ldots, U_m \in \mathcal{B}(\mathcal{Z})$ be disjoint such that $\widehat{N}_k(U_j, (0, t]) =: \pi(U_j)t < \infty (j = 1, \ldots, m)$ for all $t > 0$. From the independence property of $\{w(t)\}$ and $\{N_k(U_j, (0, t])\}_{j=1}^m$ we have

$$\begin{aligned} &E\left[\prod_{i=1}^n f_i(w(t_i)) \prod_{i=1}^n \prod_{j=1}^m g_{ij}(N_k(U_j, (0, t_i]))| \mathcal{F}_t^*\right] \\ &= E\left[\prod_{i=1}^n f_i(w(t_i))| \mathcal{F}_t\right] \prod_{j=1}^m E\left[\prod_{i=1}^n g_{ij}(N_k(U_j, (0, t_i]))| \mathcal{F}_t^*\right]. \end{aligned}$$

Furthermore, if $t_{i^*-1} \le t < t_{i^*}$, we have

$$\begin{aligned} &E\left[\prod_{i=1}^n f_i(w(t_i))| \mathcal{F}_t^*\right] \\ &= \prod_{i=1}^{i^*-1} f_i(w(t_i)) H_{n-i^*+1}(t_{i^*} - t, t_{i^*+1} - t, \ldots, t_n - t; f_{i^*}, f_{i^*+1}, \ldots, f_n)(w(t)) \end{aligned}$$

[1] $C_0(\mathcal{R}^d)$ is the Banach space of all continuous functions on $\mathcal{R}^d$ such that $\lim_{|x| \to \infty} = 0$ with the maximum norm.

with

$$\begin{cases} H_1(t;f) =: H_t f, f \in C_0(\mathcal{R}^d), \\ H_t f(x) =: \int_{\mathcal{R}^d} p(t, x-y) f(y)\, dy, \ f \in C_0(\mathcal{R}^d), \\ p(t,x) =: (2\pi t)^{-d/2} \exp[-|x|^2/2t], \\ H_n(t_1, \dots, t_n; f_1, \dots, f_n) = H_{n-1}(t_1, \dots, t_{n-1}; f_1, \dots, f_{n-2}, f_{n-1} H_{t_n - t_{n-1}} f_n) \end{cases}$$

and

$$\begin{aligned} & E\left[\prod_{i=1}^{n} g_{ij}(N_k(U_j, (0,t_i]))| \mathcal{F}_t^*\right] \\ &= \prod_{i=1}^{i^*-1} g_{ji}(N_k(U_j, (0,t_i])) \\ &\quad \times \tilde{H}^j_{n-i^*+1}(t_{i^*} - t, t_{i^*+1} - t, \dots, t_n - t; g_{i^*}, g_{i^*+1}, \dots, g_n)(N_k(U_j, (0,t])) \end{aligned}$$

with

$$\tilde{H}_1^j(t;g) =: \tilde{H}_t^j g,$$

$$\begin{aligned} (\tilde{H}_t^j g)(m) &=: \prod_{n=1}^{\infty} g(n+m) \frac{[\hat{N}_k(U_j, (0,t])]^n e^{-\hat{N}_k(U_j,(0,t])}}{n!} \\ &=: \prod_{n=m}^{\infty} g(n) \frac{[\hat{N}_k(U_j, (0,t])]^{n-m} e^{-\hat{N}_k(U_j,(0,t])}}{(n-m)!} \\ &\quad \times \tilde{H}_n^j(t_1, t_2, \dots, t_n; g_1, g_2, \dots, g_n) \\ &=: \tilde{H}^j_{n-1}(t_1, t_2, \dots, t_{n-1}; g_1, g_2, \dots, g_{n-2}, g_{n-1} \tilde{H}^j_{t_n - t_{n-1}} g_n). \end{aligned}$$

Note that $H_{n-i^*+1}(t_{i^*} - \tau, t_{i^*+1} - \tau, \dots; f_{i^*}, f_{i^*+1}, \dots, f_n)(w(t))$ and $\tilde{H}^j_{n-i^*+1} \times (t_{i^*} - \tau, t_{i^*+1} - \tau, \dots; f_{i^*}, f_{i^*+1}, \dots, f_n)(m)$ are continuous at $\tau = t$ for fixed m. Paying attention to the right-continuity property with respect to τ of the Brownian motion $w(\tau)$ and the counting measure $N_k(U_j, (0,\tau])$, we have

$$\begin{aligned} E[\bullet\bullet\,|\mathcal{F}^*_{t+0}] &= \lim_{h\to 0+} E[E[\bullet\bullet\,|\mathcal{F}^*_{t+h}]|\mathcal{F}^*_{t+0}] \\ &= E[\lim_{h\to 0+} E[\bullet\bullet\,|\mathcal{F}^*_{t+h}]|\mathcal{F}^*_{t+0}] \\ &= E[E[\bullet\bullet\,|\mathcal{F}^*_t]|\mathcal{F}^*_{t+0}] \\ &= E[\bullet\bullet\,|\mathcal{F}^*_t], \end{aligned}$$

where $\bullet\bullet$ stands for $\prod_{i=1}^n f_i(w(t_i))$ or $\prod_{i=1}^n g_{ij}(N_k(U_j, (0,t_i])), j = 1, 2, \dots, m$. This proves the desired result. □

LEMMA A2. *For any increasing sequence $\{\sigma_n\}$ of $(\mathcal{F}_t^*)$-stopping times, we have*

$$\bigvee_n \mathcal{F}^*_{\sigma_n} = \mathcal{F}^*_\sigma,$$

where $\sigma = \lim_{n\to\infty} \sigma_n$. This property of $\mathcal{F}_t^*$ is called the quasi-left-continuous property.

Proof. Next we adopt the notations introduced in the proof of Lemma A1. The following is an adaptation of [16, p. 80]. Noting the independence property of the

Brownian motion and the Poisson point process as well as their strong Markov properties, we have

$$
\begin{aligned}
&E[\bullet\bullet|\mathcal{F}_\tau^*]\\
=&E\Big[\prod_{i=1}^n f_i(w(t_i))|\mathcal{F}_\tau^*\Big]\prod_{j=1}^m E\Big[\prod_{i=1}^n g_{ij}(N_k(U_j,(0,t_i]))|\mathcal{F}_\tau^*\Big]\\
=&\Big[\sum_{l=1}^n \chi_{\{t_{l-1}\le\tau<t_l\}}\prod_{i=1}^{l-1} f_i(w(t_i))H_{n-l+1}\\
&\times(t_l-\tau,t_{l+1}-\tau,\dots,t_n-\tau;f_l,f_{l+1},\dots,f_n)(w(\tau))\\
&+\prod_{i=1}^n f_i(w(t_i))\chi_{\{t_n\le\tau\}}\Big]\prod_{j=1}^m\Big[\sum_{l=1}^n\chi_{\{t_{l-1}\le\tau<t_l\}}\prod_{i=1}^{l-1}g_{ji}(N_k(U_j,(0,t_i]))\\
&\times\widetilde{H}_{n-l+1}^j(t_l-\tau,t_{l+1}-\tau,\dots,t_n-\tau;g_{jl},g_{j(l+1)},\dots,g_{jn})(N_k(U_j,(0,\tau]))\Big],
\end{aligned}
$$

for any $\mathcal{F}_t^*$-stopping time τ.

Paying attention to insure that the predictable dual projection $\widehat{N}_k(U_j,(0,t])$ of $\mathcal{F}_t^*$-Martingale $N_k(U_j,(0,t])$ is continuous with respect to t, we see that the $\mathcal{F}_t^*$-Martingale $N_k(U_j,(0,t])$ is quasi-left-continuous, almost surely,

$$\lim_{n\to\infty}N_k(U_j,(0,\sigma_n])=N_k(U_j,(0,\sigma]),\ \text{a.s.}$$

While obviously

$$\lim_{n\to\infty}w(\sigma_n)=w(\sigma),$$

by the continuity of the sample of $w(\cdot)$ with respect to t, we conclude the proof. □

Let $\mathcal{M}_2$ denote the space of $\mathcal{H}$-valued $\mathcal{F}_t^*$-adapted square-integrable Martingales $M(\cdot)$, satisfying $M(0)=0$. Write $L^2_{\mathcal{F}^*,p}$ for $L^2_{\mathcal{F}^*,p}[[0,1];\mathcal{H}]$, and F_p^2 for $F_p^2[[0,1];\mathcal{H}]$. Define

$$
\begin{aligned}
\mathcal{M}_2^*=:\Big\{&M(t)=\int_{(0,t]}q^i(s)dw^i(s)+\iint_{\mathcal{Z}\times(0,t]}r(s,z)\widetilde{N}_k(dzds)|\,q^i(\cdot)\in L^2_{\mathcal{F}^*,p},\\
&i=1,\dots,d;r(\cdot,\cdot)\in F_p^2.\Big\}
\end{aligned}
\tag{A2}
$$

LEMMA A3. *Let $M\in\mathcal{M}_2$ be bounded and suppose that $M\cdot N$ is $\mathcal{F}_t^*$-Martingale for every $N\in\mathcal{M}_2^*$. Then $M=0$.*

Proof. Let M^i denote the ith component of $M, i=1,\dots,n$, and $U_1\in\mathcal{B}(\mathcal{Z})$ satisfy $\hat{N}_k(U_1,(0,t])=:\pi(U_1)t<\infty$ for all $t>0$. Set $W(t)=w(t)\otimes q(U_1,(0,t]), t\in[0,1]$. We see that $M^iW, i=1,2,\dots,n$, are Martingales.

Assume $|M(t)|\le\alpha$, where α is a positive constant, and set $D^i(\omega)=:1+M^i(1,\omega)/2\alpha$, $i=1,2,\dots,n$. Then $D^i(\omega)\ge\frac{1}{2}$ and $E[D^i(\omega)]=1$. Define new probability measures $\widetilde{P}^i$ on $\mathcal{F}_1^*$ by

$$\widetilde{P}^i(B)=E[D^i(\omega)\chi_B(\omega)],\quad B\in\mathcal{F}_1^*;\quad i=1,2,\dots,n.$$

Then for every $\mathcal{F}_t^*$-stopping time $\sigma\in[0,1]$,

$$
\begin{aligned}
\widetilde{E}[W(\sigma)]&=E[D^i(\omega)W(\sigma)]=E[E[D^i(\omega)|\mathcal{F}_\sigma]W(\sigma)]\\
&=E[W(\sigma)]+\frac{1}{2\alpha}E[M^i(\sigma)W(\sigma)]=E[W(\sigma)]=0,
\end{aligned}
$$

because M^iW is an $\mathcal{F}_t^*$-Martingale, $i = 1, 2, \dots, n$.

Similarly, $\widetilde{E}[W(\sigma) \otimes W^j(\sigma) - \sigma[I \otimes \pi(U_1)] = 0$, because every component of the stochastic matrix process $W(\sigma) \otimes W^j(\sigma) - \sigma[I \otimes \pi(U_1)] = \int_0^t W(s) \otimes dW(s) + \int_0^t [dW(s)] \otimes W(s) + (0I) \otimes q(U_1, (0,t])$ belongs to $\mathcal{M}_2^*$,and hence $M^i\{W(\sigma) \otimes W^j(\sigma) - \sigma[I \otimes \pi(U_1)]\}$ is a matrix $\mathcal{F}_t^*$-Martingale, $i = 1, 2, \dots, n$. That is, both $t \mapsto W(t)$ and $t \mapsto W(\sigma) \otimes W^j(\sigma) - \sigma[I \otimes \pi(U_1)]$ are $\mathcal{F}_t^*$-Martingales with respect to the probabilities $\widetilde{P}^i, i = 1, 2, \dots, n$. By Theorem 6.3 of [16], $t \mapsto w(t)$ is an $\mathcal{F}_t^*$-Brownian motion and $t \mapsto k(t)$ is an $\mathcal{F}_t^*$-Poisson point process with respect to $\widetilde{P}^i, i = 1, 2, \dots, n$. This clearly implies that $P = \widetilde{P}^i$ on $\mathcal{F}_1^*, i = 1, 2, \dots, n$; hence we must have $D = 1$ almost surely, almost everywhere, $M = 0$ almost surely. The proof is finished. □

Proof of Lemma 2.3. Lemma 2.3 asserts that $\mathcal{M}_2 = \mathcal{M}_2^*$. To prove this, we first show that every $M \in \mathcal{M}_2$ can be expressed as

$$M(t) = M_1(t) + M_2(t), \tag{A3}$$

where $M_1 \in \mathcal{M}_2^*$ and $M_2 \in \mathcal{M}_2$ satisfies the following: $M_2 \cdot N$ is an $\mathcal{F}_t^*$-Martingale for all $N \in \mathcal{M}_2^*$. Clearly, such a decomposition is unique if it exists.

Let $\mathcal{H} = \{M_1(1) | M_1 \in \mathcal{M}_2^*\}$. It is easy to see that $\mathcal{H}$ is a closed subspace of $L_2(\Omega, P)$. Let $\mathcal{H}^\perp$ be the orthogonal complement of $\mathcal{H}$. Now, let $M \in \mathcal{M}_2$ be given. Then, since $M(1) \in L_2(\Omega, P)$, we have the orthogonal decomposition $M(1) = H_1 + H_2$, where $H_1 \in \mathcal{H}$ and $H_2 \in \mathcal{H}^\perp$. By definition, H_1 is of the form

$$H_1(\omega) = \int_0^1 \Phi(s)\, dw(s) + \iint_{\mathcal{Z} \times (0,1]} r(s,z)\, \widetilde{N}_k(dzds)$$

for some $\Phi \in L_2$ and $r \in F_p^2$. Let $M_2(t)$ be the right-continuous modification of $E[H_2 | \mathcal{F}_t^*]$. Then clearly

$$M(t) = M_1(t) + M_2(t), \qquad t \in [0,1],$$

where $M_1(t) = \int_0^t \Phi(s)\, dw(s) + \iint_{\mathcal{Z} \times (0,1]} r(s,z)\, \widetilde{N}_k(dzds)$. It remains to show that for every $N \in \mathcal{M}_2^*, t \mapsto M_2(t) \cdot N(t)$ is an $\mathcal{F}_t^*$-Martingale on $[0,1]$. For this it is sufficient to show that for every $\mathcal{F}_t^*$-stopping time σ such that $\sigma \le 1$, $E[M_2(\sigma) \cdot N(\sigma)] = 0$. However, if

$$N(t) = \int_0^t \Psi(s)\, dw(s) + \iint_{\mathcal{Z} \times (0,t]} r(s,z)\, \widetilde{N}_k(dzds),$$

then

$$\begin{aligned} N^\sigma(t) &=: N(t \wedge \sigma) \\ &= \int_0^t \Psi(s) \chi_{s \le \sigma} dw(s) + \iint_{\mathcal{Z} \times (0,t]} r(s,z) \chi_{s \le \sigma}\, \widetilde{N}_k(dzds) \in \mathcal{M}_2^*, \end{aligned}$$

and hence

$$\begin{aligned} E[N(\sigma) \cdot M_2(\sigma)] &= E[N(\sigma) \cdot E[M_2(1) | \mathcal{F}_\sigma^*]] \\ &= E[N(\sigma) \cdot M_2(1)] = E[N^\sigma(1) H_2] = 0. \end{aligned}$$

Thus far, we have completed the proof of the decomposition (A3).

Let $\widetilde{\mathcal{N}} =: \{M \in \mathcal{M}_2 | M \text{ is bounded}\}$. It is easy to see that $\widetilde{\mathcal{N}}$ is dense in $\mathcal{M}_2$. Let $M \in \widetilde{\mathcal{N}}$ and $M = M_1 + M_2$ be the decomposition of (A3). Since $M_1 \in \mathcal{M}_2^*$, we

can check that

$$E \sup_{0\le t\le 1} |M_1(t)| < \infty,$$

and then $M_1(t)$ are almost surely bounded in t. Thus there exists a sequence σ_n $(= \sigma_n(M_1))$ of $\mathcal{F}_t^*$-stopping times such that $\sigma_n \in [0,1], \sigma_n \uparrow 1$, and $M_1^{\sigma_n} = (M_1(t\wedge\sigma_n))$ is a bounded Martingale, $n = 1, 2, \ldots$. As we know, $M_1^{\sigma_n} \in \mathcal{M}_2^*$ and $M^{\sigma_n} = M_1^{\sigma_n} + M_2^{\sigma_n}$ is the decomposition of (A3) for M^{σ_n} since $M_2^{\sigma_n} \cdot N$ is an $\mathcal{F}_t^*$-Martingale for every $N \in \mathcal{M}_2^*$. Set $\widehat{\mathcal{N}} =: \{M^{\sigma_n} | n = 1, 2, \ldots, M \in \widehat{\mathcal{N}}\}$. Then by Lemma A2 it is easy to see that $\widehat{\mathcal{N}}$ is dense in $\mathcal{M}_2$. Furthermore, if $M = M_1 + M_2$ is the decomposition of (A3) for $M \in \widehat{\mathcal{N}}$, then both M_1 and M_2 are bounded. Hence $M_2 = 0$ by Lemma A3. This shows that $\widehat{\mathcal{N}} \subset \mathcal{M}_2^*$.

Since $\mathcal{M}_2^*$ is closed and $\widehat{\mathcal{N}}$ is dense in $\mathcal{M}_2$, we have $\mathcal{M}_2 \subset \mathcal{M}_2^*$. The proof is complete. □

Acknowledgments. The authors thank Professor Jiongmin Yong for helpful conversations related to this subject. The authors also thank the editor and the referees for their useful suggestions to improve the first version of the paper.

REFERENCES

[1] J. S. Baras, R. J. Elliott, and M. Kohlmann, *The partially observed stochastic minimum principle,*, SIAM J. Control Optim., 27 (1989), pp. 1279–1292.

[2] A. Bensoussan, *Lectures on Stochastic Control*, Lecture Notes in Mathematics, Vol. 972, Nonlinear Filtering and Stochastic Control, Proceedings, Cortona, 1981.

[3] ———, *Perturbation Methods in Optimal Control*, Dunod, Gauthier-Villars, 1988.

[4] J. M. Bismut, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.

[5] R. K. Boel, *Optimal control of jump processes*, Electronics Research Lab. Memo M448, University of California, Berkeley, CA, July 1974.

[6] R. K. Boel and P. Varaiya, *Optimal control of jump processes*, SIAM J. Control Optim., 15 (1977), pp. 92–119..

[7] M. Davis and R. Elliott, *Optimal control of a jump process*, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 40 (1977), pp. 183–202.

[8] I. Ekeland, *Sur les problems variationels*, Acad. Sci. Paris, 275 (1972), pp. 1057–1059.

[9] ———, *Nonconvex minimization problems*, Bull. Amer. Math. Soc. (NS), 1 (1979), pp. 443–474.

[10] R. J. Elliott, *The optimal control of a stochastic system*, SIAM J. Control Optim., 15 (1977), pp. 756–778.

[11] W. H. Fleming, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[12] U. G. Haussmann, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Stud., 6 (1976), pp. 34–58.

[13] ———, *The maximum principle for optimal control of diffusions with partial information*, SIAM J. Control Optim., 25 (1987), pp. 341–361.

[14] Y. Hu, *Maximum principle of optimal control for Markov processes*, Acta Mathematica Sinica, 33 (1990), pp. 43–56.

[15] Y. Hu and S. Peng, *Maximum principle for semilinear stochastic evolution control systems*, Stochastics Stochastics Rep., 33 (1990), pp. 159–180.

[16] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Kodansha, 1989.

[17] H. J. Kushner, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control, 10 (1972), pp. 550–565.

[18] X. Li and Y. Yao, *Maximum Principle of Distributed Parameter Systems with Time Lags, Distributed Parameter Systems*, Lecture Notes in Control and Information Sciences, Vol. 75,

Springer-Verlag, New York, 1985, pp. 410–427.

[19] X. Li and J. Yong, *Necessary conditions for optimal control of distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.

[20] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[21] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.

[22] R. Rishel, *A minimum principle for controlled jump processes*, Lecture Notes in Economics and Mathematical Systems, Vol. 107, Springer-Verlag, Berlin, Heidelberg, New York, 1975, pp. 493–508.

[23] R. Situ, *A maximum principle for optimal controls of stochastic systems with random jumps*, in Proc. National Conference on Control Theory and Its Applications, Qingdao, Shandong, People's Republic of China, October 1991.

[24] X. Y. Zhou, *The connection between the maximum principle and dynamic programming in stochastic control,*, Stochastics Stochastics Rep., 35 (1990), pp. 1–13.

[25] ———, *A unified treatment of maximum principle and dynamic programming in stochastic controls*, Stochastics Stochastics Rep., 36 (1991), pp. 137–161.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 5, pp. 1476–1502, September 1994

012

NONSMOOTH OPTIMUM PROBLEMS WITH CONSTRAINTS*

ZS. PÁLES† AND V. M. ZEIDAN‡

Abstract. This paper develops second-order necessary conditions for nonsmooth infinite-dimensional optimization problems with Banach space-valued equality and real-valued inequality constraints. Another constraint in the form of a closed convex set is also present. The objective function is the maximum over a parameter of functions $f(t,z)$ that are Lipschitz in z and upper semicontinuous in t. The inequality constraints $g(s,z)$ depend on a parameter s. The technique we use is a generalization of that of Dubovitskii and Milyutin. The second-order conditions obtained here are in terms of a certain function σ that disappears when the parameters and a certain set that derives from the given convex set are absent. The presence of the function σ and that set is due to the envelope-like effect discovered by Kawasaki.

Key words. nonsmooth analysis, second-order necessary conditions, Dubovitskii–Milyutin approach, inequality constraints with parameter, envelope-like effect

AMS subject classifications. 49B27, 49B36

1. Introduction.

The goal of this paper is to develop second-order necessary conditions for nonsmooth optimum problems with constraints. The prototype of such a problem arises in control theory, when state inequality constraints are to be satisfied, or when the objective functional is the maximum of smooth functionals depending on a parameter from a compact metric space (minimax problems).

A general discussion on second- and higher-order necessary conditions can be found in Levitin, Milyutin, and Osmolovskii [16]. The Dubovitskii–Milyutin scheme for second order necessary conditions was first presented in [7]. This method was applied by Ben-Tal and Zowe [4] to abstract infinite-dimensional programming problems. The inequality-like constraint $-g(x) \in K$ treated in [4] can be considered as infinitely many (real-valued) inequality constraints, and, as was discovered by Kawasaki [12]–[15], we must encounter an envelope-like effect caused by the constraints (see also Ioffe [9], [10]). The result of this effect is that the second-order necessary conditions involve terms different from the second derivatives of the data of the problem. These new terms are probably enough to fill the gap between necessary and sufficient conditions.

In §2 we consider abstract optimum problems. Generalizing the scheme of Dubovitskii and Milyutin ([6], [7]), we deduce necessary conditions that involve the concepts of descent, admissible, and tangent variations for a one-parameter family of directions. The suitable choice of the family of directions then leads to first-, second-, and higher-order necessary conditions. The method introduced here is applicable to Pareto optimum problems as well. We also recall the separation theorem of Dubovitskii and Milyutin for convex sets, which enables us to rewrite the necessary conditions, as an analytic form, in terms of affine functionals from the adjoint sets of descent, admissible, and tangent variations. The standard results of convex analysis are also recalled here.

The object of §3 is to introduce the first- and second-order Clarke's derivative and the function σ, which plays a key role in the determination of descent and admissible

* Received by the editors April 20, 1992; accepted for publication (in revised form) April 19, 1993. The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada.

† Institute of Mathematics, L. Kossuth University, H-4010 Debrecen, Pf. 12, Hungary. The second revised version of this paper was prepared while the author worked as a Humboldt Research Fellow at the University of Saarbrücken, Germany.

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

variations. This function σ appears first in the Key Lemma, which gives necessary and sufficient conditions for a quadratic inequality with function coefficients. The connection of the function σ to the function E of Kawasaki (see [13]–[15]) is described after the proof of the Key Lemma.

There are several notions of first- and second-order differentiation that are used to investigate optimization problems. One recent technique is due to Rockafellar [18], who introduced the concept of second-order epidifferentiation. (See also Ioffe [10].) As we see in §4, for our purposes Clarke's first derivative and the induced second-order directional derivative fit the best.

In §4 we deal with second-order descent variations. The objective functional considered there is of the form

$$F(z) = \max_{t \in T} f(t, z), \qquad z \in D,$$

where T is a compact metric space and D is an open set of a Banach space Z. The set of descent variations is completely described with the help of the first- and second-order Clarke's derivative and the function σ introduced in the previous section. Section 5 contains similar results for second-order admissible variations of inequality constraints. In the same section we also treat admissible variations of convex set constraints that are of great importance to control problems. Here we derive another form of the set introduced by Kawasaki for the case of convex cones (see [12]).

Second-order tangent variations of constraints $H(z) = 0$ are considered in §6. Here H maps Z into another Banach space. Several regularity assumptions have to be assumed on H, in order that the theorem of Lyusternik [17] on the tangent space of differentiable manifolds can be applied.

Section 7 contains the main result of this paper, a second-order Lagrange multiplier rule. We can automatically specialize it to a first-order rule taking the identically zero family of directions. This rule turns out to be sufficient to handle minimax problems, Pareto optima, inequality constraints consisting of "compactly many" inequalities, convex set constraints, and also equality constraints with one Banach space-valued equality and a finite number of real-valued equalities. Therefore this result can be used effectively in control problems, as we will show in a forthcoming paper. The result generalizes most of the known versions of the existing multiplier rules. For instance, some recent results of Ben-Tal and Zowe [4] and the results of Kawasaki [12],[13] are generalized. In [12] Kawasaki makes use of the so called Mangasarian–Fromovitz condition. Since Kawasaki uses tangent cones in describing the envelope-like effect, this condition seems to be important. In our approach the convex set constraints are handled in another way; therefore the results obtained do not involve the Mangasarian–Fromovitz condition.

2. Statement of the problem—General theory of variations. In this section we treat the notions and results introduced by Dubovitskii and Milyutin [6], [7] in a unified form. The advantage of our approach is that we can deal with first-, second-, and higher-order necessary conditions using the same scheme. We also formulate a result for optimum problems in the sense of Pareto. The sets of admissible, descent, and tangent variations introduced below turn out to have a certain relation, i.e., each set can be expressed with the use of one notion only.

Let Z be a Banach space (over $\mathbf{R}$), $D \subset Z$ be open, $F : D \to \mathbf{R}$ and $Q_1, \ldots, Q_N$,

Q_{N+1} subsets of Z. The problem $(\mathcal{P})$ is to minimize $F(z)$ subject to

$$z \in \bigcap_{i=1}^{N+1} Q_i. \tag{1}$$

A point $\hat{z} \in D$ is called a (*local*) *solution of* $(\mathcal{P})$ if $\hat{z} \in \cap_{i=1}^{N+1} Q_i$ and there exists a neighbourhood U of $\hat{z}$ such that

$$F(z) \geq F(\hat{z}) \quad \text{for all} \quad z \in \left(\bigcap_{i=1}^{N+1} Q_i \right) \cap U \cap D.$$

If there are given M objective functions $F_1, \ldots, F_M : D \to \mathbf{R}$, then we can speak about Pareto optimality subject to the above constraint (1). Denote this problem by $(\mathcal{P})^*$. A point $\hat{z} \in D$ is called a (*local*) *solution* of $(\mathcal{P})^*$ if $\hat{z} \in \cap_{i=1}^{N} Q_i$ and there exists a neighbourhood U of $\hat{z}$ such that

$$\text{for all} \quad z \in \left(\bigcap_{i=1}^{N+1} Q_i \right) \cap U \cap D \quad \exists j \text{ such that} \quad F_j(z) \geq F_j(\hat{z}).$$

Defining F by $F(z) := \max\left(F_1(z) - F_1(\hat{z}), \ldots, F_M(z) - F_M(\hat{z})\right)$, one can see that $\hat{z}$ is a solution of $(\mathcal{P})^*$ if and only if it is a solution of $(\mathcal{P})$ with this function F.

A function $d : [0, \varepsilon_0] \to Z$ will be called a *one-parameter family of directions* if it is continuous at 0 with $d(0) = 0$. In what follows, we introduce several concepts of variations of order k with respect to a given one-parameter family of directions d. We note that k need not be an integer in the definitions and theorems below, although it is in the usual applications.

DEFINITION 1. *A vector* $\overline{w} \in Z$ *is a* kth-order descent variation of F at $\hat{z}$ in the direction d *if there exists an* $\overline{\varepsilon} > 0$ *such that* $\hat{z} + d(\varepsilon) + \varepsilon^k(\overline{w} + w) \in D$ *and*

$$F\left(\hat{z} + d(\varepsilon) + \varepsilon^k(\overline{w} + w)\right) < F(\hat{z}),$$

whenever $0 < \varepsilon < \overline{\varepsilon}$ *and* $||w|| < \overline{\varepsilon}$. *We denote by* $\mathcal{W}_\delta^{(k)} = \mathcal{W}_\delta^{(k)}(F; \hat{z}, d)$ *the set of all such variations* $\overline{w}$. *This set* $\mathcal{W}_\delta^{(k)}$ *is always open.*

DEFINITION 2. *A vector* $\overline{w} \in Z$ *is called a* kth-order admissible variation of $Q \subset Z$ at $\hat{z}$ in the direction d *if there exists an* $\overline{\varepsilon} > 0$ *such that*

$$\hat{z} + d(\varepsilon) + \varepsilon^k(\overline{w} + w) \in Q$$

for $0 < \varepsilon < \overline{\varepsilon}$ *and* $||w|| < \overline{\varepsilon}$. *We denote this set by* $\mathcal{W}_\alpha^{(k)} = \mathcal{W}_\alpha^{(k)}(Q; \hat{z}, d)$, *which is always open.*

DEFINITION 3. *A vector* $\overline{w} \in Z$ *is said to be a* kth-order tangent variation of $Q \subset Z$ at $\hat{z}$ in the direction d *if there exist sequences* $\varepsilon_n > 0$ *and* $w_n \in Z$ *with* $\varepsilon_n \to 0$ *and* $w_n \to 0$ *such that*

$$\hat{z} + d(\varepsilon_n) + {\varepsilon_n}^k(\overline{w} + w_n) \in Q \quad \textit{for all} \quad n \in \mathbf{N}.$$

This set will be denoted by $\mathcal{W}_\tau^{(k)} = \mathcal{W}_\tau^{(k)}(Q; \hat{z}, d)$.

Remarks. We can observe that Definitions 1 and 3 slightly differ from that of [6], [4], [8]. However, the changes here make possible the following unification: All the notions can be expressed in terms of the tangent variations.

Denote by $\rho_Q(z)$ the distance of $z \in Z$ from Q; then it is easy to check that

$$\mathcal{W}_\tau^{(k)}(Q;\hat{z},d) = \left\{ \overline{w} \mid \liminf_{\varepsilon \to 0+} \frac{\rho_Q(\hat{z} + d(\varepsilon) + \varepsilon^k \overline{w})}{\varepsilon^k} = 0 \right\}.$$

We can also write

$$\mathcal{W}_\tau^{(k)}(Q;\hat{z},d) = \operatorname*{Limsup}_{\varepsilon \to 0+} \frac{1}{\varepsilon^k}\big(Q - \hat{z} - d(\varepsilon)\big);$$

here Limsup denotes the upper set-limit introduced in [2]. The following duality relation between $\mathcal{W}_\tau^{(k)}$ and $\mathcal{W}_\alpha^{(k)}$ can be easily proved:

$$\mathcal{W}_\alpha^{(k)}(Z \setminus Q;\hat{z},d) = Z \setminus \mathcal{W}_\tau^{(k)}(Q;\hat{z},d)$$

for all $Q \subset Z$, $\hat{z} \in Z$. If Q denotes the level set $\{z \in D \mid F(z) < F(\hat{z})\}$, then we have

$$\mathcal{W}_\delta^{(k)}(F;\hat{z},d) = \mathcal{W}_\alpha^{(k)}(Q;\hat{z},d).$$

Therefore, using the concept of tangent variations, the sets of admissible and descent variations can be determined.

If $k = 1$ and $d \equiv 0$, then the above relations show that $\mathcal{W}_\tau^{(1)}$ is the contingent cone of Q at $\hat{z}$ (cf. [2, p. 121]). In this case the set $\mathcal{W}_\alpha^{(1)}$ becomes the Dubovitskii–Milyutin cone defined in [2, p. 126]. The duality relation between these two cones is stated there in Lemma 4.1.4. (See also [6] for the origin of these concepts.)

If $k = 2$ and $d(\varepsilon) \equiv \varepsilon\overline{d}$ for some constant vector $\overline{d}$, then we obtain the concept of second variations due to [7]. To obtain variations of order k, we must take $d(\varepsilon) = P(\varepsilon)$, where $P(\varepsilon)$ is a polynomial in ε of degree $k-1$ with $P(0) = 0$. This concept of higher-order tangent variations corresponds to the higher-order contingent set defined in [2, §4.7]. In [2] two other tangent sets, the adjacent and circatangent (Clarke's) sets, are also defined; however these concepts do not play any role in our approach.

We can easily check that $\mathcal{W}_\alpha^{(k)}$, $\mathcal{W}_\delta^{(k)}$ and $\mathcal{W}_\tau^{(k)}$ do not depend on the terms of d of order higher than k; i.e., if $d_1, d_2 : [0,\varepsilon_0] \to Z$ are two families of directions, such that $\|d_1(\varepsilon) - d_2(\varepsilon)\|/\varepsilon^k \to 0$ as $\varepsilon \to 0$, then

$$\mathcal{W}_\alpha^{(k)}(F;\hat{z},d_1) = \mathcal{W}_\alpha^{(k)}(F;\hat{z},d_2), \qquad \mathcal{W}_\delta^{(k)}(Q;\hat{z},d_1) = \mathcal{W}_\delta^{(k)}(Q;\hat{z},d_2),$$

and

$$\mathcal{W}_\tau^{(k)}(Q;\hat{z},d_1) = \mathcal{W}_\tau^{(k)}(Q;\hat{z},d_2).$$

Now we state a general necessary condition for optimality. This result contains that of [6] and [7] by the above remarks. Its proof follows the ideas in [4]. But it is so brief that we present it here only to render the presentation complete.

THEOREM 1. *If $\hat{z}$ is a local minimum for the problem $(\mathcal{P})$, then*

$$\mathcal{W}_\delta^{(k)}(F;\hat{z},d) \cap \left(\bigcap_{i=1}^{N} \mathcal{W}_\alpha^{(k)}(Q_i;\hat{z},d) \right) \cap \mathcal{W}_\tau^{(k)}(Q_{N+1};\hat{z},d) = \emptyset$$

for all one-parameter families of variations d and for all $k \in]0,\infty[$.

Proof. Let us proceed by contradiction. If there exist d and k so that there is an element $\overline{w}$ in the intersection above, then there exists $\overline{\varepsilon} > 0$ such that

$$\begin{aligned} \hat{z} + d(\varepsilon) + \varepsilon^k(\overline{w} + w) &\in D, \\ F\left(\hat{z} + d(\varepsilon) + \varepsilon^k(\overline{w} + w)\right) &< F(\hat{z}), \\ \hat{z} + d(\varepsilon) + \varepsilon^k(\overline{w} + w) &\in \bigcap_{i=1}^{N} Q_i \end{aligned}$$

hold for all $||w|| < \overline{\varepsilon}$ and $0 < \varepsilon < \overline{\varepsilon}$. Furthermore, there exist sequences $\varepsilon_n > 0$, $w_n \in Z$ converging to zero such that

$$\hat{z} + d(\varepsilon_n) + {\varepsilon_n}^k(\overline{w} + w_n) \in Q_{N+1}$$

for all $n \in \mathbf{N}$.

Choose n_0 such that $\varepsilon_n < \overline{\varepsilon}$ and $||w_n|| < \overline{\varepsilon}$ for all $n > n_0$. Then the sequence $z_n := \hat{z} + d(\varepsilon) + {\varepsilon_n}^k(\overline{w} + w_n)$ converges to $\hat{z}$ and

$$z_n \in \bigcap_{i=1}^{N+1} Q_i \quad \text{and} \quad F(z_n) < F(\hat{z}) \quad \text{for all} \quad n > n_0.$$

However, this is in contradiction with the optimality of $\hat{z}$. □

In a similar manner, we can obtain necessary conditions for Pareto optima using the following analogue of Theorem 1.

THEOREM 1*. *If $\hat{z}$ is a local minimum for the problem $(\mathcal{P})^*$, then*

$$\left(\bigcap_{j=1}^{M} \mathcal{W}_\delta^{(k)}(F_j; \hat{z}, d)\right) \cap \left(\bigcap_{i=1}^{N} \mathcal{W}_\alpha^{(k)}(Q_i; \hat{z}, d)\right) \cap \mathcal{W}_\tau^{(k)}(Q_{N+1}; \hat{z}, d) = \emptyset$$

for all one-parameter families of variations d and for all $k \in]0, \infty[$.

To extract useful information from the necessary condition of Theorem 1 (and Theorem 1*), the following separation theorem of Dubovitskii and Milyutin is a fundamental one.

LEMMA 1. *Let $C_0, C_1, \ldots, C_K$ be given nonempty convex sets in Z such that $C_1, \ldots, C_K$ are open. Then*

$$\bigcap_{i=0}^{K} C_i = \emptyset$$

if and only if there exist affine functions $\varphi_0, \ldots, \varphi_K : Z \to \mathbf{R}$ (not simultaneously identically constant) such that

$$\sum_{i=0}^{K} \varphi_i = 0 \quad \text{and} \quad \varphi_i|_{C_i} \geq 0, \qquad i = 0, \ldots, K.$$

Proof. For the proof see [7]. □

Remark. To apply this separation theorem to the problem involved in Theorem 1, we need conditions on the data and on the one-parameter family of directions d so that the nonemptiness and the convexity of the sets $\mathcal{W}_\alpha^{(k)}$, $\mathcal{W}_\delta^{(k)}$, and $\mathcal{W}_\tau^{(k)}$ are

assured. A one-parameter family of directions is called *critical* for the problem $(\mathcal{P})$ if all these sets are convex and nonempty.

For a nonempty convex set $C \subset Z$, we define the adjoint set of C by

$$C^+ := \{\varphi : Z \to \mathbf{R} \mid \varphi \text{ is affine and } \varphi|_C \geq 0\}.$$

Now we list some additional results from convex analysis that we use in what follows. The proofs of Lemmas 2 and 4 can be obtained by using standard arguments of convex analysis that can be found in [1] and [11]. Lemma 3 is derived from [12, Lem. 5.4].

When $C \subset Z$ is the solution set of "compactly many" convex inequalities, then its adjoint set is described in the following lemma.

LEMMA 2. *Let T be a compact metric space, D is an open set in Z and $\gamma : T \times D \to \mathbf{R}$ be an upper semicontinuous function such that it is convex and uniformly locally Lipschitz in the second variable. Denote*

$$C := \{ z \in Z \mid \gamma(t, z) < 0, \forall t \in T \}.$$

Then C is open and convex. If $C \neq \emptyset$, then

$$C^+ = \{ \varphi : Z \to \mathbf{R} \mid \varphi \text{ is affine } \exists \mu \in \mathcal{M}(T) : \varphi(z) \geq -\int_T \gamma(t, z) d\mu, \ (z \in Z) \},$$

where $\mathcal{M}(T)$ denotes the set of all nonnegative bounded Borel measures on the Borel measurable subsets of T.

If $C = \emptyset$, then there exists a nonzero $\mu \in \mathcal{M}(T)$ such that

$$\int_T \gamma(t, z) d\mu(t) \geq 0 \quad \text{for all} \quad z \in Z.$$

When C is the conical hull of a translate of $\operatorname{cone}(Q^\circ - \hat{z})$, where Q is a convex set, then the description of its adjoint set requires the following notion: A linear functional $z^* : Z \to \mathbf{R}$ is called a *supporting functional of the convex set Q at $\hat{z} \in Q$* if $z^*(z) \geq z^*(\hat{z})$ holds true for all $z \in Q$. We denote this set by $Q^*(\hat{z})$.

LEMMA 3. *Let $Q \subset Z$ be a closed convex set with nonempty interior, $\hat{z} \in Q$, $\overline{d} \in \overline{\operatorname{cone}}(Q - \hat{z})$ and*

$$C := \operatorname{cone}(\operatorname{cone}(Q^\circ - \hat{z}) - \overline{d}).$$

(Here "cone" and "$\overline{\text{cone}}$" stand for the conical and closed conical hull, respectively.) Then an affine function $\varphi(z) = z^*(z) + c$ (where $z^* \in Z^*$, $c \in \mathbf{R}$) is bounded below on C if and only if $z^* \in Q^*(\hat{z})$ and $z^*(\overline{d}) = 0$. Moreover

$$C^+ = \{ \varphi = z^* + c \mid z^* \in Q^*(\hat{z}), \ z^*(\overline{d}) = 0, \ c \geq 0 \}.$$

Define the *support function* of a set C by

$$\delta^*(z^*; C) := \sup\{z^*(c) : c \in C\}.$$

When C is the inverse image of a convex set by a surjective linear operator then the description of C^+ is contained in the following lemma.

LEMMA 4. *Let Z and Y be Banach spaces and $A : Z \to Y$ be a bounded linear operator that maps Z onto Y and let $K \subset Y$ be a nonempty convex set. Denote*

$$C := \{ z \in Z \mid Az \in K \}.$$

Then

$$C^+ = \{\, \varphi : Z \to \mathbf{R} \mid \varphi \text{ is affine and } \exists y^* \in Y^* : \varphi(z) \geq -y^* Az + \delta^*(y^*; K),\ z \in Z \,\}.$$

The subject of the following sections is the determination of second-order descent, admissible, and tangent variations and their adjoint sets.

3. Basic concepts of differentiations, auxiliary results. In this section we recall first the concept of *Clarke's generalized derivative* and introduce a more general notion for functions that are pointwise suprema of families of functions.

If F is a real-valued locally Lipschitzian functional on an open set D of Z, and $\hat{z} \in D$, then *Clarke's generalized derivative* of F at $\hat{z}$ is defined by

$$F^\circ(\hat{z}; d) := \limsup_{(z,\varepsilon)\to(\hat{z},0+)} \frac{F(z+\varepsilon d) - F(z)}{\varepsilon}, \qquad d \in Z$$

(cf. [5]). Using this notion, we introduce the second-order directional derivative

$$F^{\circ\circ}(\hat{z}; d) := \limsup_{\varepsilon\to 0+} 2\frac{F(\hat{z}+\varepsilon d) - F(\hat{z}) - \varepsilon F^\circ(\hat{z}; d)}{\varepsilon^2}.$$

If F is a Fréchet-differentiable function at $\hat{z}$ then $F^\circ(\hat{z}; d) = F'(\hat{z})d$; furthermore if F is two times differentiable at $\hat{z}$, then $F^{\circ\circ}(\hat{z}; d) = F''(\hat{z})(d, d)$.

Now we are going to deal with functionals $F : D \to \mathbf{R}$ represented in the form

$$F(z) = \sup_{t\in T} f(t, z), \tag{2}$$

where T is a compact metric space, $f : T \times D \to \mathbf{R}$ is an upper semicontinuous function in the first variable and uniformly locally Lipschitzian at $\hat{z}$ in the second variable. (This means that there exists $\varepsilon > 0$ and $K \in \mathbf{R}$ such that $\|f(t, z') - f(t, z'')\| \leq K\|z' - z''\|$ if $\|z' - \hat{z}\| < \varepsilon$ and $\|z'' - \hat{z}\| < \varepsilon$.) Note that these conditions imply the upper semicontinuity of $(t, z) \mapsto f(t, z)$. Since upper semicontinuous functions attain their suprema on compact spaces, "sup" can be changed to "max" in (2). We denote this class of functions by $\mathcal{F}(T, D)$.

The study of functionals of this form is sufficient to handle problems with very general objective functions, minimax problems, Pareto optima, and inequality constraints as well. Now we introduce the following generalization of Clarke's derivative: For functions $f : T \times D \to \mathbf{R}$ with $f \in \mathcal{F}(T, D)$, define

$$f^\circ_{[T]}(t, \hat{z}; d) := \limsup_{(\tau,z,\varepsilon)\to(t,\hat{z};0+)} \frac{f(\tau, z+\varepsilon d) - f(\tau, z)}{\varepsilon},$$

$$f^{\circ\circ}_{[T]}(t, \hat{z}; d) := \limsup_{(\tau,\varepsilon)\to(t,0+)} 2\frac{f(\tau, \hat{z}+\varepsilon d) - f(\tau, \hat{z}) - \varepsilon f^\circ_{[T]}(\tau, \hat{z}; d)}{\varepsilon^2}.$$

Clearly, $f^\circ_{[T]}$ is a real-valued and $f^{\circ\circ}_{[T]}$ is an extended real-valued function. We note that $f^\circ_{[T]}$ is usually not identical with f° (and the same is true for $f^{\circ\circ}_{[T]}$ and $f^{\circ\circ}$). If T is discrete, then $f^\circ_{[T]} = f^\circ$ and $f^{\circ\circ}_{[T]} = f^{\circ\circ}$ obviously holds. If f is Fréchet differentiable and $(t, z) \to f'(t, z)$ is continuous on $T \times Z$, then $f^\circ_{[T]}(t, \hat{z}, d) = f^\circ(t, \hat{z}, d) = f'(t, \hat{z})(d)$; moreover, if f is twice Fréchet differentiable with continuous $(t, z) \to f''(t, z)$ on $T \times Z$, then $f^{\circ\circ}_{[T]}(t, \hat{z}, d) = f^{\circ\circ}(t, \hat{z}, d) = f''(t, \hat{z})(d, d)$ can also be checked. The most important properties of $f^\circ_{[T]}$ and $f^{\circ\circ}_{[T]}$ are summarized in the following lemma.

LEMMA 5. *Let $f \in \mathcal{F}(T, D)$. Then the functions $f^\circ_{[T]}$ and $f^{\circ\circ}_{[T]}$ have the following properties:*

(i) $f^{\circ}_{[T]}$ *is upper semicontinuous on* $T \times D \times Z$.

(ii) *For fixed* $t \in T$ *and* $\hat{z} \in D$, $d \mapsto f^{\circ}_{[T]}(t, \hat{z}; d)$ *is a positively homogeneous, subadditive and (globally) Lipschitz function.*

(iii) *For fixed* $\hat{z} \in D$ *and* $d \in Z$, $t \mapsto f^{\circ\circ}_{[T]}(t, \hat{z}; d)$ *is an upper semicontinuous function on* T.

(iv) *For fixed* $t \in T$ *and* $\hat{z} \in D$, $d \mapsto f^{\circ\circ}_{[T]}(t, \hat{z}; d)$ *is a positively quadratically homogeneous function.*

Proof. The proof of properties (i), (ii) is analogous to that of [5, Thm. 2.1.1]. Property (iii) can be verified similarly to (i). Finally, the proof of (iv) is obvious. □

In what follows, we prove a result necessary for describing descent and admissible variations. We introduce the following notation: If $\mathcal{C}$ is a condition for the points of T, then $T_{\mathcal{C}}$ denotes the set of those points of T, where $\mathcal{C}$ is satisfied, e.g., $T_{f>0}$ denotes the set $\{t \in T | f(t) > 0\}$, where $f : T \to \mathbf{R}$ is an arbitrary function. If H is a subset of T, then ∂H stands for the boundary of H.

KEY LEMMA. *Let* (T, ρ) *be a compact metric space and* $a, b, c : T \to \mathbf{R}$ *be upper semicontinuous functions. Define* $\sigma : T \to [-\infty, 0]$ *by*

$$\sigma(t) := \sigma_{a,b}(t) := \begin{cases} \liminf\limits_{\substack{\tau \to t \\ a(\tau) < 0,\ b(\tau) > 0}} \dfrac{b^2(\tau)}{4a(\tau)}, & \text{if} \quad t \in T_{a=0,\, b=0} \cap \partial(T_{a<0,\, b>0}), \\ 0, & \text{otherwise.} \end{cases}$$

Then σ *is a lower semicontinuous function and* $T_{\sigma<0}$ *is nowhere dense in* T. *Furthermore, the following statements are equivalent to each other*

(i)

$$\begin{array}{rcll} a(t) & \leq & 0 & \text{for all} \quad t \in T, \\ b(t) & \leq & 0 & \text{for all} \quad t \in T_{a=0}, \\ c(t) & < & \sigma(t) & \text{for all} \quad t \in T_{a=0,\, b=0}. \end{array}$$

(ii) *There exists* $\delta > 0$ *such that for all* $t_0 \in T$, *for all sequences* $t_n \in T$ *with* $t_n \to t_0$ *and* $\varepsilon_n > 0$ *with* $\varepsilon_n \to 0$,

$$a(t_n) + \varepsilon_n b(t_n) + {\varepsilon_n}^2 (c(t_0) + \delta) \leq 0$$

holds true for sufficiently large values of $n \in \mathbf{N}$.

Proof. First we show the lower semicontinuity of σ. Let $t_0 \in T_{a=0,\, b=0}$ be arbitrary and assume that σ is not lower semicontinuous at t_0. Then there exist $\delta > 0$ and a sequence t_n such that

$$\sigma(t_n) < \sigma(t_0) - \delta \quad \text{and} \quad 0 < \rho(t_n, t_0) < 1/2n \quad \text{for} \quad n \in \mathbf{N}. \tag{3}$$

Clearly, $t_n \in T_{a=0,\, b=0} \cap \partial(T_{a<0,\, b>0})$; therefore there exists $s_n \in T_{a<0,\, b>0}$ such that $\rho(s_n, t_n) < 1/2n$ and

$$\frac{b^2(s_n)}{4a(s_n)} < \begin{cases} \sigma(t_n) + \dfrac{1}{n}, & \text{if } \sigma(t_n) > -\infty, \\ -n, & \text{if } \sigma(t_n) = -\infty. \end{cases}$$

Then $s_n \to t_0$ as $n \to \infty$ and we have

$$\liminf_{n \to \infty} \sigma(t_n) = \liminf_{n \to \infty} \frac{b^2(s_n)}{4a(s_n)} \geq \sigma(t_0).$$

This inequality contradicts (3); therefore σ must be lower semicontinuous at t_0.

By definition of σ, $T_{\sigma<0} \subset T^* := T_{a\geq 0} \cap \partial(T_{a<0})$. This set is closed, since a is upper semicontinuous. Assume that t_0 is in the interior of $T_{a\geq 0}$. Then $t_0 \notin \partial(T_{a<0})$; therefore the interior of T^* is empty, i.e., T^* is nowhere dense in T. Thus $T_{\sigma<0}$ must be nowhere dense as well.

The proof of (i) $\Rightarrow$ (ii). By the first two inequalities of (i), we have $T_{a=0,\,b=0} = T_{a\geq 0,\,b\geq 0}$, thus $T_{a=0,\,b=0}$ is a compact subset of T. By the third inequality of (i), $c - \sigma$ is negative and upper semicontinuous on $T_{a=0,\,b=0}$; therefore the least upper bound of $c - \sigma$ on this set is a negative number; denote it by -2δ. Then we have $c(t) - \sigma(t) + \delta < 0$ for $t \in T_{a=0,\,b=0}$. If (ii) is not true with this δ, then there exist $t_0 \in T$, sequences $t_n \in T$ with $t_n \to t_0$ and $\varepsilon_n > 0$ with $\varepsilon_n \to 0$ such that

$$a(t_n) + \varepsilon_n b(t_n) + {\varepsilon_n}^2(c(t_0) + \delta) > 0 \tag{4}$$

holds for infinitely many values of $n \in \mathbf{N}$. By taking subsequences if necessary, we may assume (4) for all n. Taking the limsup as $n \to \infty$ in (4), we get $a(t_0) \geq 0$; therefore $a(t_0) = 0$ by the first inequality of (i). It follows from (4) that

$$b(t_n) + \varepsilon_n(c(t_0) + \delta) > 0 \quad \text{for all} \quad n \in \mathbf{N}.$$

Taking $n \to \infty$ again, we get $b(t_0) \geq 0$. Now the second condition of (i) implies $b(t_0) = 0$. If $a(t_n)$ were 0 for some n, then (4) and $c(t_0) + \delta < \sigma(t_0) \leq 0$ would imply $b(t_n) > 0$, which contradicts the second inequality in (i), and therefore $a(t_n) < 0$ for $n \in \mathbf{N}$. Now (4) yields

$$0 < a(t_n) + (c(t_0) + \delta)\left({\varepsilon_n}^2 + \varepsilon_n \frac{b(t_n)}{c(t_0) + \delta}\right) \leq a(t_n) - \frac{b^2(t_n)}{4(c(t_0) + \delta)},$$

whence we get

$$c(t_0) + \delta > \frac{b^2(t_n)}{4a(t_n)} \quad \text{for all} \quad n \in \mathbf{N}.$$

Now taking the liminf as $n \to \infty$, we arrive at

$$c(t_0) + \delta \geq \sigma(t_0),$$

which contradicts the choice of δ.

The proof of (ii) $\Rightarrow$ (i). Assume that (ii) is valid and let t_0 in T be fixed arbitrarily. Letting $t_n = t_0$, $\varepsilon_n = 1/n$ and applying (ii), we easily the first and second inequalities of (i) and $c(t_0) + \delta \leq 0$ if $t_0 \in T_{a=0,\,b=0}$, whence we get $c(t_0) < 0$ for these values of t_0. Therefore we must show $c(t) < \sigma(t)$ only for values $t \in T_{a=0,\,b=0} \cap \partial(T_{a<0,\,b>0})$. If t_0 is in this set, then we can find a sequence $t_n \in T_{a<0,\,b>0}$ converging to t_0 and satisfying

$$\sigma(t_0) = \lim_{n\to\infty} \frac{b^2(t_n)}{4a(t_n)}.$$

Define

$$\varepsilon_n := -\frac{b(t_n)}{2c(t_0) + \delta}, \qquad n \in \mathbf{N}.$$

Since b is upper semicontinuous at t_0,

$$\limsup_{n\to\infty} b(t_n) \le b(t_0) = 0.$$

On the other hand, $b(t_n) > 0$, therefore $b(t_n) \to 0$ as $n \to \infty$. As we have proved, $c(t_0) + \delta/2 < 0$, thus $\varepsilon_n > 0$ for $n \in \mathbf{N}$ and $\varepsilon_n \to 0$ as $n \to \infty$. Applying now (ii) to these sequences $t = t_n$ and $\varepsilon = \varepsilon_n$, we obtain (for large n)

$$c(t_0) + \delta/2 < \frac{b^2(t_n)}{4a(t_n)}.$$

Taking the limit of both sides, we arrive at $c(t_0) < \sigma(t_0)$. □

Remark. In [13] Kawasaki introduced the function E defined below: Let u and v be continuous functions on T with

$$u(t) \ge 0 \quad \text{for} \quad t \in T \quad \text{and} \quad v(t) \ge 0 \quad \text{for} \quad t \in T_{u=0}.$$

Let T_0 be the set of all $t \in T$ for which there exists a sequence $\{t_n\} \in T$ with

$$u(t_n) > 0, \qquad t_n \to t \quad \text{and} \quad -v(t_n)/u(t_n) \to +\infty \quad \text{as} \quad n \to \infty.$$

Denote by $\mathcal{T}_t$ the set all those sequences. Define $E_{u,v} := E$ at $t \in T_0$ by

$$E(t) = \sup_{\{t_n\}\in\mathcal{T}_t} \limsup_{n\to\infty} v(t_n)^2/4u(t_n).$$

For $t \in T_{u=0,v=0}$, $t \notin T_0$ let $E(t) = 0$, and let $E(t) = -\infty$ otherwise.

Now we can observe that

$$\sigma_{-u,-v}(t) = -E_{u,v}(t) \quad \text{for} \quad t \in T_{u=0,v=0}.$$

Therefore the last inequality of (i) can be rewritten as

$$c(t) < -E_{-a,-b}(t) \quad \text{for all} \quad t \in T.$$

Thus the function σ could be changed to E throughout the paper. However we prefer to keep our function σ, since it takes only finite values in the regular cases.

Ioffe [10] also defines a function e that plays a similar role as E in [12]. This function is then used to investigate the second-order epidifferentiability of a function given in the form (2).

4. Second-order descent variations. In this section we deal with the determination of descent variations. We give conditions that are sufficient for vectors $\overline{w}$ to be a descent variation of a function F given by the form (2). These sufficient conditions turn out to be very close to necessary ones, e.g., in the case of $\mathcal{C}^2$ functions. This result is known in [4] for the special case where T is a singleton and f is C^2. However, when T is not a singleton a result, close in nature, is given in [11]–[13], where f is assumed C^2 in z with partial derivatives continuous in t, and the underlying space Z is of finite dimension.

THEOREM 2. *Let T be a compact metric space, $D \subset Z$ be an open set, $\hat{z} \in D$ and f be in $\mathcal{F}(T, D)$. Let $F : D \to \mathbf{R}$ be defined by (2), and $d(\varepsilon) := \varepsilon\overline{d}$, $\varepsilon > 0$. Assume that $f^{\circ\circ}_{[T]}(t, \hat{z}; \overline{d})$ is finite for $t \in T$ and denote*

$$a(t) := f(t, \hat{z}) - F(\hat{z}), \qquad b(t) := f^{\circ}_{[T]}(t, \hat{z}; \overline{d}).$$

(*Obviously,* $a(t) \le 0$, $(t \in T)$.) *If*

$$b(t) \le 0, \quad (t \in T_{a=0}) \tag{5}$$

and $\overline{w} \in Z$ *satisfies*

$$f^{\circ}_{[T]}(t, \hat{z}; \overline{w}) + \tfrac{1}{2} f^{\circ\circ}_{[T]}(t, \hat{z}; \overline{d}) < \sigma_{a,b}(t) \quad \text{for} \quad t \in T_{a=0,b=0}, \tag{6}$$

(*where* $\sigma_{a,b}$ *is the function introduced in the previous section*). *Then*

$$\overline{w} \in \mathcal{W}^{(2)}_{\delta}(F; \hat{z}, d). \tag{7}$$

Conversely, assume that $f^{\circ}_{[T]}$, $f^{\circ\circ}_{[T]}$, *and* $\overline{d}$ *satisfy*

$$f^{\circ}_{[T]}(t, \hat{z}; w) = \lim_{(\tau,z,\varepsilon)\to(t,\hat{z},0+)} \frac{f(\tau, z + \varepsilon w) - f(\tau, z)}{\varepsilon}, \qquad (t \in T,\ w \in Z),$$

$$f^{\circ\circ}_{[T]}(t, \hat{z}; \overline{d}) = \lim_{(\tau,\varepsilon)\to(t,0+)} 2\frac{f(\tau, \hat{z} + \varepsilon\overline{d}) - f(\tau, \hat{z}) - \varepsilon f^{\circ}_{[T]}(\tau, \hat{z}; \overline{d})}{\varepsilon^2}, \qquad (t \in T),$$

and there exists w^* *such that* $f^{\circ}_{[T]}(t, \hat{z}; w^*) < 0$ *for all* $t \in T_{a=0,\,b=0}$. *Then* (5) *and* (6) *are also necessary in order that* (7) *be valid.*

Proof. In the proof we apply the following identity twice:

$$\begin{aligned} f(t, \hat{z} + \varepsilon\overline{d} + \varepsilon^2 w) = f(t, \hat{z}) &+ \varepsilon f^{\circ}_{[T]}(t, \hat{z}; \overline{d}) \\ &+ \varepsilon^2 \left(\frac{f(t, \hat{z} + \varepsilon\overline{d} + \varepsilon^2 w) - f(t, \hat{z} + \varepsilon\overline{d})}{\varepsilon^2} \right) \\ &+ \varepsilon^2 \left(\frac{f(t, \hat{z} + \varepsilon\overline{d}) - f(t, \hat{z}) - \varepsilon f^{\circ}_{[T]}(t, \hat{z}; \overline{d})}{\varepsilon^2} \right). \end{aligned} \tag{8}$$

Assume that $\overline{d}, \overline{w}$ satisfy the assumptions of the theorem, and denote

$$c(t) := f^{\circ}_{[T]}(t, \hat{z}; \overline{w}) + \frac{1}{2} f^{\circ\circ}_{[T]}(t, \hat{z}; \overline{d}) \quad \text{for} \quad t \in T.$$

Then a, b, c satisfy the assumptions of the Key Lemma and statement (i) also holds. Therefore we have (ii) satisfied for a certain positive value of δ. Now we are going to show that there exists $\overline{\varepsilon} > 0$ such that

$$F(\hat{z}) \ge f(t, \hat{z} + \varepsilon\overline{d} + \varepsilon^2\overline{w}) + \varepsilon^2\delta/2, \qquad 0 < \varepsilon < \overline{\varepsilon},\ t \in T. \tag{9}$$

We proceed by contradiction. If this inequality were not true for any $\overline{\varepsilon} > 0$, then, taking $\overline{\varepsilon} = 1/n$, we can find sequences ε_n, t_n such that $\varepsilon_n \to 0$ and

$$F(\hat{z}) < f(t_n, \hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2\overline{w}) + {\varepsilon_n}^2\delta/2, \tag{10}$$

holds for all $n \in \mathbf{N}$. Taking subsequences if necessary, we may assume that t_n converges to a point $t_0 \in T$. Applying (8) and the definition of $f^{\circ}_{[T]}$, $f^{\circ\circ}_{[T]}$, it follows from (10) that

$$0 < f(t_n, \hat{z}) - F(\hat{z}) + \varepsilon_n f^{\circ}_{[T]}(t_n, \hat{z}; \overline{d}) + {\varepsilon_n}^2 \left(f^{\circ}_{[T]}(t_0, \hat{z}; \overline{w}) + \tfrac{1}{2} f^{\circ\circ}_{[T]}(t_0, \hat{z}; \overline{d}) + \delta \right)$$

for $n > n_0$. However this inequality contradicts (ii) of the Key Lemma; thus (9) must be valid. The local Lipschitz property of f at $\hat{z}$ and (9) then yields that

$$F(\hat{z}) > F(\hat{z} + \varepsilon\overline{d} + \varepsilon^2(\overline{w} + w))$$

for small ε and w. Therefore $\overline{w} \in \mathcal{W}_\delta^{(2)}(F, \hat{z}; d)$.

To prove the converse of the statement, assume that $f_{[T]}^\circ$ and $f_{[T]}^{\circ\circ}$ satisfy the conditions of the theorem. If $\overline{w}$ is a descent variation of F at $\hat{z}$ in the direction d, then

$$F(\hat{z}) > f(t, \hat{z} + \varepsilon\overline{d} + \varepsilon^2(\overline{w} + w)) \quad \text{for} \quad 0 < \varepsilon < \overline{\varepsilon},\ \|w\| < \overline{\varepsilon},\ t \in T. \tag{11}$$

Using the homogeneity of $f_{[T]}^\circ$, we can find w^* such that $f_{[T]}^\circ(t, \hat{z}; -w^*) < 0$ for $t \in T_{a=0,\, b=0}$ and $\|w^*\| < \overline{\varepsilon}$. Since $t \mapsto f_{[T]}^\circ(t, \hat{z}; -w^*)$ is upper semicontinuous, hence, for some positive δ,

$$f_{[T]}^\circ(t, \hat{z}; -w^*) \le -2\delta \quad \text{for all} \quad t \in T_{a=0,\, b=0}. \tag{12}$$

Let $t_n \in T$ and $\varepsilon_n > 0$ be arbitrary sequences with $t_n \to t_0 \in T$ and $\varepsilon_n \to 0$. By the assumptions of the necessity part of the theorem and (8), we get from (11)

$$0 > f(t_n, \hat{z}) - F(\hat{z}) + \varepsilon_n f_{[T]}^\circ(t_n, \hat{z}; \overline{d}) + {\varepsilon_n}^2 \left(f_{[T]}^\circ(t_0, \hat{z}; \overline{w} + w^*) + \tfrac{1}{2} f_{[T]}^{\circ\circ}(t_0, \hat{z}; \overline{d}) - \delta \right).$$

for $n \ge n_0$. We have proved that statement (ii) of the Key Lemma is now satisfied with a, b described above and

$$c(t) := f_{[T]}^\circ(t, \hat{z}; \overline{w} + w^*) + \tfrac{1}{2} f_{[T]}^{\circ\circ}(t, \hat{z}; \overline{d}) - 2\delta, \qquad (t \in T).$$

Therefore (5) and, for $t \in T_{a=0,\, b=0}$

$$f_{[T]}^\circ(t, \hat{z}; \overline{w} + w) + \tfrac{1}{2} f_{[T]}^{\circ\circ}(t, \hat{z}; \overline{d}) - 2\delta < \sigma_{a,b}(t)$$

must be valid. By (12) and the subadditivity of $f_{[T]}^\circ$, we have

$$f_{[T]}^\circ(t, \hat{z}; \overline{w}) \le f_{[T]}^\circ(t, \hat{z}; \overline{w} + w) + f_{[T]}^\circ(t, \hat{z}; -w) \le f_{[T]}^\circ(t, \hat{z}; \overline{w} + w) - 2\delta,$$

whence we get (6), which was to be proved. □

Remark. Introducing the notation

$$\mathcal{V}_\delta^{(2)}(F, \hat{z}; \overline{d}) := \{\, \overline{w} \in Z \mid f_{[T]}^\circ(t, \hat{z}; \overline{w}) + \tfrac{1}{2} f_{[T]}^{\circ\circ}(t, \hat{z}; \overline{d}) < \sigma_{a,b}(t),\ t \in T_{a=0,\, b=0} \,\},$$

the statement of the theorem says that

$$\mathcal{V}_\delta^{(2)}(F; \hat{z}, \overline{d}) \subset \mathcal{W}_\delta^{(2)}(F; \hat{z}, d). \tag{13}$$

Using the properties of $f_{[T]}^\circ$, we can see that $\mathcal{V}_\delta^{(2)} = \mathcal{V}_\delta^{(2)}(F; \hat{z}, \overline{d})$ is always an open convex set (that can be empty). In order that $\mathcal{V}_\delta^{(2)}$ be nonempty, it is enough to assume that $\sigma_{a,b}$ has only finite values and that there exists $w^* \in Z$ such that $f_{[T]}^\circ(t, \hat{z}; w^*) < 0$ for all $t \in T_{a=0,\, b=0}$. In this case, we can find a positive δ such that $f_{[T]}^\circ(t, \hat{z}; w^*) < -\delta$. On the other hand, $t \mapsto \sigma_{a,b}(t) - \frac{1}{2} f_{[T]}^{\circ\circ}(t, \hat{z}; \overline{d})$ is lower semicontinuous; therefore it is bounded below by a negative real constant r. Now we can check that the element $\overline{w} = \lambda w^*$ is in $\mathcal{V}_\delta^{(2)}$ if $\lambda > -r/\delta$.

Since $T_{a=0,\,b=0}$ is compact, the set $\mathcal{V}_\delta^{(2)}$ is a special case of the set C dealt with in Lemma 2. Therefore the adjoint set of $\mathcal{V}_\delta^{(2)}$ can be obtained by the help of that result. By (13), we have $\mathcal{W}_\delta^{(2)^+} \subset \mathcal{V}_\delta^{(2)^+}$, i.e., the affine functionals from the adjoint set of $\mathcal{W}_\delta^{(2)}$ can be described.

We note that in [12]–[14] Kawasaki obtained the closure of the above set $\mathcal{V}_\delta^{(2)}$, since he investigated descent directions in the tangential sense, not in the sense of Definition 1.

5. Second-order admissible variations. In this section we are going to investigate two types of constraints $z \in Q$ and their corresponding admissible variations. The first case is when Q is given by a family of inequalities

$$Q := \{\, z \in D \mid g(s,z) \le 0\,, \forall s \in S \,\}, \tag{14}$$

The second case to be considered below is when Q is a closed convex set with nonempty interior.

For inequality constraints we have the following theorem.

THEOREM 3. *Let S be a compact metric space, $D \subset Z$ be an open set, and $g \in \mathcal{F}(S,D)$. Let Q be defined by* (14), *$\hat{z} \in Q$, and $d(\varepsilon) := \varepsilon\overline{d}$, $\varepsilon > 0$. Assume that $g^{\circ\circ}_{[S]}(s,\hat{z};\overline{d})$ is finite for $s \in S$ and denote*

$$a(s) := g(s,\hat{z}), \qquad b(s) := g^{\circ}_{[S]}(s,\hat{z};\overline{d}).$$

(Obviously, $a(s) \le 0$, $(s \in S)$.) If $a(s) < 0$ for all $s \in S$, then $\mathcal{W}_\alpha^{(2)}(Q;\hat{z},d) = Z$. If $S_{a=0} \neq \emptyset$,

$$b(s) \le 0, \qquad (s \in S_{a=0}) \tag{15}$$

and $\overline{w} \in Z$ satisfies

$$g^{\circ}_{[S]}(s,\hat{z};\overline{w}) + \tfrac{1}{2} g^{\circ\circ}_{[S]}(s,\hat{z};\overline{d}) < \sigma_{a,b}(s), \quad \text{for} \quad s \in S_{a=0,b=0}, \tag{16}$$

(where $\sigma_{a,b}$ is the function introduced in §3). *Then*

$$\overline{w} \in \mathcal{W}_\alpha^{(2)}(Q;\hat{z},d). \tag{17}$$

Conversely, if $g^{\circ}_{[S]}$, $g^{\circ\circ}_{[S]}$ and $\overline{d}$ satisfy similar conditions as $f^{\circ}_{[T]}$ and $f^{\circ\circ}_{[T]}$ in the converse part of Theorem 2, *then* (15) *and* (16) *are also necessary for* (17) *to be valid.*

Proof. If $a(s) = g(s,\hat{z}) < 0$ for all $s \in S$, then using the Lipschitz property of g, we can find $\varepsilon^* > 0$ such that $g(s,z) < 0$ for all $s \in S$ if $||z - \hat{z}|| < \varepsilon^*$. In this case, $\overline{\varepsilon} > 0$ can be determined such that $\big|\big|(\hat{z} + \varepsilon\overline{d} + \varepsilon^2(\overline{w} + w)) - \hat{z}\big|\big| < \varepsilon^*$ if $0 < \varepsilon < \overline{\varepsilon}$, $||w|| < \overline{\varepsilon}$, whenever $\overline{d}$ and $\overline{w}$ are fixed vectors. Then $g(s, \hat{z} + \varepsilon\overline{d} + \varepsilon^2(\overline{w} + w)) < 0$ for $0 < \varepsilon < \overline{\varepsilon}$, $||w|| < \overline{\varepsilon}$; therefore we have $\overline{w} \in \mathcal{W}_\alpha^{(2)}(Q;\hat{z},d)$.

If $S_{a=0}$ is not empty, then using Theorem 2 with g instead f, we can check that (15) and (16) imply $\overline{w} \in \mathcal{W}_\delta^{(2)}(G;\hat{z},d)$, where G is defined by $G(z) := \sup_{s\in S} g(s,z)$, $z \in D$. It easy to see that $\mathcal{W}_\delta^{(2)}(G;\hat{z},d) \subset \mathcal{W}_\alpha^{(2)}(Q;\hat{z},d)$, therefore (17) holds. The converse statement can be proved on the line that was followed in the proof of Theorem 2. □

Remark. The set described by (16) turns out to be convex and open, and its adjoint set can be obtained with the help of Lemma 2. Therefore nonnegative affine functions on $\mathcal{W}_\alpha^{(2)}(Q;\hat{z},d)$ can also be determined.

If $Q \subset Z$ is an arbitrary closed set with nonempty interior then define the signed distance of $z \in Z$ from Q by

$$\rho_Q(z) := \begin{cases} \inf\{||w - z|| \,|w \in Q\}, & \text{if} \quad z \in Z \setminus Q, \\ -\inf\{||w - z|| \,|w \in Z \setminus Q\}, & \text{if} \quad z \in Q. \end{cases}$$

Then it is easy to check that ρ_Q is a Lipschitz function on Z with Lipschitz constant 1, and Q can be described as the level set of ρ_Q:

$$Q = \{z \in Z | \rho_Q(z) \leq 0\}.$$

Therefore Theorem 3 implies the following corollary.

COROLLARY. *Let $Q \subset Z$ be an arbitrary closed set with nonempty interior, $\hat{z} \in Q$, and $d(\varepsilon) := \varepsilon\overline{d}$, $\varepsilon > 0$. If either $z \in Q^\circ$, or $z \in \partial Q$ and $\rho_Q{}^\circ(\hat{z}; \overline{d}) < 0$, then $\mathcal{W}_\alpha^{(2)}(Q; \hat{z}, d) = Z$. If $z \in \partial Q$, $\rho_Q{}^\circ(\hat{z}; \overline{d}) = 0$, $\rho_Q{}^{\circ\circ}(\hat{z}; \overline{d})$ is finite and $\overline{w} \in Z$ satisfies*

$$\rho_Q{}^\circ(\hat{z}; \overline{w}) + \frac{1}{2}\rho_Q{}^{\circ\circ}(\hat{z}; \overline{d}) < 0,$$

then $\overline{w} \in \mathcal{W}_\alpha^{(2)}(Q; \hat{z}, d)$.

Proof. We must only deal with the case when $z \in \partial Q$, $\rho_Q{}^\circ(\hat{z}; \overline{d}) = 0$, $\rho_Q{}^{\circ\circ}(\hat{z}; \overline{d})$ is finite. Then let S be a one-element set, say $S := \{0\}$, and define $g(0, z) := \rho_Q(z)$. Now the conditions of Theorem 3 are satisfied, and clearly $\sigma_{a,b} = 0$, since the set where $\sigma < 0$ must be nowhere dense in S. Thus (16) reduces to the inequality of the theorem, and therefore Theorem 3 yields $\overline{w} \in \mathcal{W}_\alpha^{(2)}(Q; \hat{z}, d)$. □

As we have seen in the previous corollary, Theorem 3 could be used to obtain the description of admissible variations of convex sets. However, in this case, a direct approach is more convenient here; see below.

If $Q \subset Z$ is a closed convex set with nonempty interior and $\hat{z} \in Q$, $\overline{d} \in Z$, then we define the following two sets:

$$Q^\circ(\hat{z}, \overline{d}) := \bigcup_{\overline{\varepsilon} > 0} \bigcap_{\substack{\varepsilon < \overline{\varepsilon} \\ ||w|| < \overline{\varepsilon}}} \left[\frac{1}{\varepsilon^2}(Q - \hat{z} - \varepsilon\overline{d}) + w\right]$$

and

$$Q(\hat{z}, \overline{d}) := \bigcup_{\delta(\,.\,)} \bigcap_{\varepsilon > 0} \left[\frac{1}{\varepsilon^2}(Q - \hat{z} - \varepsilon\overline{d}) + \delta(\varepsilon)B\right]$$

where, in the latter, the union is taken over all function $\delta :]0, \infty[\to [0, \infty[$ with $\lim_{\varepsilon \to 0} \delta(\varepsilon) = 0$ and B denotes the closed unit ball.

The set $Q(\hat{z}, \overline{d})$, when $Q = K$, a convex cone, was introduced and investigated by Kawasaki [12]. This set turned out to be important in the description of the envelope-like effect for the constraint $z \in K$. We can observe that, taking $d(\varepsilon) \equiv \varepsilon\overline{d}$, the relation

$$Q(\hat{z}, \overline{d}) \subset \mathcal{W}_\tau^{(2)}(Q; \hat{z}, d)$$

holds. (Equality cannot be stated here in most of the cases.)

The set $Q^\circ(\hat{z};\overline{d})$ is in a stronger connection with admissible variations, namely, it follows directly from the definition that

$$Q^\circ(\hat{z},\overline{d}) = \mathcal{W}_\alpha^{(2)}(Q;\hat{z},d).$$

The determination of the set $Q(\hat{z},\overline{d})$ seems more difficult, since there we have to take the union over a class of functions, not over real numbers. In what follows we list the most important properties of these two sets and we clarify the relationship between them.

We can easily see that both sets are convex, $Q^\circ(\hat{z},\overline{d})$ is open (since it is the set of admissible variations), $Q(\hat{z},\overline{d})$ is closed (for the proof see [12, Lem. 3.4] when Q is convex cone). Furthermore,

$$Q^\circ(\hat{z},\overline{d}) \subset Q(\hat{z},\overline{d}).$$

This inclusion is proved in a much sharper form in Theorem 4 below. If one of these sets is not empty, then $\overline{d} \in \overline{\text{cone}}(Q-\hat{z})$. Therefore this relation is necessary in order that $\overline{d}$ be critical.

THEOREM 4. *Let $Q \subset Z$ be a closed convex set with nonempty interior, $\hat{z} \in Q$ and $\overline{d} \in Z$ arbitrary. Then*

(i) $Q(\hat{z},\overline{d}) + \text{cone}(\text{cone}(Q^\circ - \hat{z}) - \overline{d}) \subset Q^\circ(\hat{z},\overline{d})$;

(ii) $Q^\circ(\hat{z},\overline{d}) \subset \text{cone}(\text{cone}(Q^\circ - \hat{z}) - \overline{d})$; *moreover, if $\overline{d} \in \text{cone}(Q-\hat{z})$, then the inclusion can be replaced by equality here*;

(iii) $\overline{Q^\circ(\hat{z},\overline{d})} = Q(\hat{z},\overline{d})$.

Proof.

Proof of (i). When $Q(\hat{z},\overline{d}) = \emptyset$, then there is nothing to prove (since we adopt the convention $\emptyset + H = \emptyset$). Let $\overline{w} \in Q(\hat{z},\overline{d})$ and $w_0 \in K := \text{cone}(\text{cone}(Q^\circ - \hat{z}) - \overline{d})$. Then there exists a function $\Delta :]0,\infty[\to Z$ with $\lim_{\varepsilon\to 0}\Delta(\varepsilon) = 0$ and $\lambda,\mu,\delta > 0$ such that

$$\hat{z} + \varepsilon\overline{d} + \varepsilon^2(\overline{w} + \Delta(\varepsilon)) \in Q \quad \text{for} \quad \varepsilon > 0$$

and

$$\hat{z} + \lambda\overline{d} + \lambda\mu(w_0 + w) \in Q \quad \text{for} \quad ||w|| < \delta.$$

Taking the convex combination of these inclusions and $\hat{z} \in Q$ with the weights

$$\frac{\mu^2}{(\varepsilon+\mu)^2}, \quad \frac{\varepsilon^2\mu}{(\varepsilon+\mu)^2\lambda}, \quad 1 - \frac{\mu^2}{(\varepsilon+\mu)^2} - \frac{\varepsilon^2\mu}{(\varepsilon+\mu)^2\lambda}$$

(where ε is so small that $\varepsilon(\mu-\lambda) < 2\lambda\mu$), whence follows that all the weights are positive), then we obtain

$$\hat{z} + \frac{\varepsilon\mu}{\varepsilon+\mu}\overline{d} + \left(\frac{\varepsilon\mu}{\varepsilon+\mu}\right)^2(\overline{w} + w_0 + w + \Delta(\varepsilon)) \in Q \tag{18}$$

for $||w|| < \delta$ and small $\varepsilon > 0$. If ε is small, then we also have

$$\frac{\delta}{2}B \subset \{w + \Delta(\varepsilon) \mid ||w|| < \delta\}.$$

Introducing the new variable $\theta = \varepsilon\mu/(\varepsilon + \mu)$, we can see that (18) reduces to

$$\hat{z} + \theta\overline{d} + \theta^2(\overline{w} + w_0 + w) \in Q \quad \text{for} \quad 0 < \theta < \theta_0,\ ||w|| < \delta/2.$$

Thus we obtain $\overline{w} + w_0 \in Q^\circ(\hat{z}, \overline{d})$, which was to be proved.

Proof of (ii). If $\overline{w} \in Q^\circ(\hat{z}, \overline{d})$, then

$$\overline{w} \in \frac{1}{\varepsilon^2}(Q^\circ - \hat{z} - \varepsilon\overline{d}) = \frac{1}{\varepsilon}\left(\frac{1}{\varepsilon}(Q^\circ - \hat{z}) - \overline{d}\right) \subset K$$

for small $\varepsilon > 0$. To prove the reversed inclusion when $\overline{d} \in \operatorname{cone}(Q - \hat{z})$, observe that in this case, $0 \in Q(\hat{z}, \overline{d})$. (If $\hat{z} + \overline{\varepsilon}\overline{d} \in Q$, then the convexity of Q and $\hat{z} \in Q$ yields $\hat{z} + \varepsilon\overline{d} \in Q$ for $\varepsilon < \overline{\varepsilon}$, and hence $0 \in Q(\hat{z}, \overline{d})$ easily follows.) Taking out the zero element of $Q(\hat{z}, \overline{d})$, the inclusion stated in (i) implies the reversed inclusion in (ii).

Proof of (iii). Using the closure of the inclusion (i) and taking $0 \in \overline{K}$, we get $Q(\hat{z}, \overline{d}) \subset \overline{Q^\circ(\hat{z}, \overline{d})}$. The reversed inclusion is a consequence of the elementary properties listed earlier. □

Remarks. The (iii) statement of the theorem shows that the two sets $Q^\circ(\hat{z}, \overline{d})$ and $Q(\hat{z}, \overline{d})$ can be used equivalently in the description of admissible variations. It follows from (i), (ii) and (iii), that the following relations are also valid:

(iv) $Q(\hat{z}, \overline{d}) + \overline{\operatorname{cone}}(\operatorname{cone}(Q - \hat{z}) - \overline{d}) \subset Q(\hat{z}, \overline{d})$;

(v) $Q(\hat{z}, \overline{d}) \subset \overline{\operatorname{cone}}(\operatorname{cone}(Q - \hat{z}) - \overline{d})$, moreover, if $\overline{d} \in \operatorname{cone}(Q - \hat{z})$, then the inclusion can be replaced by equality here.

These inclusions (iv) and (v) were already proved by Kawasaki in the case when Q is a convex cone (see [12, Lems. 5.3, 5.7]).

When we must determine the adjoint set of $Q^\circ(\hat{z}, \overline{d})$, then Lemma 3 should be applied. The relation (i) of the theorem shows that an affine function from the adjoint set is necessarily bounded below on $\operatorname{cone}(\operatorname{cone}(Q - \hat{z}) - \overline{d})$, thus its linear part satisfies the conditions listed in Lemma 3.

6. Second-order tangent variations. In this section we deal with tangent variations of sets Q determined by Banach space valued equations

$$Q := \{\, z \in D \mid H(z) = 0 \,\}.$$

The main tool we have to apply here is the following result of Lyusternik [17] (see also [1]).

LEMMA 6. *Let Z and Y be Banach spaces, $D \subset Z$ be an open set, $\hat{z} \in D$ and $H : D \to Y$ be strictly Fréchet differentiable at $\hat{z}$. Assume that $H'(\hat{z})$ maps Z onto Y. Then there exists a neighbourhood of U of $\hat{z}$, a positive constant K and a function $h : U \to Z$ such that*

$$H(z + h(z)) = H(\hat{z}) \quad \text{and} \quad ||h(z)|| \le K\,||H(z) - H(\hat{z})|| \tag{19}$$

holds for $z \in U$.

We note that this result is in fact the same as Graves's implicit function theorem (cf. Aubin and Frankowska [2]).

The *strict Fréchet differentiability of H at $\hat{z}$* means the following: There exists a bounded linear operator $H'(\hat{z}) : Z \to Y$ such that for all $\varepsilon > 0$, there exists $\delta(\epsilon)$ with

$$||H(z') - H(z'') - H'(\hat{z})(z' - z'')|| \le \varepsilon\,||z' - z''||$$

whenever z' and z'' satisfy $||z' - \hat{z}|| < \delta(\varepsilon)$ and $||z'' - \hat{z}|| < \delta(\varepsilon)$. It is easy to see that

$$H'(\hat{z})\overline{d} = \lim_{(z,\varepsilon)\to(\hat{z},0+)} \frac{H(z+\varepsilon\overline{d}) - H(z)}{\varepsilon}$$

holds for $\overline{d} \in Z$.

If H is strictly Fréchet differentiable at $\hat{z}$, $\overline{d} \in Z$, then we introduce the second-order weak directional derivative of H by the formula

$$H''(\hat{z};\overline{d}) := \left\{ z \in Z \mid \liminf_{\varepsilon\to 0+} \left\| z - 2\frac{H(\hat{z}+\varepsilon\overline{d}) - H(\hat{z}) - \varepsilon H'(\hat{z})\overline{d}}{\varepsilon^2} \right\| = 0 \right\}.$$

In other words, using the concept of the upper set-limit of [2],

$$H''(\hat{z};\overline{d}) = \operatorname*{Limsup}_{\varepsilon\to 0+} \left\{ 2\frac{H(\hat{z}+\varepsilon\overline{d}) - H(\hat{z}) - \varepsilon H'(\hat{z})\overline{d}}{\varepsilon^2} \right\}.$$

This set is possibly empty. If it is not empty, then we say that H is *twice weakly directionally differentiable at $\hat{z}$ in the direction $\overline{d}$*. If the above relations hold with "limsup" and "Liminf," respectively, then we speak about strong directional differentiability. In that case $H''(\hat{z};\overline{d})$ clearly consists of one point. When H is $\mathcal{C}^2$ near $\hat{z}$ then we can see that $H''(\hat{z};\overline{d}) = \{H''(\hat{z})(\overline{d},\overline{d})\}$.

The statement of the following theorem is well known for twice continuously differentiable functions [4]. The proof follows the standard argument in an improved form, where the above notion of the second-order directional derivative is exploited.

THEOREM 5. *Let Z and Y be Banach spaces, $D \subset Z$ be an open set, $\hat{z} \in D$, $d(\varepsilon) := \varepsilon\overline{d}$, and $H : D \to Y$ be strictly Fréchet differentiable at $\hat{z}$. Assume that $H(\hat{z}) = 0$ and $H'(\hat{z})$ maps Z onto Y. Define the set Q by* (18). *Then $\overline{w} \in \mathcal{W}_T^{(2)}(Q;\hat{z},d)$ if and only if $H'(\hat{z})\overline{d} = 0$, H is twice weakly directionally differentiable at $\hat{z}$ in the direction $\overline{d}$ and*

$$0 \in H'(\hat{z})\overline{w} + \tfrac{1}{2}H''(\hat{z};\overline{d}). \tag{20}$$

Proof. Assume that $\overline{w}$ is a tangent variation of Q at $\hat{z}$. Then we have sequences $\varepsilon_n > 0$ and $w_n \in Z$ converging to 0 such that

$$H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2(\overline{w} + w_n)) = 0 = H(\hat{z})$$

for $n \in \mathbf{N}$. Using the strict Fréchet differentiability of H, it follows from this equation that

$$\frac{H'(\hat{z})(\varepsilon_n\overline{d} + {\varepsilon_n}^2(\overline{w} + w_n))}{\varepsilon_n} \to 0 \qquad \text{as} \quad n \to \infty;$$

that is, $H'(\hat{z})\overline{d} = 0$. Then we have

$$\begin{aligned} 0 = {} & \frac{H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2(\overline{w} + w_n)) - H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2 w_n)}{{\varepsilon_n}^2} \\ & + \frac{H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2 w_n) - H(\hat{z} + \varepsilon_n\overline{d})}{{\varepsilon_n}^2} + \frac{H(\hat{z} + \varepsilon_n\overline{d}) - H(\hat{z}) - \varepsilon_n H'(\hat{z})\overline{d}}{{\varepsilon_n}^2} \end{aligned}$$

for $n \in \mathbf{N}$. Now observe that the first term converges to $H'(\hat{z})\overline{w}$ and the second to zero as $n \to \infty$. Therefore the third term must have a limit as well, i.e., $H''(\hat{z};\overline{d})$ is not empty. Taking the limit, we get (20).

To prove the sufficiency of the conditions, assume $H'(\hat{z})\overline{d} = 0$ and (20). This latter means that there exists a sequence $\varepsilon_n > 0$ with $\varepsilon_n \to 0$ such that

$$\lim_{n\to\infty} \frac{H(\hat{z} + \varepsilon_n\overline{d}) - H(\hat{z}) - \varepsilon_n H'(\hat{z})\overline{d}}{{\varepsilon_n}^2} = -H'(\hat{z})\overline{w}.$$

By Lyusternik's theorem, we have a function h that satisfies (19). Define

$$w_n := \frac{1}{{\varepsilon_n}^2} h(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2\overline{w})$$

for large values of n. Then the first equation in (19) yields

$$H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2(\overline{w} + w_n)) = H(\hat{z}) = 0,$$

i.e.,

$$\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2(\overline{w} + w_n) \in Q \quad \text{for} \quad n > n_0.$$

We have only to show that $w_n \to 0$ as $n \to \infty$. Applying the inequality of (19), we get

$$\begin{aligned} ||w_n|| &\le K \left|\left| \frac{H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2\overline{w}) - H(\hat{z})}{{\varepsilon_n}^2} \right|\right| \\ &= K \left|\left| \frac{H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2\overline{w}) - H(\hat{z} + \varepsilon_n\overline{d})}{{\varepsilon_n}^2} + \frac{H(\hat{z} + \varepsilon_n\overline{d}) - H(\hat{z})}{{\varepsilon_n}^2} \right|\right|. \end{aligned}$$

Taking the limit $n \to \infty$, we arrive at

$$\begin{aligned} &\limsup_{n\to\infty} ||w_n|| \\ &\quad \le K \left|\left| \lim_{n\to\infty} \frac{H(\hat{z} + \varepsilon_n\overline{d} + {\varepsilon_n}^2\overline{w}) - H(\hat{z} + \varepsilon_n\overline{d})}{{\varepsilon_n}^2} + \lim_{n\to\infty} \frac{H(\hat{z} + \varepsilon_n\overline{d}) - H(\hat{z})}{{\varepsilon_n}^2} \right|\right| \\ &\quad = K\,||H'(\hat{z})\overline{w} - H'(\hat{z})\overline{w}|| = 0, \end{aligned}$$

which was to be shown. □

The set of second-order tangent variations has turned out to be the set $(H'(\hat{z}))^{-1}(-\frac{1}{2}H''(\hat{z};\overline{d}))$. To guarantee the convexity of this set it is enough to assume that $H''(\hat{z};\overline{d})$ is convex. However, as we will see later, we can avoid this assumption if we take a convex subset of $-\frac{1}{2}H''(\hat{z};\overline{d})$ and form its inverse image by $H'(\hat{z})$. (In most of the applications the set $H''(\hat{z};\overline{d})$ consists of one point only.) In this way we usually do not obtain the whole set $\mathcal{W}_\tau^{(2)}$, but we obtain a big convex part of it that is still disjoint from the sets of descent and admissible variations. To describe the adjoint set of inverse images of nonempty convex sets, we must use Lemma 4 from the second section.

7. The Lagrange multiplier rule. In this section we give a multiplier rule for a large class of optimum problems. The general problem to be considered here is the following specification of $(\mathcal{P})$.

Assume that Z and Y are Banach spaces (over $\mathbf{R}$), $D \subset Z$ is nonempty and open, T and S are compact metric spaces, $f : T \times D \to \mathbf{R}$ and $g : S \times D \to \mathbf{R}$, $H : D \to Y$, furthermore, Q is a closed convex subset of Z with nonempty interior. Denote

$$F(z) = \sup_{t\in T} f(t, z).$$

The problem $(\overline{\mathcal{P}})$ is to minimize $F(z)$ subject to

$$z \in D: \quad g(s,z) \leq 0, \quad (s \in S), \quad z \in Q, \quad H(z) = 0.$$

Let $Q_1 := \{z \in D | g(s,z) \leq 0,\ s \in S\}$, $Q_2 := Q$ and $Q_3 := \{z \in D | H(z) = 0\}$. Then $(\overline{\mathcal{P}})$ is a special case of $(\mathcal{P})$ with $N = 2$.

Introducing the notation

$$G(z) = \sup_{s \in S} g(s,z),$$

the constraint $g(s,z) \leq 0$, $(s \in S)$ can be rephrased as $G(z) \leq 0$. However, as we will see later, the "pointwise" analysis of this constraint yields more effective necessary conditions than the analysis of the single inequality.

A point $\hat{z} \in D$ is called an *admissible point* for the problem $(\overline{\mathcal{P}})$ if $g(s,\hat{z}) \leq 0$, $(s \in S)$, $\hat{z} \in Q$ and $H(\hat{z}) = 0$.

A point $\hat{z} \in D$ is called a *regular point* for the problem $(\overline{\mathcal{P}})$ if

— $f \in \mathcal{F}(T,U)$ for some neighbourhood $U \subset D$ of $\hat{z}$;

— $g \in \mathcal{F}(S,U)$ for some neighbourhood $U \subset D$ of $\hat{z}$;

— H is strictly Fréchet differentiable at $\hat{z}$ and the range of the linear operator $H'(\hat{z})$ is a closed subspace of Y.

Let $\hat{z}$ be regular admissible point for the problem $(\overline{\mathcal{P}})$. Introduce the following notation: If $\overline{d}$ is an arbitrary direction then let

$$\sigma_{[f]}(t,\hat{z};\overline{d}) := \sigma_{a_f,b_f}(t), \quad \text{with } a_f(t) := f(t,\hat{z}) - F(\hat{z}), \qquad b_f(t) := f^{\circ}_{[T]}(t,\hat{z};\overline{d})$$

and

$$\sigma_{[g]}(s,\hat{z};\overline{d}) := \sigma_{a_g,b_g}(s), \quad \text{with } a_g(s) := g(s,\hat{z}), \qquad b_g(s) := g^{\circ}_{[S]}(s,\hat{z};\overline{d}).$$

A direction $\overline{d}$ is called a *regular direction* at $\hat{z}$ for our problem $(\overline{\mathcal{P}})$ if

— $f^{\circ\circ}_{[T]}(t,\hat{z};\overline{d})$ and $\sigma_{[f]}(t,\hat{z};\overline{d})$ are finite on T;

— $g^{\circ\circ}_{[S]}(s,\hat{z};\overline{d})$ and $\sigma_{[g]}(s,\hat{z};\overline{d})$ are finite on S;

— H is twice weakly directionally differentiable at $\hat{z}$ in the direction $\overline{d}$, that is $H''(\hat{z};\overline{d})$ is nonempty;

— $Q^{\circ}(\hat{z},\overline{d}) \neq \emptyset$.

A direction $\overline{d}$ is called a *critical direction* at $\hat{z}$ if

— $f^{\circ}_{[T]}(t,\hat{z};\overline{d}) \leq 0$ for $t \in T_{f(t,\hat{z})=F(\hat{z})}$;

— $g^{\circ}_{[S]}(s,\hat{z};\overline{d}) \leq 0$ for $s \in S_{g(s,\hat{z})=0}$;

— $H'(\hat{z};\overline{d}) = 0$;

— $\overline{d} \in \overline{\text{cone}}(Q - \hat{z})$.

We can see that zero is always a regular and critical direction at $\hat{z}$ for $(\overline{\mathcal{P}})$.

Our main result is the following theorem.

THEOREM 6. *Let $\hat{z}$ be a regular solution of the above problem $(\overline{\mathcal{P}})$. Then for all regular critical directions $\overline{d}$ and convex set $K \subset H''(\hat{z};\overline{d})$ there exist Lagrange multipliers $\mu \in \mathcal{M}(T)$, $\nu \in \mathcal{M}(S)$, $y^* \in Y^*$, and $z^* \in Z^*$ such that at least one of ν, μ, y^* is different from zero and the following relations hold:*

$$\text{supp } \mu \subset T_{f(t,\hat{z})=F(\hat{z})}, \qquad \text{supp } \nu \subset S_{g(s,\hat{z})=0}, \tag{21}$$

$$-z^* \in Q^*(\hat{z}) \quad \textit{and} \quad z^*(\overline{d}) = 0,$$

$$\int_T f^{\circ}_{[T]}(t,\hat{z};z)d\mu(t) + \int_S g^{\circ}_{[S]}(s,\hat{z};z)d\nu(s) + y^*(H'(\hat{z})z) + z^*(z) \geq 0 \tag{22}$$

for all $z \in Z$ *and*

$$\int_T \left(f^{\circ\circ}_{[T]} - 2\sigma_{[f]}\right)(t,\hat{z};\bar{d})d\mu(t) + \int_S \left(g^{\circ\circ}_{[S]} - 2\sigma_{[g]}\right)(s,\hat{z};\bar{d})d\nu(s) - d^*(-y^*;K) - 2\delta^*(z^*;Q^{\circ}(\hat{z},\bar{d})) \geq 0. \tag{23}$$

Proof. First we eliminate the trivial cases, when the existence of multipliers μ, ν, y^*, z^* is more or less obvious. Assume that $\bar{d}$ is a fixed regular critical direction, and $K \subset H''(\hat{z};\bar{d})$, then consider the following cases.

Case A. Define

$$a_f(t) := f(t,\hat{z}) - F(\hat{z}), \qquad b_f(t) := f^{\circ}_{[T]}(t,\hat{z};\bar{d})$$

and

$$V_f := \{\, w \in Z \mid f^{\circ}_{[T]}(t,\hat{z};w) + \tfrac{1}{2}f^{\circ\circ}_{[T]}(t,\hat{z};\bar{d}) < \sigma_{[f]}(t,\hat{z};\bar{d}),\ t \in T_{a_f=0,\,b_f=0} \,\}.$$

Now we show that if $V_f = \emptyset$ then there exists a nonzero μ such that the conditions of the theorem hold with this μ and $\nu = 0$, $y^* = 0$, $z^* = 0$.

Assume that $V_f = \emptyset$. With

$$\gamma(t,w) := f^{\circ}_{[T]}(t,\hat{z};w) + \tfrac{1}{2}f^{\circ\circ}_{[T]}(t,\hat{z};\bar{d}) - \sigma_{[f]}(t,\hat{z};\bar{d})$$

and $T_{a_f=0,\,b_f=0}$ instead of T, the last statement of Lemma 2 yields a nonzero measure $\mu \in \mathcal{M}(T)$ with supp $\mu \subset T_{a_f=0,\,b_f=0}$ such that

$$\int_T \left[f^{\circ}_{[T]}(t,\hat{z};w) + \frac{1}{2}f^{\circ\circ}_{[T]}(t,\hat{z};\bar{d}) - \sigma_{[f]}(t,\hat{z};\bar{d})\right] d\mu(t) \geq 0$$

for all $w \in Z$. Putting $w = 0$, we get (23). Substituting $w = \lambda z$, dividing by λ and taking the limit $\lambda \to \infty$ in the resulting inequality, we obtain (22).

Case B. Define

$$a_g(s) := g(s,\hat{z}), \qquad b_g(s) := g^{\circ}_{[S]}(s,\hat{z};\bar{d})$$

and

$$V_g := \{\, w \in Z \mid g^{\circ}_{[S]}(s,\hat{z};w) + \tfrac{1}{2}g^{\circ\circ}_{[S]}(s,\hat{z};\bar{d}) < \sigma_{[g]}(s,\hat{z};\bar{d}),\ s \in S_{a_g=0,\,b_g=0} \,\}.$$

If $V_g = \emptyset$ then an argument similar to Case A shows that there exists a nonzero measure $\nu \in \mathcal{M}(S)$ such that the conditions of the theorem are satisfied with $\mu = 0$, $y^* = 0$, $z^* = 0$ and this measure ν.

Case C. Assume that the range of $H'(\hat{z})$ is a proper subspace of Y. Then, since it is also closed by our regularity assumptions, there exists a nonzero linear functional $y_0 \in Y^*$ that is identically zero on the range of $H'(\hat{z})$. Then the requirements of the theorem can be satisfied with $\mu = 0$, $\nu = 0$, $z^* = 0$ and with $y^* = y_0$ or $y^* = -y_0$.

Summarizing our observations, the only case we have to deal with is when V_f and V_g are nonempty sets and the range of $H'(\hat{z})$ is the whole space Y.

Define

$$Q_1 := \{z \in D | g(s,z) \leq 0, \forall s \in S\}, \quad Q_2 := Q, \quad Q_3 := \{z \in D | H(z) = 0\}.$$

The point $\hat{z}$ is a solution of $(\overline{\mathcal{P}})$ if and only if it is a solution of $(\mathcal{P})$ with these sets Q_1, Q_2, and Q_3 and $N = 2$. Therefore, by Theorem 1, we have

$$\mathcal{W}_\delta^{(2)}(F;\hat{z},d) \cap \mathcal{W}_\alpha^{(2)}(Q_1;\hat{z},d) \cap \mathcal{W}_\alpha^{(2)}(Q_2;\hat{z},d) \cap \mathcal{W}_\tau^{(2)}(Q_3;\hat{z},d) = \emptyset.$$

On the other hand, Theorems 2–5 yield

$$V_f \subset \mathcal{W}_\delta^{(2)}(F;\hat{z},d), \qquad V_g \subset \mathcal{W}_\alpha^{(2)}(Q_1,\hat{z},d),$$

and

$$V_H := \{w \in Z | H'(\hat{z})w \in -\tfrac{1}{2}K\} \subset \mathcal{W}_\tau^{(2)}(Q_3;\hat{z},d),$$

since $\overline{d}$ is a critical direction. Let

$$V_Q := \mathcal{W}_\alpha^{(2)}(Q_2;\hat{z},d) = Q^\circ(\hat{z},\overline{d}) \neq \emptyset.$$

Thus

$$V_f \cap V_g \cap V_Q \cap V_H = \emptyset.$$

Since we have eliminated Cases A–C, these sets are nonempty; moreover, they are convex and the first three sets are open. Now Lemma 1, the Dubovitskii–Milyutin separation theorem, can be applied. Then there exist affine functions $\varphi_f, \varphi_g, \varphi_Q, \varphi_H : Z \to \mathbf{R}$ (not simultaneously identically constant) such that

$$\varphi_f \in V_f{}^+, \quad \varphi_g \in V_g{}^+, \quad \varphi_Q \in V_Q{}^+, \quad \varphi_H \in V_H{}^+,$$

and

$$\varphi_f + \varphi_g + \varphi_Q + \varphi_H = 0. \tag{24}$$

Applying Lemma 2, we can find measures $\mu \in \mathcal{M}(T)$ and $\nu \in \mathcal{M}(S)$ with $\operatorname{supp}\mu \subset T_{a_f=0,\,b_f=0}$ and $\operatorname{supp}\nu \subset S_{a_g=0,\,b_g=0}$ such that

$$\varphi_f(w) \geq -\int_T \left[f^\circ_{[T]}(t,\hat{z};w) + \frac{1}{2}f^{\circ\circ}_{[T]}(t,\hat{z};\overline{d}) - \sigma_{[f]}(t,\hat{z};\overline{d})\right] d\mu(t),$$

$$\varphi_g(w) \geq -\int_S \left[g^\circ_{[S]}(s,\hat{z};w) + \frac{1}{2}g^{\circ\circ}_{[S]}(s,\hat{z};\overline{d}) - \sigma_{[g]}(s,\hat{z};\overline{d})\right] d\nu(s).$$

Write the affine functional φ_Q into the form $\varphi_Q = -z^* + c$, where $z^* \in Z^*$ and $c \in \mathbf{R}$. Then, by Lemma 3 and Theorem 4, we have $-z^* \in Q^*(\hat{z})$ and $z^*(\overline{d}) = 0$. Since $\varphi_Q \in \big(Q^\circ(\hat{z},\overline{d})\big)^+$, we also have

$$\varphi_Q(w) = -z^*(w) + c \geq -z^*(w) + \delta^*(z^*;Q^\circ(\hat{z},\overline{d})).$$

It follows from Lemma 4 that there exist a linear functional $y^* \in Y^*$ such that

$$\varphi_H(w) \geq -y^*(H'(\hat{z})w) + \tfrac{1}{2}\delta^*(-y^*,K).$$

Then (24) yields

$$\begin{aligned} 0 \leq \quad & \int_T \left[f^\circ_{[T]}(t,\hat{z};w) + \frac{1}{2}f^{\circ\circ}_{[T]}(t,\hat{z};\overline{d}) - \sigma_{[f]}(t,\hat{z};\overline{d})\right] d\mu(t) \\ & + \int_S \left[g^\circ_{[S]}(s,\hat{z};w) + \frac{1}{2}g^{\circ\circ}_{[S]}(s,\hat{z};\overline{d}) - \sigma_{[g]}(s,\hat{z};\overline{d})\right] d\nu(s) \\ & + y^*(H'(\hat{z})w) - \frac{1}{2}\delta^*(-y^*;K) + z^*(w) - \delta^*(z^*;Q^\circ(\hat{z},\overline{d})) \end{aligned}$$

for all $w \in Z$.

Putting $w = 0$ into this equation, we obtain (23). To get (22), substitute $w = \lambda z$, divide by λ, and take the limit $\lambda \to \infty$.

To complete the proof of the theorem, we must show that μ, ν, y^* cannot be zero at the same time. In fact, if they were zero, then φ_f, φ_g, and φ_H would be nonnegative affine functions on Z. Then they must be constants, and by (24), φ_Q is then also a constant. However this contradicts the statement of the Dubovitskii–Milyutin separation theorem. □

Now we formulate two important corollaries of Theorem 6. In the first result the functions $\sigma_{[f]}$ and $\sigma_{[g]}$ and z^* do not play any role, but we have only a first-order necessary condition.

COROLLARY 1. *Let $\hat{z}$ be a regular solution of the above problem $(\overline{\mathcal{P}})$. Then there exist Lagrange multipliers $\mu \in \mathcal{M}(T)$, $\nu \in \mathcal{M}(S)$, and $y^* \in Y^*$ such that at least one of them is different from zero,* (21) *holds, and*

$$\int_T f^\circ_{[T]}(t, \hat{z}; w)d\mu(t) + \int_S g^\circ_{[S]}(s, \hat{z}; w)d\nu(s) + y^*(H'(\hat{z})w) \geq 0$$

for $w \in Q - \hat{z}$.

Proof. Let $\overline{d} = 0$. Then $\overline{d}$ is trivially a critical direction for the problem $(\overline{\mathcal{P}})$ at $\hat{z}$ and $H''(\hat{z}; \overline{d}) = \{0\}$; furthermore, $Q^\circ(\hat{z}, \overline{d}) = \operatorname{cone}(Q^\circ - \hat{z})$. Therefore $\overline{d} = 0$ is also regular for $(\overline{\mathcal{P}})$. Now, applying Theorem 6 for $\overline{d} = 0$ and $K = \{0\}$, we arrive at the above statement. □

The second corollary deals with the following generalization $(\hat{\mathcal{P}})$ of problem $(\overline{\mathcal{P}})$: Instead of the constraint $z \in Q$ we take $G(z) \in Q$, where Q is assumed to be a closed convex subset (with nonempty interior) of the Banach space W and $G : D \to W$. The notion of the admissibility of a point $\hat{z} \in D$ is defined analogously.

A point $\hat{z}$ is called regular if it is regular for problem $(\hat{\mathcal{P}})$ and, in addition, G is strictly Fréchet differentiable at $\hat{z}$.

A direction $\overline{d}$ is regular for $(\hat{\mathcal{P}})$ at $\hat{z}$ if it satisfies the first two conditions on regularity of directions to problem $(\overline{\mathcal{P}})$; furthermore, the mapping $(H, G) : D \to Y \times W$ is twice weakly differentiable at $\hat{z}$ in the direction $\overline{d}$ and $Q^\circ(G(\hat{z}), G'(\hat{z})\overline{d}) \neq \emptyset$. We can check that in order that the assumption on the second-order directional differentiability be satisfied it is necessary that both H and G be twice weakly differentiable, and it is sufficient that one of them be weakly and the other be strongly twice directionally differentiable at $\hat{z}$.

A direction $\overline{d}$ is called critical at $\hat{z}$ if the conditions of criticality for problem $(\overline{\mathcal{P}})$ are satisfied except the last one. Instead of the last one, we require now the following condition: $G'(\hat{z})\overline{d} \in \overline{\operatorname{cone}}(Q - G(\hat{z}))$.

Clearly, when $G(z) \equiv z$ then this problem $(\hat{\mathcal{P}})$ reduces to $(\overline{\mathcal{P}})$, and also the notions defined above reduce to that of $(\overline{\mathcal{P}})$, respectively.

COROLLARY 2. *Let $\hat{z}$ be a regular solution of the above problem $(\hat{\mathcal{P}})$. Then for all regular critical directions $\overline{d}$ and convex set $K \subset (H, G)''(\hat{z}; \overline{d})$ there exist Lagrange multipliers $\mu \in \mathcal{M}(T)$, $\nu \in \mathcal{M}(S)$, $y^* \in Y^*$, and $w^* \in W^*$ such that at least one of ν, μ, y^* is different from zero,* (21) *is valid, and the following relations hold:*

$$-w^* \in Q^*(G(\hat{z})) \quad \text{and} \quad w^*(G'(\hat{z})\overline{d}) = 0,$$

$$\int_T f^\circ_{[T]}(t, \hat{z}; z)d\mu(t) + \int_S g^\circ_{[S]}(s, \hat{z}; z)d\nu(s) + y^*(H'(\hat{z})z) + w^*(G'(\hat{z})z) \geq 0$$

for all $z \in Z$ *and*

$$\int_T \left(f^{\circ\circ}_{[T]} - 2\sigma_{[f]}\right)(t, \hat{z}; \overline{d})d\mu(t) + \int_S \left(g^{\circ\circ}_{[S]} - 2\sigma_{[g]}\right)(s, \hat{z}; \overline{d})d\nu(s)$$
$$-\delta^*(-(y^*, w^*); K) - 2\delta^*(w^*; Q^\circ(G(\hat{z}), G'(\hat{z})\overline{d})) \geq 0.$$

Proof. The proof of this corollary can be done with the following reduction trick: Introduce a new artificial variable $w \in W$, and write the constraint $G(z) \in Q$ as $G(z) - w = 0$ and $w \in Q$. Thus we obtain an optimum problem for the new variable (z, w). It is easy to check that the problem $(\hat{\mathcal{P}})$ rewritten in this form can be handled already with Theorem 6. The multipliers obtained for this new problem then exactly yield the statement of Corollary 2. □

Remarks. 1. Putting $z = \overline{d}$ in (22), we can see that the stronger relations

$$\operatorname{supp} \mu \subset T_{f(t,\hat{z})=F(\hat{z}),\, f^\circ_{[T]}(t,\hat{z};\overline{d})=0}, \qquad \operatorname{supp} \nu \subset S_{g(s,\hat{z})=0,\, g^\circ_{[S]}(s,\hat{z};\overline{d})=0},$$

are also valid instead of (21).

If the critical direction in problem $(\overline{\mathcal{P}})$ is taken from the smaller set $\operatorname{cone}(Q - \hat{z})$, then $Q^\circ(\hat{z}, \overline{d})$ turns out to be a cone (by Theorem 4); therefore the term containing z^* in the second-order necessary condition vanishes and the last condition for the regularity of $\overline{d}$, $Q^\circ(\hat{z}, d) \neq \emptyset$, holds. (An analogous statement is true for Corollary 2.)

When the multipliers are uniquely determined, then we can see that $\delta^*(-y^*;K)$ can be replaced by $\delta^*(-y^*; H''(\hat{z}, \overline{d}))$. This is important when H is twice weakly but not strongly directionally differentiable.

2. If there are given M objective functions $F_1, \ldots, F_M$ of the form

$$F_i(z) = \max_{t \in T_i} f_i(t, z), \qquad i = 1, \ldots, M;\ z \in D,$$

i.e., when the problem is a Pareto optimum problem, then introducing

$$T := \bigcup_{i=1}^{M} \{i\} \times T_i,$$
$$f((i,t), z) := f_i(t, z) - F_i(\hat{z}), \qquad (i,t) \in T,\ z \in D,$$

this Pareto optimum problem is equivalent to $(\overline{\mathcal{P}})$ with this set T and function f. Now the criticality and regularity of a direction $\overline{d}$ with respect to the objective function means that $\overline{d}$ is critical and regular with respect to $F_1, \ldots, F_M$. Applying Theorem 6, we obtain now that μ should be replaced by $\mu_1, \ldots, \mu_M$; f by f_i, T by T_i and the subformulae

$$\int_T f^\circ_{[T]}(t, \hat{z}; w)d\mu(t) \quad \text{and} \quad \int_T \left(f^{\circ\circ}_{[T]} - 2\sigma_{[f]}\right)(t, \hat{z}; \overline{d})d\mu(t)$$

should be replaced by

$$\sum_{i=1}^{M} \int_{T_i} f^\circ_{i\,[T_i]}(t, \hat{z}; w)d\mu_i(t) \quad \text{and} \quad \sum_{i=1}^{M} \int_{T_i} \left(f^{\circ\circ}_{i\,[T_i]} - 2\sigma_{[f_i]}\right)(t, \hat{z}; \overline{d})d\mu_i(t),$$

respectively, in the formulation of Theorem 6. If the sets $T_1, \ldots, T_M$ are singletons (one-element sets), then the measures μ_i can be interpreted simply as nonnegative

scalars, and the integration becomes multiplication by this scalar. In this case (since T_i is discrete) the functions $\sigma_{[f_i]}$ are zero and therefore do not occur in the formulation of the theorem.

3. A similar reasoning shows how Theorem 6 is transformed when, instead of $g(s,z) \leq 0$, there are given N inequalities:

$$g_i(s,z) \leq 0, \qquad i = 1,\dots,N.$$

4. If instead of $H(z) = 0$, we are given the following equality constraints:

$$H_0(z) = 0, \qquad h_1(z) = 0,\dots,h_N(z) = 0,$$

where $H_0 : D \to Y$ is strictly Fréchet differentiable at $\hat{z}$ with closed range $H_0'(\hat{z})(Z)$, and $h_i : D \to \mathbf{R}$ are also strictly differentiable functions at $\hat{z}$, then introduce

$$H(z) := (H_0(z), h_1(z),\dots,h_N(z)), \qquad z \in D.$$

Clearly $H : D \to Y \times \mathbf{R}^N$ is again strictly differentiable at $\hat{z}$ and it still satisfies the closed-range property by the closed image theorem of [1]. Now the equality constraints can equivalently be expressed as $H(z) = 0$; therefore Theorem 6 can be applied in this situation. To formulate the result that we can obtain now, $y \in H''(\hat{z};\overline{d})$ and $y^* \in Y^*$ should be replaced by $y_0 \in H_0''(\hat{z};\overline{d})$, $y_1 \in h_1''(\hat{z};\overline{d}),\dots,y_N \in h_N''(\hat{z};\overline{d})$, and $y^* \in Y^*$, $\lambda_1,\dots,\lambda_N \in \mathbf{R}$, respectively. The formulae

$$y^*(H'(\hat{z})w) \geq 0 \quad \text{and} \quad y^*(y)$$

should be replaced by

$$\sum_{i=1}^{N} \lambda_i h_i'(\hat{z})w + y^*(H_0'(\hat{z})w) \geq 0 \quad \text{and} \quad \sum_{i=1}^{N} \lambda_i y_i + y^*(y_0),$$

respectively, in the text of Theorem 6.

5. Using the above remarks, the results obtained in Theorem 6, or in Corollary 1, are more general than that of [5, Thm. 6.1.1], [1, Thm. 3.2.2], [11, Thm. 1, Chap. 1], [19, Thm. 48.B], and other known versions of the Lagrange multiplier rule. In [5], only the finite-dimensional case of equality constraints is considered; however the functions involved need not be strictly Fréchet differentiable, only locally Lipschitzian at $\hat{z}$. In our corollary we deal with infinite-dimensional equality constraints, but the strict Fréchet differentiability and the closed-range property is required. In [1, Thm. 3.4.2], second-order conditions are developed, where neither f in the objective function nor g in the inequality constraints is "time dependent." Furthermore, the set Q there is Z.

Theorem 6 and Corollary 2 include as a special case some recent results of Ben-Tal and Zowe [4] and Kawasaki [12]. Kawasaki already improved the result of Ben-Tal and Zowe, pointing out that when the constraint qualification $\overline{d} \in \operatorname{cone}(Q - \hat{z})$ is replaced by the less restrictive condition $\overline{d} \in \overline{\operatorname{cone}}(Q - \hat{z})$, then we must encounter an envelope-like effect, and as a result, a new term in the second-order necessary condition appears. However, Kawasaki could prove this result only assuming the twice continuous differentiability of the data and also the so-called Mangasarian–Fromovitz condition. The regularity conditions of our results are much weaker. Moreover, we did not need the Mangasarian–Fromovitz condition since we treated the convex set

constraint using admissible variations instead of tangent variations. The possibility of omitting this condition was already noted by Ioffe [9].

6. The first-order necessary conditions of Corollary 1 could also be expressed in the following subdifferential form:

$$0 \in \int_T \partial_{[T]} f(t, \hat{z}) d\mu(t) + \int_S \partial_{[S]} g(s, \hat{z}) d\nu(s) + y^* \circ H'(\hat{z}) - Q^*(\hat{z}),$$

where $\partial_{[T]} f(t, \hat{z})$ is the weak* convex hull of those linear functionals $z^* : Z \to \mathbf{R}$ for which there exists sequences $z_n \to \hat{z}$, $t_n \to t$ and $z_n{}^* \in \partial f(t_n, z_n)$ such that $z_n{}^*$ converges to z^* in the weak* topology. Here $\partial f(t_n, z_n)$ denotes the Clarke's subgradient of the function $z \to f(t_n, z)$ at $z = z_n$ (see [5, §2.1]). The set $\partial_{[S]} g(s, \hat{z})$ is defined analogously. These notions are introduced similarly in [5, §2.8]. It might happen that the second-order conditions of Theorem 6 can also be expressed in a subdifferential form if we define second-order subdifferentials in a suitable manner.

7. The question about the sufficiency of our necessary conditions is open. However in convex optimization problems or in smooth problems we can become convinced that they are close to sufficiency.

Now we present a simple example that shows that the presence of the function σ cannot be eliminated from the above multiplier rule.

Example. Let T be the closed unit ball in $\mathbf{R}^2$ and define

$$f(u, v, x, y) := (2u - 1)x + 2vy - u^2 - v^2, \qquad (u, v) \in T,\ (x, y) \in \mathbf{R}^2,$$
$$F(x, y) := \max_{(u,v) \in T} f(u, v, x, y), \qquad (x, y) \in \mathbf{R}^2.$$

It is easy to observe that $F(x, y) = x^2 + y^2 - x$ for $\|(x, y)\| < 1$; moreover, the maximum is then attained at $u = x$ and $v = y$. Let us consider the following optmization problem:

$$\text{Minimize } F(x, y) \quad \text{subject to } x - y^2 = 0.$$

For directions $(h, k) \in \mathbf{R}^2$ and (x, y), $(u, v) \in T$, we can easily obtain

$$f^{\circ}_{[T]}(u, v, x, y; h, k) = (2u - 1)h + 2vk, \quad \text{and} \quad f^{\circ\circ}_{[T]}(u, v, x, y; h, k) = 0.$$

Now we compute the function $\sigma_{[f]}$. Let x, y, h, and k be fixed as above and define

$$a(u, v) := f(u, v, x, y) - F(x, y) = -(x - u)^2 - (y - v)^2,$$
$$b(u, v) := f^{\circ}_{[T]}(u, v, x, y; h, k) = (2u - 1)h + 2vk.$$

for $(u, v) \in T$. Then $\sigma_{[f]}(u, v, x, y; h, k) = \sigma_{a,b}(u, v)$. By definition, we have that $\sigma_{a,b}(u, v) = 0$ if $(u, v) \notin T_{a=0,\, b=0}$. This latter set is nonempty if and only if $(2x - 1)h + 2yk = 0$ and then it is the singleton $\{(x, y)\}$. Let us determine $\sigma_{a,b}(x, y)$, when $(2x - 1)h + 2yk = 0$. By definition again,

$$\sigma_{a,b}(x, y) = \liminf_{\substack{(u,v) \to (x,y) \\ a(u,v) < 0,\ b(u,v) > 0}} \frac{b^2(u, v)}{4a(u, v)}$$
$$= \liminf_{\substack{(u,v) \to (x,y) \\ (2u-1)h + 2vk > 0}} \frac{((2u - 1)h + 2vk)^2}{-4[(x - u)^2 + (y - v)^2]}$$

$$= \liminf_{\substack{(u,v)\to(x,y)\\(2u-1)h+2vk>0}} \frac{(2(u-x)h+2(v-y)k)^2}{-4[(x-u)^2+(y-v)^2]}$$

$$= \liminf_{\substack{(u,v)\to(0,0)\\uh+vk>0}} \frac{(uh+vk)^2}{-[u^2+v^2]} = -\limsup_{\substack{(u,v)\to(0,0)\\uh+vk>0}} \frac{(uh+vk)^2}{[u^2+v^2]} = -(h^2+k^2).$$

(The last equality is a consequence of the Cauchy–Schwarz inequality.) Thus

$$\sigma_{[f]}(u,v,x,y;h,k) = -(h^2+k^2) \quad \text{for} \quad u=x,\ v=y, \quad \text{if } (2x-1)h+2yk=0$$

and is equal to zero otherwise.

Define $H(x,y) := x - y^2$. Then $H'(x,y)(h,k) = h - 2yk$ and $H''(x,y;h,k) = \{-2k^2\}$.

Now we can apply the multiplier rule of Theorem 6 to our problem. We are going to show that $(\hat{x},\hat{y}) = (0,0)$ is the only point where the necessary conditions of Theorem 6 are satisfied. Let $(\hat{x},\hat{y})$ be a solution of the above problem and (h,k) be a regular critical direction at $(\hat{x},\hat{y})$. Clearly, all directions (h,k) are regular now; therefore (h,k) is a critical direction if

$$(2\hat{x}-1)h + 2\hat{y}k \le 0 \quad \text{and} \quad h - 2\hat{y}k = 0.$$

By Theorem 6, there exists a measure μ and $\lambda := y^* \in \mathbf{R}$ (not all zero) such that

$$\int_T (2u-1)s + 2vtd\mu(u,v) + \lambda(s-2\hat{y}t) \ge 0, \quad \text{for} \quad (s,t)\in\mathbf{R}^2,$$

$$\int_T 2(h^2+k^2)d\mu(u,v) + \lambda(-2k^2) \ge 0$$

and the support of the measure μ is at most the one-element set $(\hat{x},\hat{y})$. Therefore integration on T by μ can be interpreted now as multiplication by a nonnegative scalar μ at the point $(u,v) = (\hat{x},\hat{y})$. Thus the above inequalities can be rewritten as

$$\mu[(2\hat{x}-1)s + 2\hat{y}t] + \lambda(s-2\hat{y}t) \ge 0, \quad \text{for} \quad (s,t)\in\mathbf{R}^2,$$
$$2\mu(h^2+k^2) + \lambda(-2k^2) \ge 0.$$

The first inequality yields

$$\mu(2\hat{x}-1) + \lambda = 0 \quad \text{and} \quad \mu\hat{y} = \lambda\hat{y}. \tag{25}$$

The multiplier μ must be different from zero, otherwise λ would also be zero (by the first equality in (25)). If $\hat{y}$ is not zero, then we get $\mu = \lambda$ from the second equation, and then the first yields $\mu\hat{x} = 0$. This means that $\hat{x}$ and also $\hat{y}$ must be zero. Then $\mu = \lambda = 1$ is a solution for the multipliers and the critical directions at $(0,0)$ are now of the form $(0,k)$, where $k\in\mathbf{R}$ is arbitrary. We can also see that the second-order inequality condition is satisfied with equality now. Therefore $(\hat{x},\hat{y}) = (0,0)$ is the only solution candidate of our problem. On the other hand, if $x = y^2$, then $F(x,y) = x^2 + y^2 - x = y^4 \ge 0$, thus $(0,0)$ is a solution.

Acknowledgment. We thank the referee, who called our attention to some recent works, especially those by Kawasaki.

REFERENCES

[1] V. M. ALEKSEEV, S. V. FOMIN, AND V. M. TIHOMIROV, *Optimal Control*, Nauka, Moscow, 1979. (In Russian.)

[2] J. P. AUBIN, AND H. FRANKOWSKA, *Set-valued analysis*, in Systems and Control: Foundations and Applications, Vol. 2, Birkhäuser, Boston, Basel, Berlin, 1990.

[3] A. BEN-TAL, *Second order theory of extremum problems*, in Extremal methods and system analysis, A. V. Fiacco and K. Kortanek, eds., Springer-Verlag, Berlin, 1980, pp. 336–356.

[4] A. BEN-TAL AND J. ZOWE, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming Study, 19 (1982), pp. 39–76.

[5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Series Monographs Adv. Texts, John Wiley, New York, 1983.

[6] A. YA. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems with constraints*, Dokl. Akad. Nauk SSSR, 149 (1963), pp. 759–762; Soviet Math. Dokl., 4 (1963), pp. 452–455.

[7] ———, *Second variations in extremal problems with constraints*, Dokl. Akad. Nauk SSSR, 160 (1965), pp. 18–21.

[8] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Lecture Notes in Econom. Math. Systems, Vol. 67, Springer-Verlag, New York, Berlin, 1972.

[9] A. D. IOFFE, *On some recent developments in the theory of second order optimality conditions*, in Optimization, S. Dolecki, ed., Lecture Notes in Math., Vol. 1405, Springer-Verlag, New York, Berlin, 1989, p. 55–68.

[10] ———, *Variational analysis of a composite function: a formula for the lower second order epi-derivative*, J. Math. Anal. Appl., 160 (1991), pp. 379–405.

[11] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of extremal problems*, North-Holland, Amsterdam, 1979.

[12] H. KAWASAKI, *An envelope like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Programming, 41 (1988), pp. 73–96.

[13] ———, *The upper and second order directional derivatives for a sup-type function*, Math. Programming, 41 (1988), pp. 327–339.

[14] ———, *Second order necessary optimality conditions for minimizing a sup type function*, Math. Programming, 49 (1991), pp. 213–229.

[15] ———, *Second order necessary and sufficient optimality conditions for minimizing a sup type function*, Appl. Math. Optim., 26 (1992), pp. 195–220.

[16] E. S. LEVITIN, A. A. MILYUTIN, AND N. P. OSMOLOVSKII, *Higher order conditions for a local minimum in problems with constraints*, Uspehi Math. Nauk, 33 (1978), pp. 85–148.

[17] L. A. LYUSTERNIK, *On extremum conditions for functionals*, Math. Sbornik, 41 (1934), pp. 390–401.

[18] R. T. ROCKAFELLAR, *Second order optimality conditions in nonlinear programming obtained by way of epi-derivatives*, Math. Oper. Res., 14 (1989), pp. 462–484.

[19] E. ZEIDLER, *Nonlinear Functional Analysis and its Applications*, Vol III, Springer-Verlag, New York, Berlin, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1503–1541, November 1994

ROBUST STABILITY OF LINEAR EVOLUTION OPERATORS ON BANACH SPACES*

D. HINRICHSEN† AND A. J. PRITCHARD‡

Abstract. This paper introduces a stability radius for a wide class of linear infinite-dimensional time-varying systems under structured time-varying perturbations. A framework is presented that allows the same degree of unboundedness in the perturbations as in the generator of the nominal model.

Key words. infinite-dimensional systems, time-varying, evolution operators, Cauchy problem, robust stability, structured perturbations

AMS subject classifications. 93D09, 93C73, 93G50, 93C25

Notation.

$X, \underline{X}, \overline{X}, U_i, Y_i, U, Y$	Banach spaces over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ with norms $\|\cdot\|_X$, etc.
$\mathcal{L}(X,Y)$ (respectively, $\mathcal{L}(X)$)	Banach space of bounded linear operators from X to Y (respectively, on X) provided with the operator norm $\|\cdot\|_{\mathcal{L}(X,Y)}$ (respectively, $\|\cdot\|_{\mathcal{L}(X)}$)
$\mathcal{U}(X)$	$\{L \in \mathcal{L}(X);\ L$ invertible in $\mathcal{L}(X)\}$, the set of invertible bounded linear operators on X provided with the operator norm
$C(s,t;X)$	The set of X-valued continuous functions on $[s,t]$, $s \le t \le \infty$
$PC(\mathbb{R}_+, \mathcal{L}(X,Y))$	The set of piecewise continuous $\mathcal{L}(X,Y)$-valued operator functions on $\mathbb{R}_+ = [0,\infty)$
$PC_b(\mathbb{R}_+, \mathcal{L}(X,Y))$	The set of bounded operator functions in $PC(\mathbb{R}_+, \mathcal{L}(X,Y))$
$PC^1(s,t;\mathcal{U}(X))$	The set of continuous piecewise continuously differentiable functions from $[s,t]$ into the set $\mathcal{U}(X)$ of invertible linear operators on X
$L^p(s,t;X)$	The set of strongly measurable p-integrable X-valued functions on $[s,t]$, $1 \le p < \infty$
$L^p_{loc}(s,\infty;X)$	The set of strongly measurable locally p-integrable X-valued functions on $[s,\infty)$
$L^\infty(s,t;X)$	The set of strongly measurable functions $h : [s,t] \to X$ such that $\sup_{\tau\in[s,t]} \|h(\tau)\|_X < \infty$, where sup denotes the essential supremum
$\underline{N}$	$= \{1, \dots, N\}$

1. Introduction. A systematic theory of infinite-dimensional time-varying differentiable equations

$$\dot{x}(t) = A(t)x(t), \qquad t \ge 0, \tag{1}$$

where the $A(t)$ are unbounded linear operators on a Banach space X, was initiated in the fifties by Kato [14]. He approximated the fundamental solution of (1) by fundamental solutions

* Received by the editors May 6, 1992; accepted for publication (in revised form) May 5, 1993.
† Institut für Dynamische Systeme, Universität Bremen, D-2800 Bremen 33, Germany.
‡ Control Theory Centre, University of Warwick, Coventry CV4 7AL, United Kingdom.

corresponding to piecewise constant generators. Using the theory of holomorphic semigroups Tanabe [25] constructed a fundamental solution for (1) by representing the system generator as a time-varying perturbation of a time-invariant generator. For both approaches an essential assumption is that $A(t)$ generates a C_0-semigroup for each $t \geq 0$. This assumption, which we do not make, is also of basic importance in more recent treatments of evolution equations (see [26], [20, Chap. 4], [8, Chap. 7]). The fundamental questions concerning existence and uniqueness of solutions, construction of evolution operators, and the well-posedness of the Cauchy problem are still subjects of current research. In broad correspondence to the approaches of Kato and Tanabe, two sets of assumptions have evolved: one for hyperbolic-type equations and the other for parabolic-type equations (see [26], [20], [8]). A different approach has been developed by Lions [19], who assumed that $A(t)$ is defined via a time-varying bilinear form. In spite of these efforts the existence theory for solutions of time-varying equations (1) is by no means as well developed as that of time-invariant differential equations, where necessary and sufficient conditions for unique solvability are given by the Hille–Yosida theorem.

The first attempts to develop a *stability theory* for time-varying linear differential equations in Banach spaces go back to the late forties. In [16] Krein extended results of Liapunov on second-order systems with periodic coefficients to a Hilbert space setting. Over the next two decades the work on the stability of time-varying infinite-dimensional systems was mainly restricted to equations (1) with *bounded* $A(t) \in \mathcal{L}(X)$, $t \geq 0$ [4]. Early contributions to the stability theory of systems with *unbounded* operators $A(t)$ can be found in [5] and [26].

Our main objective is to investigate the *robustness* of exponentially stable systems (1) when the unbounded generator $A(\cdot)$ is subjected to various types of unbounded perturbations. A prerequisite for this is to secure the existence and uniqueness of solutions of the perturbed equation. However, we emphasize that in our analysis the question of existence and uniqueness is treated jointly with the problem of exponential stability. A separate treatment is possible (along the lines indicated in Remark 3.4, below) but is beyond the scope of the present paper. Clearly, such an analysis would be of independent interest and desirable from a systematic point of view.

Problems of robust stability rely for their solution on methods from both stability theory and perturbation theory. For the time-invariant case, there is a well-developed perturbation theory [9], [16], [20], although the effects of *large perturbations* on the system behaviour have not been so frequently studied in the literature [23], [24]. In comparison, the situation for time-varying systems is bleak. Phillips [21] has given an example where A generates a C_0-semigroup on X, $\Delta(\cdot) \in C(\mathbb{R}_+, \mathcal{L}(X))$, and for certain initial states $x \in D(A)$ there are no differentiable functions $x(\cdot)$ such that $\dot{x}(t) = (A + \Delta(t))x(t)$, $t \geq 0$, $x(0) = x$. Because we want to consider quite general perturbations, we cannot, therefore, insist that the perturbed system has differentiable solutions for every $x \in D(A)$.

We now proceed to specify the perturbation structures that we consider in this paper. Suppose that $A(t)$ is subjected to perturbations of the *output feedback type*, that is, there are Banach spaces $\underline{X}, \overline{X}, U, Y$ with

$$\underline{X} \subset X \subset \overline{X}, \tag{2}$$

such that the perturbed system equations are of the form

$$\dot{x}(t) = [A(t) + D(t)\Delta(t)E(t)]x(t), \qquad t \geq 0, \tag{3}$$

where $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y, U))$ is an unknown *bounded* time-varying disturbance operator and $D(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(U, \overline{X}))$, $E(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(\underline{X}, Y))$ are given operator-valued functions that describe the structure and unboundedness of the perturbation.

We assume that the inclusions in (2) are continuous with dense ranges so that $E(t) \in \mathcal{L}(\underline{X}, Y)$ may be viewed as an unbounded operator from X to Y and $D(t) \in \mathcal{L}(U, \overline{X})$ as an unbounded operator from U to X. Moreover, because $E(\cdot)$, $D(\cdot)$ are only assumed to be piecewise continuous, the perturbations may also be unbounded *in time*. The size of the disturbance operator $\Delta(\cdot)$ is measured by

$$\|\Delta(\cdot)\|_\infty = \operatorname*{ess\,sup}_{t\geq 0} \|\Delta(t)\|_{\mathcal{L}(Y,U)}.$$

The idea of a *stability radius* was introduced in [11] for time-invariant finite-dimensional systems. It is the size of the smallest perturbation Δ, which results in a time-invariant system (3) that is not exponentially stable. A fairly complete theory of the stability radius has been developed for both finite- and infinite-dimensional time-invariant systems over the complex field $\mathbb{C}$ [11], [24]. For time-varying systems the results are less satisfactory and there is no computable formula available for the stability radius. In the finite-dimensional case lower bounds have been obtained in terms of an associated input–output operator, and these have been improved by the use of scaling techniques [10], [13]. Here we extend this approach to infinite-dimensional time-varying systems.

The organization of the paper is as follows. To deal with uncertain parameters in the model we need a framework that allows for the same degree of unboundedness in the perturbations $D(t)\Delta(t)E(t)$ as in the generator $A(t)$. This is developed in §2 where we recall some notions from infinite-dimensional systems theory [3] and discuss various relationships between a generator $A(\cdot)$ and an associated evolution operator $\Phi(\cdot,\cdot)$. We have already seen that it would be overrestrictive to assume that the perturbed system (3) defines a well-posed Cauchy problem. Therefore, we renounce this requirement and introduce a mild version of the perturbed system equation. This leads us to the basic concept of a *mild evolution operator*. A disturbance Δ is called *admissible* if the mild version of the perturbed system equation defines a mild evolution operator. Based on this notion we introduce stability radii for a wide class of perturbed systems.

In the two subsequent sections we derive the main results of this paper. In §3 we specify conditions that enable us to determine a lower bound for the stability radius of an exponentially stable mild evolution operator $\Phi(\cdot,\cdot)$ under perturbations of a given structure (D, E). This lower bound is $\|\mathbb{L}_0\|^{-1}$, where $\mathbb{L}_0$ is the input–output operator of the system

$$\begin{aligned} x(t) &= \Phi(t,0)x(0) + \int_0^t \Phi(t,s)D(s)u(s)ds, \\ y(t) &= E(t)x(t), \qquad t \geq 0. \end{aligned}$$

The central problem here is to construct a perturbed mild evolution operator for every disturbance Δ satisfying $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0\|^{-1}$. The conditions imposed in §3 are rather complicated, so we specialize them to time-invariant systems and compare them with ones that have been given in the literature. In particular, we see that if (A, D, E) is a regular system in the sense of Weiss [29], then our conditions hold. Moreover, under a slight additional assumption we are able to prove that the lower bound $\|\mathbb{L}_0\|^{-1}$ is in fact equal to the stability radius in a time-invariant setting *when the ground field* $\mathbb{K}$ *is complex*. This theorem considerably improves the result in [23]. We also show that a time-invariant system cannot be destabilized by smaller time-varying perturbations than by constant disturbances.

Because of its generality, a mild evolution operator is not necessarily representable as the fundamental solution of a time-varying differential equation, nor is it necessarily associated with a generator. Therefore, we introduce (in §2) the more restrictive notion of a *weak evolution operator*, which has a generator and is associated with a differential equation (although this

equation is not necessarily well posed). This class of models lies between the class of strong evolution operators defined by well-posed Cauchy problems and the general class of mild evolution operators. In §4 we investigate under which conditions the perturbed mild evolution operator that exists and is unique for Δ satisfying $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0\|^{-1}$ is in fact a weak evolution operator. It turns out that the unboundedness of the perturbations has to be restricted to roughly "half of the unboundedness" of the weak generator. The precise result is given in Theorem 4.2. Specializing to time-invariant system, we show that if the system (A, D, E) is in the so-called Pritchard–Salamon class [22], then again our conditions hold. We also examine in some detail the case where the nominal model is an analytic semigroup.

In §5 the notion of a Bohl transformation is extended to an infinite-dimensional context and the behaviour of the stability radius under these transformations is investigated. We show that scalar Bohl transformations of the state may be used to improve the tightness of the lower bound. We also consider *multiperturbations* of the form

$$\dot{x} = A(t)x(t) + \sum_{i=1}^{N} D_i(t)\Delta_i(t)E_i(t)x(t), \qquad t \geq 0 \tag{4}$$

and introduce a stability radius for this wider class of structured perturbations. Equation (4) may be written in the form (3) where $\Delta(\cdot)$ has a diagonal structure. We are, therefore, able to apply the results of §§3 and 4 to obtain a lower bound for the stability radius of multiperturbation problems. Moreover, we show that scalar Bohl transformations of the state and arbitrary scalings of the inputs and outputs may be used to improve this lower bound.

In §6 the applicability of the abstract results and conditions are illustrated by three examples: an uncertain time-varying system (1) with $A(t) \in \mathcal{L}(X)$ unbounded in time, a perturbed distributed parameter system described by a three-dimensional heat equation with uncertain conductivity (space dependent), and an interconnected system with perturbed subsystems and uncertain couplings.

2. Definitions. We begin by introducing a number of concepts that enable us to define the meaning of the nominal equation (1) and the perturbed system equation (3). The concept of a weak evolution operator (Definition 2.4) is fundamental in §4, while the weaker concept of a mild evolution operator is basic in §3.

In the literature $A(\cdot)$ is usually supposed to be strongly continuous [17], [8], [26]. However, to allow for perturbations with jumps, we assume that $A(\cdot)$ is piecewise continuous. More precisely, the following is a standing assumption for the nominal model (1).

Assumption. For all $t \geq 0$, $A(t)$ is a linear operator on X, its domain $D(A)$ is dense in X and is independent of t. There exists a discrete subset $J \subset \mathbb{R}_+$ (set of jump points) such that, for all $x \in D(A)$, $A(\cdot)x$ is continuous on $\mathbb{R}_+\backslash J$ and the one-sided limits $\lim_{t\downarrow\tau} A(t)x$, $\lim_{t\uparrow\tau} A(t)x$ exist at each jump point $\tau \in J$.

Consider the Cauchy problem

$$\begin{aligned} \dot{x}(t) &= A(t)x(t), \qquad t \geq s, \\ x(s) &= x, \end{aligned} \tag{5}$$

where the initial state $x \in D(A)$ and the initial time $s \geq 0$ are given.

DEFINITION 2.1. *A function $x(\cdot) \in C(s, \infty; X)$ is said to be a* strong solution *of* (5) *on $[s, \infty)$ if $x(t) \in D(A)$ for all $t \geq s$, $x(\cdot)$ is strongly differentiable on $[s, \infty)\backslash J$ with right-sided derivative in s and satisfies* (5) *in the following sense*:

(i) (5) *holds for every $t \in (s, \infty)\backslash J$;*

(ii) $$\left.\begin{aligned} \frac{\partial_+ x}{\partial t}(\tau) &:= \lim_{h\downarrow 0} \frac{x(\tau+h)-x(\tau)}{h} = \lim_{t\downarrow\tau} A(t)x(t), \\ \frac{\partial_- x}{\partial t}(\tau) &:= \lim_{h\downarrow 0} \frac{x(\tau)-x(\tau-h)}{h} = \lim_{t\uparrow\tau} A(t)x(t), \end{aligned}\right\} \tau \in J\backslash\{s\};$$

(iii) $x(s) = x$ *and* $\frac{\partial_+ x}{\partial t}(s) = \lim_{t\downarrow s} A(t)x(t)$.

Note that a strong solution can only exist if $x \in D(A)$. It is not clear at this point whether such a solution of (5) exists and whether it is uniquely determined. Let $\Gamma = \{(t,s);\ 0 \le s \le t < \infty\}$.

DEFINITION 2.2. *The Cauchy problem* (5) *is said to be* well posed *on* Γ *if the following hold*:

(i) *For every* $s \ge 0$, $x \in D(A)$, (5) *has a unique strong solution* $x(\cdot, s)$ *on* $[s, \infty)$;

(ii) $x(t,s)$ *is continuous and* $(\partial x/\partial t)(t,s)$ *is piecewise continuous in* $t \in [s,\infty)$ *and in* $s \in [0,t]$ *with jump points only in* J;

(iii) *The solution* $x(t,s)$ *depends continuously on the initial value* x *locally uniformly in* s *and in* t, *that is, if* $x_n \in D(A)$ *and* $\lim_{n\to\infty} x_n = 0$, *then the corresponding solutions* $x_n(t,s)$ *converge to zero locally uniformly in* t *for fixed* s *and locally uniformly in* s *for fixed* t *as* $n \to \infty$.

This definition is slightly weaker than the definition of a well-posed Cauchy problem in the literature [17, II§2], [8].

PROPOSITION 2.3. *Suppose the Assumption holds and that the Cauchy problem* (5) *is well posed. Then there is a (unique) family* $\Phi = (\Phi(t,s))_{(t,s)\in\Gamma}$ *of bounded linear operators* $\Phi(t,s) \in \mathcal{L}(X)$ *such that the solutions of* (5) *are given by* $x(t,s) = \Phi(t,s)x$, *for all* $x \in D(A)$, $s \ge 0$. *Moreover,* Φ *has the following properties*:

(i) $\Phi(t,t) = I$, $t \in \mathbb{R}_+$;

(ii) $\Phi(t,\sigma)\Phi(\sigma,s) = \Phi(t,s)$, $0 \le s \le \sigma \le t < \infty$;

(iii) $D(A)$ *is* $\Phi(t,s)$*-invariant for all* $(t,s) \in \Gamma$, $t \mapsto A(t)\Phi(t,s)x$ *is piecewise continuous on* $[s,\infty)$ *with jump points only in* J *for all* $x \in D(A)$, *and*

$$\Phi(t,s)x - x = \int_s^t A(\rho)\Phi(\rho,s)x\,d\rho, \qquad (t,s) \in \Gamma; \tag{6}$$

(iv) *For all* $x \in D(A)$, $s \mapsto \Phi(t,s)A(s)x$ *is piecewise continuous on* $[0,t]$ *with jump points only in* J, *and*

$$\Phi(t,s)x - x = \int_s^t \Phi(t,\rho)A(\rho)x\,d\rho, \qquad (t,s) \in \Gamma; \tag{7}$$

(v) $\Phi(\cdot,s)$ *is strongly continuous on* $[s,\infty)$ *and* $\Phi(t,\cdot)$ *is strongly continuous on* $[0,t]$.

Proof. The existence and uniqueness of the family $(\Phi(t,s))_{(t,s)\in\Gamma}$ satisfying $x(t,s) = \Phi(t,s)x$ and (i), (ii) is easily established [17]. If (x_k) is a sequence in $D(A)$ converging to $x \in X$ in X, then $\Phi(\cdot,s)x_k$ converges to $\Phi(\cdot,s)x$ uniformly on every compact subset in $[s,\infty)$ by condition (iii) of Definition 2.2. Hence $\Phi(\cdot,s)x$ is continuous on $[s,\infty)$. Similarly, it can be shown that $\Phi(t,\cdot)$ is strongly continuous on $[0,t]$, whence (v).

As a consequence of (v), the sets $\{\Phi(t,s)x;\ s \in [0,t]\}$ and $\{\Phi(t,s)x;\ t \in [s,\tau]\}$ are bounded for arbitrary $t \ge 0$, respectively, $\tau \ge s$ and $x \in X$. By the theorem of Banach–Steinhaus it follows that the sets $\{\Phi(t,s);\ s \in [0,t]\}$ and $\{\Phi(t,s);\ t \in [s,\tau]\}$ are uniformly bounded in $\mathcal{L}(X)$. Using the same arguments as in [17, II§3] it can be shown that $\Phi(\cdot,s)x$ and $\Phi(t,\cdot)x$ are strongly differentiable on $[s,\infty)\backslash J$, $[0,t]\backslash J$, respectively, and

$$\frac{\partial\Phi(t,s)x}{\partial t} = A(t)\Phi(t,s)x, \qquad t \in [s,\infty)\backslash J,\ x \in D(A), \tag{8}$$

$$\frac{\partial \Phi(t,s)x}{\partial s} = -\Phi(t,s)A(s)x, \qquad s \in [0,t]\backslash J,\ x \in D(A). \tag{9}$$

The partial derivatives in (8), (9) are taken in the unilateral sense at $t = s$ and $s = 0$, $s = t$, respectively, and at all jump points $\tau \in J$ of $A(\cdot)$. Then the equations in (8), (9) are also satisfied at the jump points if their right-hand sides (RHS) are replaced by the corresponding unilateral limits. For $x \in D(A)$ the map $t \mapsto A(t)\Phi(t,s)x$ is piecewise continuous on $[s,\infty)$ and continuous on $[s,\infty)\backslash J$ by condition (ii) in Definition 2.2 and $s \mapsto \Phi(t,s)A(s)x$ is piecewise continuous on $[0,t]$ and continuous on $[0,t]\backslash J$ by the Assumption and the uniform boundedness of $\Phi(t,\cdot)$ on $[0,t]$. This proves (iii) and (iv). □

The operator family Φ in the previous proposition is called the *evolution operator* associated with the well-posed Cauchy problem (5).

DEFINITION 2.4. *Let $A = (A(t))_{t\geq 0}$ be a family of linear operators on X satisfying the Assumption and $\Phi = (\Phi(t,s))_{(t,s)\in\Gamma}$ be a family of bounded linear operators $\Phi(t,s) \in \mathcal{L}(X)$ such that conditions* (i)–(v) *of Proposition* 2.3 *are satisfied. Then we say that Φ is a* strong evolution operator *with generator $A(\cdot)$. If only conditions* (i), (ii), (iv), (v) *are satisfied, Φ is called a* weak evolution operator *with (weak) generator $A(\cdot)$.*

Note that the two equations (8), (9) hold if Φ is a strong evolution operator, whereas only (9) holds if Φ is a weak evolution operator. In both cases the generator $A(\cdot)$ is uniquely determined outside of J by the evolution operator via

$$A(t)x = -\frac{\partial_- \Phi(t,s)x}{\partial s}\bigg|_{s=t} = -\lim_{h\downarrow 0}\frac{x - \Phi(t,t-h)x}{h}, \qquad t \in \mathbb{R}_+\backslash J,\ x \in D(A).$$

However, observe that there is some arbitrariness in the definition of the domain $D(A)$. In fact, Definition 2.4 does not give a definition of a strong (weak) evolution operator that is separated from and independent of its generator and vice versa. For a systematic study of the relationship between these two objects (e.g., in the spirit of the Hille–Yosida theorem for semigroups, see [20]) this would be unsatisfactory. But such a study is not intended in this paper and for our purpose a joint definition of the two concepts is sufficient.

To see that any strong evolution operator is uniquely determined by its generator we need the following lemma.

LEMMA 2.5. *Suppose $\Phi = (\Phi(\rho))_{\rho\in[s,t]}$ is a family of operators in $\mathcal{L}(X)$ such that $\rho \mapsto \Phi(\rho)z$ is continuous on $[s,t]$ for all $z \in X$ and differentiable on (s,t) for all $z \in D \subset X$. Moreover, assume that $x(\cdot) : (s,t) \to X$ is differentiable and has values in D. Then $y(\rho) = \Phi(\rho)x(\rho)$ is differentiable on (s,t) and the following "product rule" holds:*

$$y'(\rho) = \Phi'(\rho)x(\rho) + \Phi(\rho)x'(\rho),$$

where $\Phi'(\rho)x(\rho)$ denotes the derivative of $\tau \mapsto \Phi(\tau)x(\rho)$ at $\tau = \rho$.

Proof. For any $\rho \in (s,t)$ and $h \neq 0$ such that $\rho + h \in (s,t)$, we have

$$\begin{aligned}\frac{1}{h}[y(\rho+h) - y(\rho)] = {} & \frac{1}{h}[\Phi(\rho+h)x(\rho+h) - \Phi(\rho+h)x(\rho)] \\ & + \frac{1}{h}[\Phi(\rho+h)x(\rho) - \Phi(\rho)x(\rho)].\end{aligned} \tag{10}$$

As $h \to 0$ the second term converges toward $\Phi'(\rho)x(\rho)$ by assumption. Setting $x_h(\rho) = \frac{1}{h}[x(\rho+h) - x(\rho)]$ the first term in (10) is

$$\Phi(\rho+h)[x_h(\rho) - x'(\rho)] + \Phi(\rho+h)x'(\rho)$$

and the first term tends toward zero as $h \to 0$ because of the uniform boundedness of $(\Phi(\rho))_{\rho\in[s,t]}$ in $\mathcal{L}(X)$ (by the theorem of Banach–Steinhaus), whilst the second term tends toward $\Phi(\rho)x'(\rho)$. This concludes the proof. □

If Φ is a weak evolution operator on X, it is not clear whether there exists a strong solution of the Cauchy problem (5) for every $x \in D(A)$. However, if one exists, then it is unique and it is given by the evolution operator.

PROPOSITION 2.6. *Suppose $(\Phi(t,s))_{(t,s)\in\Gamma}$ is a weak evolution operator on X with generator $A(\cdot)$ and $x(\cdot)$ solves (5) for given $s \geq 0$, $x \in D(A)$. Then $x(t) = \Phi(t,s)x$.*

Proof. For $[\bar{s},\bar{t}] \subset \mathbb{R}_+\backslash J$ by the previous lemma $y(\rho) = \Phi(\bar{t},\rho)x(\rho)$ is differentiable on $(\bar{s},\bar{t})$ with derivative

$$y'(\rho) = -\Phi(\bar{t},\rho)A(\rho)x(\rho) + \Phi(\bar{t},\rho)A(\rho)x(\rho) = 0.$$

Thus,

$$x(\bar{t}) - \Phi(\bar{t},\bar{s})x(\bar{s}) = \int_{\bar{s}}^{\bar{t}} \frac{\partial}{\partial\rho}[\Phi(\bar{t},\rho)x(\rho)]d\rho = 0.$$

Now let $J \cap [s,\infty) = \{t_k\}$, $t_1 < t_2 < \cdots$ and set $t_0 = s$. By continuity we obtain from the above that $x(\bar{t}) = \Phi(\bar{t},\bar{s})x(\bar{s})$ for any $\bar{s} \leq \bar{t}, \bar{s}, \bar{t} \in [t_k, t_{k+1}]$, $k \in \mathbb{N}$. Hence by induction, $x(t) = \Phi(t,s)x$ follows, using the properties (i) and (ii) of an evolution operator. □

PROPOSITION 2.7. *If Φ is a strong evolution operator on X with generator A, then the Cauchy problem (5) is well posed.*

Proof. By Definition 2.4 and (8), $x(t) = \Phi(t,s)x$ is a strong solution of (5) on $[s,\infty)$ for every $s \geq 0$, $x \in D(A)$, and this solution is uniquely determined by the previous proposition. This shows condition (i) in Definition 2.2. Condition (ii) follows from the Assumption and (iii), (iv), (v) in Proposition 2.3. Finally, by condition (v) in Proposition 2.3 and the theorem of Banach–Steinhaus, the sets $\{\Phi(t,s);\ 0 \leq s \leq t\}$ and $\{\Phi(t,s);\ s \leq t \leq \tau\}$ are bounded in $\mathcal{L}(X)$ for every $t \geq 0$, respectively $\tau \geq s$. This implies (iii) in Definition 2.2. □

COROLLARY 2.8. *If $(\Phi(t,s))_{(t,s)\in\Gamma}$ and $(\hat{\Phi}(t,s))_{(t,s)\in\Gamma}$ are two strong evolution operators with the same generator $A(\cdot)$, then they are equal.*

Various sufficient conditions under which a family $(A(t))_{t\in\mathbb{R}_+}$ of linear operators on X generates a strong evolution operator can be obtained from the results in [27].

We now turn to the perturbed system equation (3), that is, we consider the Cauchy problem

$$\begin{aligned} \dot{x}(t) &= (A(t) + D(t)\Delta(t)E(t))x(t), \qquad t \geq s, \\ x(s) &= x. \end{aligned} \tag{11}$$

In view of the above results it seems natural to assume that the nominal Cauchy problem (5) is well posed and to look for conditions on the perturbations Δ that guarantee that the perturbed Cauchy problem (11) is again well posed with an associated strong evolution operator Φ_Δ. However, even in the case where the nominal system is time-invariant and the time-varying perturbations $D(t)\Delta(t)E(t)$ are bounded ($\underline{X} = X = \overline{X}$), one cannot guarantee that (11) has strong solutions. More precisely, if a time-invariant closed operator A generates a strongly continuous semigroup $(S(t))_{t\geq 0}$ on X (hence a strong evolution operator $(S(t-s))_{(t,s)\in\Gamma}$), $D(t) = E(t) = I$, $t \geq 0$, and $\Delta(\cdot) \in C(\mathbb{R}_+;\mathcal{L}(X))$, then it is possible that (11) has no differentiable solutions for certain initial conditions $x(s) = x \in D(A)$ [21]. However, we can show [3] that $A + \Delta(\cdot)$ always generates a weak evolution operator with domain $D(A + \Delta) = D(A)$.

Our main aim in this paper is to define and examine a stability radius for (1) when subjected to a wide range of unbounded perturbations. If the perturbed Cauchy problem is required to

be well posed, we have just seen that the perturbation class needs to be restricted. Instead of following this path we consider a *mild* version of the perturbed Cauchy problem (11) that allows us to admit perturbations $D(t)\Delta(t)E(t)$ of the same degree of unboundedness as that of $A(t)$. In §4 we analyze the degree of unboundedness that can be admitted if we require the perturbed equation (11) to define a *weak* evolution operator.

To motivate the mild version of the perturbed Cauchy problem, suppose that (11) has a strong solution $x(\cdot)$ with values in $D(A)$, for some given $s \geq 0$, $x \in D(A)$. Applying Lemma 2.5, we obtain

$$\begin{aligned} x(t) - \Phi(t,s)x &= \int_s^t \frac{\partial}{\partial \rho}[\Phi(t,\rho)x(\rho)]d\rho \\ &= \int_s^t \Phi(t,\rho)[-A(\rho) + A(\rho) + D(\rho)\Delta(\rho)E(\rho)]x(\rho)d\rho \\ &= \int_s^t \Phi(t,\rho)D(\rho)\Delta(\rho)E(\rho)x(\rho)d\rho, \qquad (t,s) \in \Gamma. \end{aligned}$$

Or

$$x(t) = \Phi(t,s)x + \int_s^t \Phi(t,\rho)D(\rho)\Delta(\rho)E(\rho)x(\rho)d\rho, \qquad t \geq s. \tag{12}$$

This is the "mild" version of the perturbed Cauchy problem (11), and to make sense of it we specify some standing hypotheses. Because the generator $A(\cdot)$ does not appear in (12), this equation can be considered the perturbed system equation for a wider class of dynamical models described by *mild evolution operators*.

DEFINITION 2.9 (Mild evolution operator). $\Phi(\cdot,\cdot) : \Gamma \to \mathcal{L}(X)$ *is a* mild evolution operator *on* X *if the following hold*:

(i) $\Phi(t,t) = I$, $t \in \mathbb{R}_+$;

(ii) $\Phi(t,\sigma)\Phi(\sigma,s) = \Phi(t,s)$, $0 \leq s \leq \sigma \leq t < \infty$;

(iii) $\Phi(\cdot,s)$ *is strongly continuous on* $[s,\infty)$ *and* $\Phi(t,\cdot)$ *is strongly continuous on* $[0,t]$.

If there exist $M > 0$, $\omega > 0$ *such that*

$$\|\Phi(t,s)\| \leq Me^{-\omega(t-s)}, \qquad t \geq s \geq 0, \tag{13}$$

then the mild evolution operator $\Phi(\cdot,\cdot)$ *is said to be* exponentially stable.

Remark 2.10. By definition, every weak evolution operator is a mild one. A mild evolution operator Φ is said to be *time-invariant* if $\Phi(t,s) = \Phi(t-s,0)$, $(t,s) \in \Gamma$. But then $S(t) := \Phi(t,0)$ defines a strongly continuous semigroup on X, and as a consequence we see that the concepts of mild, weak, and strong evolution operators coincide in the time-invariant setting [3].

If Φ is a mild evolution operator, then the sets $\{\Phi(t,s);\ s \in [0,t]\}$ and $\{\Phi(t,s);\ t \in [s,\tau]\}$ are bounded in $\mathcal{L}(X)$ for every $t \geq 0$, respectively $\tau \geq s$. However, in contrast to the semigroup case, there may not exist exponential bounds for $\Phi(\cdot,s)$ on $[s,\infty)$. To capture this property, Bohl [1] introduced the concept of a Bohl exponent, which we now extend to an infinite-dimensional setting [4].

DEFINITION 2.11 (Bohl exponent). *The* (upper) Bohl exponent $\beta(\Phi)$ *of a mild evolution operator* $\Phi(t,s)$ *is given by*

$$\beta(\Phi) = \inf\{\omega \in \mathbb{R};\ \exists M_\omega > 0 : t \geq s \geq 0 \Rightarrow \|\Phi(t,s)\| \leq M_\omega e^{\omega(t-s)}\}.$$

It is possible that $\beta(\Phi) = \pm\infty$. If $\Phi(t,s) = S(t-s)$, where $(S(t))_{t\in\mathbb{R}_+}$ is an analytic semigroup on X with generator A, then

$$\beta(\Phi) = \sup_{\lambda\in\sigma(A)} \operatorname{Re}\lambda.$$

The following proposition collects some basic properties of Bohl exponents used throughout this paper.

PROPOSITION 2.12. *Suppose that Φ is a mild evolution operator on X. Then we have the following.*

(i) *The Bohl exponent of Φ is finite if and only if* $\sup_{0\le t-s\le 1}\|\Phi(t,s)\| < \infty$.

(ii) *If $\beta(\Phi) < \infty$, then*

$$\beta(\Phi) = \limsup_{s,t-s\to\infty} \frac{\ln\|\Phi(t,s)\|}{t-s}.$$

(iii) *If $\beta(\Phi) < \infty$, $1 \le p < \infty$ the following statements are equivalent:*

(a) *Φ is exponentially stable;*

(b) *$\beta(\Phi) < 0$;*

(c) *there exists a constant c such that*

$$\int_s^\infty \|\Phi(t,s)x\|^p\,dt \le c^p\|x\|^p \quad \text{for all } s \ge 0,\ x \in X.$$

Proof. (i), (ii), and the equivalence "(a) $\Leftrightarrow$ (b)" in (iii) can be proved in a similar manner to that in [4]. The implication "(a) $\Rightarrow$ (c)" being trivial, it only remains to show that (c) implies (a). This is proved in [4] for the case where Φ is generated by a bounded A and in [5] for the general case with $p = 2$. A slick proof for the case $p = 2$ is given in [3] and it is easy to see that this proof can be extended to any p, $1 \le p < \infty$. □

The Bohl exponent of (1) is said to be *strict* if "lim sup" can be replaced by "lim" in (ii).

Remark 2.13. Suppose that Φ is a mild evolution operator on X and $D(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(U,X))$, $E(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(X,Y))$. Then the triplet (Φ, D, E) defines a time-varying linear dynamical system with input space U, state space X, and output space Y. The state trajectory generated by an initial condition $x(s) = x \in X$ and a control function $u(\cdot) \in L^1(s,\tau;U), \tau > s$ is

$$\varphi(t,s,x,u(\cdot)) = \Phi(t,s)x + \int_s^t \Phi(t,\rho)D(\rho)u(\rho)d\rho, \qquad t \in [s,\tau], \tag{14}$$

with associated output function

$$\begin{aligned} y(t,s,x,u(\cdot)) &= E(t)\varphi(t,s,x,u(\cdot)) \\ &= E(t)\Phi(t,s)x + E(t)\int_s^t \Phi(t,\rho)D(\rho)u(\rho)d\rho, \qquad t \in [s,\tau]. \end{aligned} \tag{15}$$

It can be shown that if the Bohl exponent $\beta(\Phi) < \infty$, then $x(\cdot) = \varphi(\cdot,s,x,u(\cdot))$ is continuous on $[s,\tau]$. Indeed, if $s \le \hat{t} \le t \le \tau$, then

$$\begin{aligned} x(t) - x(\hat{t}) = [\Phi(t,s) - \Phi(\hat{t},s)]x &+ \int_s^{\hat{t}} [\Phi(t,\hat{t}) - I_X]\Phi(\hat{t},\rho)D(\rho)u(\rho)d\rho \\ &+ \int_{\hat{t}}^t \Phi(t,\rho)D(\rho)u(\rho)d\rho. \end{aligned} \tag{16}$$

If we fix $\hat{t}$, the first and the second term go to zero as $t \downarrow \hat{t}$ by the strong continuity of $\Phi(\cdot, s)$. Now because $\beta(\Phi) < \infty$, there exists M_τ such that $\|\Phi(t,\rho)\| \leq M_\tau, s \leq \rho \leq t \leq \tau$. Hence the last term goes to zero by the estimate

$$\left\| \int_{\hat{t}}^{t} \Phi(t,\rho)D(\rho)u(\rho)d\rho \right\| \leq M_\tau \int_{\hat{t}}^{t} \|D(\rho)u(\rho)\|d\rho \to 0 \quad \text{as } t \downarrow \hat{t}.$$

So $x(\cdot)$ is continuous from the right. Now fix t and let $\hat{t} \uparrow t$. The first and the last term in (16) go to zero by the same arguments as before. Extending the integrand of the second term by zero to the fixed interval $[s,t]$, it can be written in the form $\int_s^t f(\hat{t},\rho)d\rho$, where $f(\hat{t},\rho)$ tends pointwise (in ρ) to zero as $\hat{t} \uparrow t$ (by strong continuity of $\Phi(t,\cdot)$) and its norm is bounded by the integrable function $2M_\tau\|D(\rho)u(\rho)\|_X$. Therefore, $\int_s^t f(\hat{t},\rho)d\rho \to 0$ as $\hat{t} \uparrow t$. Hence $x(\cdot)$ is continuous from the left.

Because in our context the operators $D(t)$ and $E(t)$ are unbounded, the integral in (14) is not well defined without further assumptions. We introduce the following hypotheses (which will be supplemented later). Throughout this paper we suppose that p is given a real number with $1 \leq p \leq \infty$.

Hypothesis 1. $\underline{X}$, X, $\overline{X}$ are Banach spaces such that $\underline{X} \subset X \subset \overline{X}$ and the canonical injections $\underline{X} \hookrightarrow X$, $X \hookrightarrow \overline{X}$ are continuous with dense ranges. $D(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(U, \overline{X}))$, $E(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(\underline{X}, Y))$.

Hypothesis 2. $(\Phi(t,s))_{(t,s)\in\Gamma}$ is a mild evolution operator on X, and for all $t \geq s \geq 0$, $\Phi(t,s)$ can be extended to a bounded linear operator on $\overline{X}$ (again denoted by $\Phi(t,s)$).

Hypothesis 3. For every $u(\cdot) \in L^p(0,t;U)$, $t > 0$, the map $\Phi(t,\cdot)D(\cdot)u(\cdot)$ from $[0,t]$ to $\overline{X}$ is integrable in $\overline{X}$.

An important role will be played by the *input to state* operators $\mathbb{M}_s$, $s \geq 0$,

$$(\mathbb{M}_s u)(t) = \int_s^t \Phi(t,\rho)D(\rho)u(\rho)d\rho, \qquad t \geq s, \qquad u(\cdot) \in L^p(s,\infty;U). \tag{17}$$

Note that because of Hypothesis 3, the integral in (17) is well defined (in $\overline{X}$). The following hypothesis will be needed later and contains stronger assumptions than required for this section.

Hypothesis 4. For every $u(\cdot) \in L^p(s,\infty;U)$, $s \geq 0$, $(\mathbb{M}_s u)(t) \in \underline{X}$ for almost every $t \geq s$, $t \mapsto (\mathbb{M}_s u)(t)$ is continuous on $[s,\infty)$ with respect to the norm $\|\cdot\|_X$, and there exists an exponentially bounded $k(t) \geq 0$ such that

$$\|(\mathbb{M}_s u)(t)\|_X \leq k(t)\|u(\cdot)\|_{L^p(s,t;U)}, \qquad t \geq s \geq 0. \tag{18}$$

DEFINITION 2.14. *Given a perturbation $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$, a continuous function $x(\cdot) : [s,\infty) \to X$ is said to be a* mild solution *of* (12) *on $[s,\infty)$ if $x(t) \in \underline{X}$ for almost all $t \geq s$, $E(\cdot)x(\cdot)$ is L^p-integrable on every interval $[s,t]$ and* (12) *is satisfied for all $t \geq s$.*

Note that at this stage we cannot be sure that a mild solution $x(\cdot)$ of (12) exists and is unique, even if $x \in \underline{X}$.

Remark 2.15. Under Hypotheses 1 and 2, $(\Phi(t,s))_{(t,s)\in\Gamma}$ satisfies conditions (i), (ii) of Definition 2.9 on $\overline{X}$. Moreover, Φ is a mild evolution operator on $\overline{X}$ if and only if, in addition, $\{\Phi(t,s); s \in [0,t]\}$ and $\{\Phi(t,s); t \in [s,\tau]\}$ are bounded in $\mathcal{L}(\overline{X})$ for every $t \geq 0$, respectively, $s \geq 0$, $\tau \geq s$. The necessity is clear. Now assume that the above sets are bounded. For any $x \in \overline{X}$, let (x_k) be a sequence in X that converges to x in $\overline{X}$. Because

$$\|\Phi(t,s)x - \Phi(t,s)x_k\|_{\overline{X}} \leq \sup_{s\in[0,t]} \|\Phi(t,s)\|_{\mathcal{L}(\overline{X})}\|x - x_k\|_{\overline{X}}, \qquad s \in [0,t],$$

$\Phi(t,\cdot)x : [0,t] \to \overline{X}$ is a uniform limit of the continuous functions $\Phi(t,\cdot)x_k : [0,t] \to \overline{X}$; hence it is continuous. The continuity of $\Phi(\cdot,s)x$ on $[s,\infty)$ for $x \in \overline{X}$ is proved similarly.

Now suppose that there exists a mild evolution operator $\Phi_\Delta(\cdot,\cdot)$ on X such that for all $t \geq s \geq 0$,

$$\Phi_\Delta(\rho,s)x \in \underline{X}, \quad \text{for any } x \in \underline{X} \quad \text{and} \quad \text{a.e. } \rho \in [s,t], \tag{19}$$

$$\|E(\cdot)\Phi_\Delta(\cdot,s)x\|_{L^p(s,t;Y)} \leq k\|x\|_X, \qquad x \in \underline{X} \quad (k \geq 0 \text{ a constant}), \tag{20}$$

$$\Phi_\Delta(t,s)x = \Phi(t,s)x + \int_s^t \Phi(t,\rho)D(\rho)\Delta(\rho)E(\rho)\Phi_\Delta(\rho,s)x\,d\rho, \qquad x \in \underline{X}. \tag{21}$$

If $x \in X$, there exists a sequence (x_k) in $\underline{X}$ that converges to x in X. By (20) the corresponding sequence $(E(\cdot)\Phi_\Delta(\cdot,s)x_k|_{[s,t]})$ in $L^p(s,t;Y)$ is Cauchy for all $t > s$. We denote the limit in $L^p_{loc}(s,\infty;Y)$ by

$$E(\cdot)\Phi_\Delta(\cdot,s)x = \lim_{k\to\infty} E(\cdot)\Phi_\Delta(\cdot,s)x_k. \tag{22}$$

With this notation we conclude from Hypotheses 3 and 4 that $x(t) = \Phi_\Delta(\cdot,s)x$ is a mild solution of the perturbed equation (12) for all $x \in X$ (see Definition 2.14).

DEFINITION 2.16 (Perturbed mild evolution operator). *Suppose Hypotheses* 1–4 *hold and* $\Delta \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$. *If there exists a unique mild evolution operator* $\Phi_\Delta(\cdot,\cdot)$ *on* X *satisfying* (19)–(21) *for all* $t \geq s \geq 0$, *then* $\Delta(\cdot)$ *is called an* admissible *perturbation for* $\Phi(\cdot,\cdot)$ *under the perturbation structure* (D,E) *and* $\Phi_\Delta(\cdot,\cdot)$ *is called the perturbed mild evolution operator corresponding to the admissible perturbation* $\Delta(\cdot)$.

We are now in a position to introduce the stability radius as a measure of robust stability for exponentially stable mild evolution operators.

DEFINITION 2.17. *Suppose* Φ *is exponentially stable and Hypothesis* 1–4 *hold. The* stability radius *of* $\Phi(\cdot,\cdot)$ *with respect to the perturbation structure* (D,E) *is defined by*

$$r(\Phi;D,E) = \sup\{r \in \mathbb{R}_+;\ \forall \Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U)) : \|\Delta\|_\infty \leq r \Rightarrow \Delta \text{ is admissible and } \Phi_\Delta(\cdot,\cdot) \text{ is exponentially stable}\}.$$

3. A lower bound for the stability radius. Throughout this section we assume that $\Phi(\cdot,\cdot)$ is an exponentially stable mild evolution operator on X and $1 \leq p < \infty$, $\mathbb{M}_s$ defined by (17). In addition to Hypotheses 1–4 we also require the following.

Hypothesis 5. For all $s \geq 0$, $\mathbb{M}_s \in \mathcal{L}(L^p(s,\infty;U),\ L^p(s,\infty;X))$.

Hypothesis 6. $\Phi(t,s)x \in \underline{X}$, for any $x \in \underline{X}$ and almost every $t \geq s$, $s \geq 0$.

Hypothesis 7. For every $x \in \underline{X}$, $t > 0$,

$$\lim_{s\to t} \|E(\cdot)\Phi(\cdot,s)x\|_{L^p(s,t;Y)} = 0, \tag{23}$$

and there exists a constant $K > 0$ such that $E(\cdot)\Phi(\cdot,s)x : [s,\infty) \to Y$ is p-integrable and

$$\|E(\cdot)\Phi(\cdot,s)x\|_{L^p(s,\infty;Y)} \leq K\|x\|_X, \qquad x \in \underline{X}, \quad s \geq 0. \tag{24}$$

Then just as in (22), we define $y_0(\cdot,s,x) \in L^p(s,\infty;Y)$, for $x \in X$, by

$$y_0(\cdot,s,x) = E(\cdot)\Phi(\cdot,s)x = \lim_{k\to\infty} E(\cdot)\Phi(\cdot,s)x_k, \tag{25}$$

where (x_k) is some sequence in $\underline{X}$ converging to x in X. Using this definition it is easy to see that (23) holds for every $x \in X$, $t > 0$. In fact, for any $\varepsilon > 0$ there exists, by (24), $l \in \mathbb{N}$ such that $\|E(\cdot)\Phi(\cdot,s)x - E(\cdot)\Phi(\cdot,s)x_l\|_{L^p(s,t;Y)} < \varepsilon/2$ for all $s \in [0,t]$ and there exists $\delta \in (0,t)$ such that $\|E(\cdot)\Phi(\cdot,s)x_l\|_{L^p(s,t;Y)} < \varepsilon/2$ for $s \in [t-\delta, t]$, whence $\|E(\cdot)\Phi(\cdot,s)x\|_{L^p(s,t;Y)} < \varepsilon$ for $s \in [t-\delta, t]$.

Our final hypothesis concerns the input–output operator $\mathbb{L}_s$, where

$$(\mathbb{L}_s u)(t) = E(t)\int_s^t \Phi(t,\rho)D(\rho)u(\rho)d\rho, \qquad t \geq s \geq 0,\ u(\cdot) \in L^p(s,\infty;U). \tag{26}$$

Hypothesis 8. $\mathbb{L}_s \in \mathcal{L}(L^p(s,\infty;U),\ L^p(s,\infty;Y))$, $s \geq 0$.

Remark 3.1. (i) If $E(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(X,Y))$, $\underline{X} = X$, then Hypothesis 6 and (23) are automatically satisfied and (24) is a consequence of the exponential stability of $\Phi(\cdot,\cdot)$.

(ii) If $D(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(U,X))$, $\overline{X} = X$, then Hypotheses 2, 3, and 5, and all but the first statement in Hypothesis 4 are satisfied.

(iii) If the conditions in both (i) and (ii) are satisfied, then Hypotheses 1–8 hold.

We have the following result.

THEOREM 3.2. *Suppose Hypotheses* 1–8 *hold and* Φ *is exponentially stable on* X. *Then*

$$r(\Phi; D, E) \geq \|\mathbb{L}_0\|^{-1}. \tag{27}$$

Proof. Let $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$ be such that $\|\Delta\|_\infty < \|\mathbb{L}_0\|^{-1}$. We begin by constructing the operators $\Phi_\Delta(t,s)$, $(t,s) \in \Gamma$. Consider the equations

$$y_\Delta(\cdot,s,x) = y_0(\cdot,s,x) + (\mathbb{L}_s\Delta_s y_\Delta(\cdot,s,x))(\cdot), \tag{28}$$

where $s \geq 0$ and $y_0(\cdot,s,x)$ is defined by (25). Here $\Delta_s : L^p(s,\infty;Y) \to L^p(s,\infty;U)$ is the multiplication operator described by $\Delta(\cdot)$. It is not difficult to show that $\|\mathbb{L}_s\| \leq \|\mathbb{L}_{\hat{s}}\|$, $\|\mathbb{M}_s\| \leq \|\mathbb{M}_{\hat{s}}\|$, $\|\Delta_s\| \leq \|\Delta_{\hat{s}}\|$ for $s \geq \hat{s}$. Because $\|\Delta_s\|_\infty$ is the operator norm of the multiplication operator Δ_s, $\mathbb{L}_s\Delta_s$ is a contraction on $L^p(s,\infty;Y)$. Furthermore, because $y_0(\cdot,s,x) \in L^p(s,\infty;Y)$ by (25), there must exist a unique solution $y_\Delta(\cdot,s,x) \in L^p(s,\infty;Y)$ of (28) for each $s \geq 0$, $x \in X$. Moreover,

$$\begin{aligned}\|y_\Delta(\cdot,s,x)\|_{L^p(s,\infty;Y)} &= \|(I - \mathbb{L}_s\Delta_s)^{-1}y_0(\cdot,s,x)\|_{L^p(s,\infty;Y)} \\ &\leq (1 - \|\mathbb{L}_0\|\|\Delta_0\|)^{-1}\|y_0(\cdot,s,x)\|_{L^p(s,\infty;Y)}, \qquad x \in X.\end{aligned}$$

Hence by Hypothesis 7, there exists a $K_\Delta > 0$ such that

$$\|y_\Delta(\cdot,s,x)\|_{L^p(s,\infty;Y)} \leq K_\Delta\|x\|_X, \qquad s \geq 0, \qquad x \in X, \tag{29}$$

and $x \mapsto y_\Delta(\cdot,s,x)$ is bounded linear operator from X to $L^p(s,\infty;Y)$. Now define $x_\Delta(\cdot,s,x)$ by

$$x_\Delta(\cdot,s,x) = \Phi(\cdot,s)x + (\mathbb{M}_s\Delta_s y_\Delta(\cdot,s,x))(\cdot), \tag{30}$$

for $s \geq 0$, $x \in X$. Then Hypotheses 4 and 5 and the exponential stability of $\Phi(\cdot,\cdot)$ imply

$$\|x_\Delta(t,s,x)\|_X \leq k_\Delta(t)\|x\|_X, \qquad t \geq s \geq 0, \qquad x \in X, \tag{31}$$

$$\|x_\Delta(\cdot,s,x)\|_{L^p(s,\infty;X)} \leq c\|x\|_X, \qquad s \geq 0, \qquad x \in X \tag{32}$$

for some constant $c > 0$ (independent of t, s) and some exponentially bounded $k_\Delta(t)$. In particular, $\Phi_\Delta(t,s) : x \mapsto x_\Delta(t,s,x)$ is a bounded linear operator on X for all $t \geq s \geq 0$ and

$$\|\Phi_\Delta(t,s)\|_{\mathcal{L}(X)} \leq k_\Delta(t) \qquad t \geq s \geq 0. \tag{33}$$

Next, we show that Φ_Δ satisfies (19)–(21). It follows from Hypotheses 6 and 4 that $x_\Delta(t,s,x) \in \underline{X}$ for all $x \in \underline{X}$ and almost every $t \geq s$, $s \geq 0$, whence (19). By Hypotheses 7, 4, and 8, $E(\cdot)x_\Delta(\cdot,s,x) \in L^p(s,\infty;Y)$ is well defined for $x \in X$, $s \geq 0$ (making use of (25)), and we have, from (28),

$$E(\cdot)x_\Delta(\cdot,s,x) = y_0(\cdot,s,x) + (\mathbb{L}_s\Delta_s y_\Delta(\cdot,s,x))(\cdot) = y_\Delta(\cdot,s,x), \qquad x \in X, \qquad s \geq 0.$$

Thus,

$$E(\cdot)\Phi_\Delta(\cdot,s)x = (I - \mathbb{L}_s\Delta_s)^{-1}E(\cdot)\Phi(\cdot,s)x = y_\Delta(\cdot,s,x), \qquad x \in X, \qquad s \geq 0. \tag{34}$$

So (20) and (21) follow from (29) and (30).

Now we show that Φ_Δ is uniquely determined by the above properties. Suppose that $(\hat{\Phi}_\Delta(t,s))_{(t,s)\in\Gamma}$ is another mild evolution operator on X satisfying (19)–(21). Then $\hat{y}(\cdot,s,x) = E(\cdot)\hat{\Phi}_\Delta(\cdot,s)x \in L^p(s,\infty;Y)$ is well defined for every $x \in \underline{X}$, $s \geq 0$ and satisfies $\hat{y}(\cdot,s,x) = y_0(\cdot,s,x) + (\mathbb{L}_s\Delta_s\hat{y}(\cdot,s,x))(\cdot)$, so that $E(\cdot)\hat{\Phi}_\Delta(\cdot,s)x = y_\Delta(\cdot,s,x)$. Hence (21) implies

$$\hat{\Phi}_\Delta(t,s)x = \Phi(t,s)x + \int_s^t \Phi(t,\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)d\rho = \Phi_\Delta(t,s)x, \qquad x \in \underline{X}.$$

It remains to be proven that Φ_Δ is an exponentially stable mild evolution operator. Clearly, property (i) of Definition 2.9 is satisfied. To prove the evolution property (ii), let $t \geq \hat{s} \geq s \geq 0$. Then

$$\begin{aligned}\Phi_\Delta(t,s)x = \Phi(t,\hat{s})\Phi(\hat{s},s)x \; &+ \int_{\hat{s}}^t \Phi(t,\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)d\rho \\ &+ \int_s^{\hat{s}} \Phi(t,\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)d\rho.\end{aligned} \tag{35}$$

Because $\Phi(t,\hat{s}) \in \mathcal{L}(\overline{X})$, we have

$$\int_s^{\hat{s}} \Phi(t,\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)d\rho = \Phi(t,\hat{s})\int_s^{\hat{s}} \Phi(\hat{s},\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)d\rho,$$

and therefore we obtain from (35) and (30)

$$\Phi_\Delta(t,s)x = \Phi(t,\hat{s})\Phi_\Delta(\hat{s},s)x + \int_{\hat{s}}^t \Phi(t,\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)\,d\rho. \tag{36}$$

Thus, $y(\cdot) = y_\Delta(\cdot,s,x) = E(\cdot)\Phi_\Delta(\cdot,s)x$ satisfies

$$y(\cdot) = y_0(\cdot,\hat{s},\Phi_\Delta(\hat{s},s)x) + (\mathbb{L}_{\hat{s}}\Delta_{\hat{s}}y(\cdot))(\cdot).$$

Because this equation has a unique solution in $L^p(\hat{s},\infty;Y)$, we conclude that

$$y_\Delta(\cdot,s,x) = y_\Delta(\cdot,\hat{s},\Phi_\Delta(\hat{s},s)x) = E(\cdot)\Phi_\Delta(\cdot,\hat{s})\Phi_\Delta(\hat{s},s)x.$$

Substituting in (36) for $y_\Delta(\cdot, s, x)$ and applying (30), we obtain $\Phi_\Delta(t,s)x = \Phi_\Delta(t,\hat{s})\Phi_\Delta(\hat{s},s)x$.

We now show that $\Phi_\Delta(\cdot,s)x = x_\Delta(\cdot,s,x) \in C(s,\infty;X)$ for all $x \in X$, $s \geq 0$, and $\Phi_\Delta(t,\cdot)x = x_\Delta(t,\cdot,x) \in C(0,t;X)$ for all $x \in X$, $t > 0$. The former statement is immediate because the first term on the RHS of (30) is continuous by Hypothesis 2 and the second term is continuous by Hypothesis 4. To prove the latter statement, let $t \geq \hat{s} \geq s$, then

$$\begin{aligned} x_\Delta(t,s,x) - x_\Delta(t,\hat{s},x) &= \Phi(t,s)x - \Phi(t,\hat{s})x + \Phi(t,\hat{s})\int_s^{\hat{s}} \Phi(\hat{s},\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)d\rho \\ &\quad + \int_{\hat{s}}^t \Phi(t,\rho)D(\rho)\Delta(\rho)[y_\Delta(\rho,s,x) - y_\Delta(\rho,\hat{s},x)]\,d\rho. \end{aligned}$$

So by Hypothesis 4,

$$\begin{aligned} \|x_\Delta(t,s,x) - x_\Delta(t,\hat{s},x)\|_X &\leq \|\Phi(t,s)x - \Phi(t,\hat{s})x\|_X \\ &\quad + M\left\| \int_s^{\hat{s}} \Phi(\hat{s},\rho)D(\rho)\Delta(\rho)y_\Delta(\rho,s,x)d\rho \right\|_X \\ &\quad + k(t)\|\Delta_0\|\|y_\Delta(\cdot,s,x) - y_\Delta(\cdot,\hat{s},x)\|_{L^p(\hat{s},t;Y)}, \end{aligned} \tag{37}$$

where M is taken from (13). Now

$$\begin{aligned} y_\Delta(\rho,s,x) - y_\Delta(\rho,\hat{s},x) &= E(\rho)[\Phi(\rho,s)x - \Phi(\rho,\hat{s})x] \\ &\quad + E(\rho)\Phi(\rho,\hat{s})\int_s^{\hat{s}} \Phi(\hat{s},\sigma)D(\sigma)\Delta(\sigma)y_\Delta(\sigma,s,x)d\sigma \\ &\quad + E(\rho)\int_{\hat{s}}^{\rho} \Phi(\rho,\sigma)D(\sigma)\Delta(\sigma)[y_\Delta(\sigma,s,x) - y_\Delta(\sigma,\hat{s},x)]d\sigma. \end{aligned}$$

By Hypothesis 7,

$$\begin{aligned} &\|y_\Delta(\cdot,s,x) - y_\Delta(\cdot,\hat{s},x)\|_{L^p(\hat{s},\infty;Y)} \\ &\leq K\left[\|(\Phi(\hat{s},s) - I_X)x\| + \left\| \int_s^{\hat{s}} \Phi(\hat{s},\sigma)D(\sigma)\Delta(\sigma)y_\Delta(\sigma,s,x)d\sigma \right\|\right] \\ &\quad + \|\mathbb{L}_{\hat{s}}\|\|\Delta_{\hat{s}}\|\|y_\Delta(\cdot,s,x) - y_\Delta(\cdot,\hat{s},x)\|_{L^p(\hat{s},\infty;Y)}. \end{aligned} \tag{38}$$

First we fix s. The first and second terms on the RHS of (37) go to zero as $\hat{s} \downarrow s$ by the strong continuity of $\Phi(t,\cdot)$ and by Hypothesis 4. By the same argument, the term in brackets on the RHS of (38) goes to zero as $\hat{s} \downarrow s$. Because $\|\mathbb{L}_{\hat{s}}\|\|\Delta_{\hat{s}}\| \leq \|\mathbb{L}_0\|\|\Delta_0\| < 1$, (38) implies $\|y_\Delta(\cdot,s,x) - y_\Delta(\cdot,\hat{s},x)\|_{L^p(\hat{s},\infty;Y)} \to 0$ as $\hat{s} \downarrow s$. Hence, the third term in (37) goes to zero as $\hat{s} \downarrow s$.

Now fix $\hat{s}$. By Hypothesis 4,

$$\begin{aligned} &\left\| \int_s^{\hat{s}} \Phi(\hat{s},\sigma)D(\sigma)\Delta(\sigma)y_\Delta(\sigma,s,x)d\sigma \right\|_X \\ &\quad \leq k(\hat{s})\|\Delta_0\|\|y_\Delta(\cdot,s,x)\|_{L^p(s,\hat{s};Y)} \\ &\quad \leq k(\hat{s})\|\Delta_0\|(1 - \|\mathbb{L}_0\|\,\|\Delta_0\|)^{-1}\|y_0(\cdot,s,x)\|_{L^p(s,\hat{s};Y)} \to 0 \end{aligned}$$

as $s \uparrow \hat{s}$ by Hypothesis 7. Hence the second term on the RHS of (37) and the term in brackets on the RHS of (38) tend to zero as $s \uparrow \hat{s}$. It follows as above that the third and the first term on the RHS of (37) go to zero as $s \uparrow \hat{s}$. This proves $x_\Delta(t,\cdot,x) \in C(0,t;X)$.

Altogether we see that $\Phi_\Delta(\cdot,\cdot)$ satisfies (i)–(iii) of Definition 2.9, that is, $\Phi_\Delta(\cdot,\cdot)$ is a mild evolution operator that is exponentially bounded on Γ because of (33). Hence $\beta(\Phi_\Delta) < \infty$ and from (32),

$$\int_s^\infty \|\Phi_\Delta(t,s)x\|^p dt \le c^p \|x\|_X^p, \qquad s \ge 0, \qquad x \in X. \tag{39}$$

Applying Proposition 2.12 it follows that $\Phi_\Delta(\cdot,\cdot)$ is exponentially stable on $[0,\infty)$ and this completes the proof. □

Remark 3.3. If we extend $y_\Delta(\cdot,s,x) = E(\cdot)\Phi_\Delta(\cdot,s)x \in L^p(s,\infty;Y)$ trivially to $\mathbb{R}_+$, for every $s \ge 0$ and $x \in X$,

$$\bar{y}(t,s,x) = \begin{cases} E(t)\Phi_\Delta(t,s)x & \text{a.e. } t \ge s, \\ 0 & t \in [0,s). \end{cases}$$

The above proof shows that the map $s \mapsto \bar{y}(\cdot,s,x)$ from $\mathbb{R}_+$ into $L^p(\mathbb{R}_+,Y)$ is continuous for every $x \in X$.

Remark 3.4. (i) All the previous definitions and results can be stated analogously for an arbitrary time interval $[t_0,\infty)$ instead of the fixed interval $[0,\infty)$. Thus, if we assume $\|\Delta\|_\infty < \|\mathbb{L}_{t_0}\|^{-1}$ for some $t_0 \ge 0$ in the previous proof, then the existence, uniqueness, and exponential stability of the perturbed mild evolution operator Φ_Δ *on the interval* $[t_0,\infty)$ follows. Contrary to the finite-dimensional case [10], however, we are not able to extend this evolution operator backward to the interval $[0,\infty)$. Therefore we cannot exclude the possibility that the stability radius may be increased by restricting the data (Φ,D,E) to intervals $[t_0,\infty)$, $t_0 > 0$.

(ii) We have analyzed the existence of a perturbed evolution operator and its properties in the context of robust stability. In so doing we have not separated the problem of existence from that of stability. Therefore, we assumed Φ to be exponentially stable and imposed conditions in Hypotheses 7, 5, and 8, which enabled us to first construct $y_\Delta(\cdot,s,x)$ via (28) and then $x_\Delta(\cdot,s,x)$ by (30) on an infinite interval $[s,\infty)$. Our basic method could also have been applied on finite time intervals without the assumption that Φ is exponentially stable. For example, we would seek $y_\Delta(\cdot,s,x)$ as a fixed point of (28) in $L^p(s,t;Y)$. Then $x_\Delta(\cdot,s,x)$ is determined on $[s,t]$ via (30). Proceeding stepwise on successive intervals—under suitable assumptions that ensure that the sums of the lengths of the intervals diverge—we construct $x_\Delta(\cdot,s,x)$ on $[s,\infty)$.

(iii) In the finite-dimensional case it has been shown [10] that if $p = 2$, then $\|\mathbb{L}_{t_0}\|^{-1}$ is the largest value of ρ for which the differential Riccati equation

$$\dot{P}(t) + A^*(t)P(t) + P(t)A(t) - \rho^2 E^*(t)E(t) - P(t)D(t)D^*(t)P(t) = 0, \quad t \ge t_0 \tag{40}$$

has a bounded Hermitian solution. We believe it should be possible to extend this result to systems satisfying the assumptions of Theorem 3.2, but the differential Riccati equation must be replaced by an integral Riccati equation

$$P(t)x = -\int_t^\infty \Phi^*(s,t)[\rho^2 E^*(s)E(s) + P(s)D(s)D^*(s)P(s)]\Phi(s,t)x\,ds, \qquad x \in \underline{X}, \quad t \ge t_0.$$

In the following we apply Theorem 3.2 to a time-invariant setting. Heretofore, our development is equally valid for the fields $\mathbb{R}$ or $\mathbb{C}$; however, for the second part of the next theorem it is essential that $\mathbb{K} = \mathbb{C}$, because our construction of a destabilizing perturbation $\Delta \in \mathcal{L}(Y,U)$ presupposes that U, Y are *complex* Hilbert spaces.

THEOREM 3.5. (i) *Suppose the system* $(S(\cdot), D, E)$ *satisfies the following conditions*:

(a) *Hypothesis* 1 *holds with* D, E *time-invariant*, $S(\cdot)$ *is an exponentially stable* C_0-*semigroup on* X, *and* $S(t)$ *can be extended to a bounded linear operator on* $\overline{X}$ *for every* $t \geq 0$, *such that* $S(t-\cdot)Du(\cdot)$ *is integrable in* $\overline{X}$ *on* $[0,t]$, *for every* $u(\cdot) \in L^p(0,\infty;U)$.

(b) *For every* $u(\cdot) \in L^p(0,\infty;U)$, *the input-state map* $\mathbb{M}_0$,

$$(\mathbb{M}_0 u)(t) = \int_0^t S(t-\rho)Du(\rho)d\rho, \qquad t \geq 0, \qquad u(\cdot) \in L^p(0,\infty;U)$$

satisfies $(\mathbb{M}_0 u)(t) \in \underline{X}$ *for almost every* $t \geq 0$, *and there exists* $\tau > 0$, $k > 0$ *such that*

$$\|(\mathbb{M}_0 u)(\tau)\|_X \leq k\|u(\cdot)\|_{L^p(0,\tau;U)}, \qquad u(\cdot) \in L^p(0,\infty;U). \tag{41}$$

(c) $S(t)x \in \underline{X}$ *for any* $x \in \underline{X}$ *and almost every* $t \geq 0$, $ES(\cdot)x : [0,\infty) \to Y$ *is* p-*integrable for every* $x \in \underline{X}$, *and there exists a constant* $K > 0$ *such that*

$$\|ES(\cdot)x\|_{L^p(0,\infty;Y)} \leq K\|x\|_X, \qquad x \in \underline{X}.$$

(d) *The input–output operator* $\mathbb{L}_0$,

$$(\mathbb{L}_0 u)(t) = E\int_0^t S(t-\rho)Du(\rho)d\rho, \qquad t \geq 0, \qquad u(\cdot) \in L^p(0,\infty;U),$$

satisfies $\mathbb{L}_0 \in \mathcal{L}(L^p(0,\infty;U),\ L^p(0,\infty;Y))$.

Then

$$r(S(\cdot);D,E) \geq \|\mathbb{L}_0\|^{-1}_{\mathcal{L}(L^p(0,\infty;U),L^p(0,\infty;Y))}. \tag{42}$$

(ii) *Suppose additionally that* $\mathbb{K} = \mathbb{C}$, U, Y *are Hilbert spaces*, $p = 2$, *and*

(e) S *extends to an exponentially stable semigroup on* $\overline{X}$ *with generator* $A^{\overline{X}}$ *such that* $D(A^{\overline{X}}) \subset \underline{X}$, *where the embedding is dense and continuous with respect to the graph norm on* $D(A^{\overline{X}})$. *Then*

$$r^c(S;D,E) = r(S;D,E) = \|\mathbb{L}_0\|^{-1}_{\mathcal{L}(L^2(0,\infty;U),L^2(0,\infty;Y))} = [\sup_{\omega\in\mathbb{R}}\|H(i\omega)\|_{\mathcal{L}(U,Y)}]^{-1}, \tag{43}$$

where $r^c(S;D,E)$ *is defined in the same way as* $r(S;D,E)$ *except the perturbations* Δ *are restricted to be constant and* $H(s) = E(sI - A^{\overline{X}})^{-1}D$, $s \in \mathbb{C}\backslash\sigma(A^{\overline{X}})$.

Proof. (i) It is shown in [27] that (a) and (b) imply that for every $u(\cdot) \in L^p(0,\infty;U)$ and $t \geq 0$, $(\mathbb{M}_0 u)(\cdot) \in C(0,t;X)$, and there exists a constant $k > 0$ such that

$$\|(\mathbb{M}_0 u)(t)\|_X \leq k\|u(\cdot)\|_{L^p(0,t;U)}, \qquad t \geq 0.$$

Moreover, it follows from the exponential stability of $S(\cdot)$ that $\mathbb{M}_0 \in \mathcal{L}(L^p(0,\infty;U), L^p(0,\infty;X))$ [29]. Also

$$\lim_{s\to t}\|ES(\cdot - s)x\|_{L^p(s,t;Y)} = \lim_{s\to t}\|ES(\cdot)x\|_{L^p(0,t-s;Y)} = 0, \qquad x \in \underline{X},\ t > 0.$$

So (a)–(d) imply Hypotheses 1–8 for the system $(S(\cdot), D, E)$, and hence we may apply Theorem 3.2 to obtain (42).

(ii) Now assume the hypotheses of the second part of the theorem. In [29] it is shown that under these assumptions,

$$\|\mathbb{L}_0\|_{\mathcal{L}(L^2(0,\infty;U),L^2(0,\infty;Y))} = \sup_{s\in\mathbb{C}_0}\|H(s)\|_{\mathcal{L}(U,Y)}, \tag{44}$$

where $\mathbb{C}_0 = \{s \in \mathbb{C};\ \mathrm{Re}\, s > 0\}$, $H(s) = E_L(sI - A^{\overline{X}})^{-1}D$, and E_L is the Lebesgue extension of E defined by

$$E_L x = \lim_{\tau \to 0} E\frac{1}{\tau}\int_0^\tau S(s)x\, ds,$$

with domain $D(E_L) = \{x \in X;$ the above lim exists$\}$. Suppose $x \in D(A^{\overline{X}}) \subset \underline{X}$, $(A^{\overline{X}})^{-1}x = y$, then $y \in D((A^{\overline{X}})^2)$, and because S is a semigroup on the Banach space $D(A^{\overline{X}})$ with generator $A^{\overline{X}}|D((A^{\overline{X}})^2)$, we have

$$\begin{aligned}
Ex &= E\, A^{\overline{X}} y,\\
&= E \lim_{\tau \to 0} \frac{1}{\tau}(S(\tau) - I)y \quad \text{(where the limit is taken in } D(A^{\overline{X}})),\\
&= \lim_{\tau \to 0} E\frac{1}{\tau}(S(\tau) - I)y \quad \text{(where the limit is taken in } Y),\\
&= \lim_{\tau \to 0} E\frac{1}{\tau}\int_0^\tau \frac{dS(s)y}{ds}\, ds = \lim_{\tau \to 0} E\frac{1}{\tau}\int_0^\tau S(s)A^{\overline{X}} y\, ds,\\
&= \lim_{\tau \to 0} E\frac{1}{\tau}\int_0^\tau S(s)x\, ds = E_L x.
\end{aligned}$$

Hence $D(A^{\overline{X}}) \subset D(E_L)$ and $Ex = E_L x$ for all $x \in D(A^{\overline{X}})$. By the exponential stability of S on $\overline{X}$, there exists $\delta > 0$ such that the resolvent $(sI - A^{\overline{X}})^{-1}$ is analytic on $\mathbb{C}_{-\delta} = \{s \in \mathbb{C};\ \mathrm{Re}\, s > -\delta\}$. Hence $(sI - A^{\overline{X}})^{-1}Du \in D(A^{\overline{X}})$ for all $s \in \mathbb{C}_{-\delta}$, $u \in U$ and

$$H(s) = E(sI - A^{\overline{X}})^{-1}D, \qquad s \in \mathbb{C}_{-\delta}$$

is analytic on $\mathbb{C}_{-\delta}$. It follows from the maximum principle and (44) that

$$\|\mathbb{L}_0\|_{\mathcal{L}(L^2(0,\infty;U),L^2(0,\infty;Y))} = \sup_{\omega \in \mathbb{R}} \|H(\imath\omega)\|_{\mathcal{L}(U,Y)}. \tag{45}$$

Now suppose, by way of contradiction, that $\|H(\imath\omega)\|^{-1}_{\mathcal{L}(U,Y)} < r^c(S; D, E)$ and choose $\varepsilon > 0$ small enough to satisfy $[\sup_{\omega \in \mathbb{R}} \|H(\imath\omega)\| - \varepsilon]^{-1} < r^c(S; D, E)$. There exist $\omega_0 \in \mathbb{R}$, $u \in U$ such that $\|u\| = 1$ and $\|H(\imath\omega_0)u\| \geq \sup_{\omega \in \mathbb{R}} \|H(\imath\omega)\| - \varepsilon$. Define $\Delta \in \mathcal{L}(Y, U)$ by

$$\Delta y = \frac{\langle y,\, H(\imath\omega_0)u\rangle u}{\|H(\imath\omega_0)u\|^2}, \qquad y \in Y.$$

Then $\|\Delta\| = \|H(\imath\omega_0)u\|^{-1}$, and so

$$\left[\sup_{\omega \in \mathbb{R}} \|H(\imath\omega)\|\right]^{-1} \leq \|\Delta\| \leq \left[\sup_{\omega \in \mathbb{R}} \|H(\imath\omega)\| - \varepsilon\right]^{-1}.$$

Setting $y(t) = e^{\imath\omega_0 t}H(\imath\omega_0)u$, we obtain $\Delta y(t) = e^{\imath\omega_0 t}u$ and

$$\int_0^t S(t - \rho)D\Delta y(\rho)d\rho = e^{\imath\omega_0 t}\int_0^t e^{-\imath\omega_0(t-\rho)}S(t - \rho)Du\, d\rho.$$

But $x = (\imath\omega_0 I - A^{\overline{X}})^{-1}\, Du \in D(A^{\overline{X}})$ satisfies

$$\frac{d}{d\rho}[e^{-\imath\omega_0(t-\rho)}S(t - \rho)x] = e^{-\imath\omega_0(t-\rho)}S(t - \rho)(\imath\omega_0 I - A^{\overline{X}})x = e^{-\imath\omega_0(t-\rho)}S(t - \rho)Du.$$

Hence

$$\int_0^t S(t-\rho)D\Delta y(\rho)d\rho = e^{i\omega_0 t}\int_0^t \frac{d}{d\rho}[e^{-i\omega_0(t-\rho)}S(t-\rho)x]d\rho$$
$$= e^{i\omega_0 t}x - S(t)x,$$

and $y(t) = e^{i\omega_0 t}Ex$. So $x(t) = e^{i\omega_0 t}x$ solves

$$x(t) = S(t)x + \int_0^t S(t-\rho)D\Delta Ex(\rho)d\rho.$$

But because $\|\Delta\| < r^c(S; D, E)$, the mild perturbed evolution operator Φ_Δ must be exponentially stable and $x(t) = \Phi_\Delta(t,0)x$. This is a contradiction and so $[\sup_{\omega\in\mathbb{R}}\|H(i\omega)\|]^{-1} \geq r^c(S; D, E)$. Equation (43) now follows from (45), (42), and the fact that, by definition, $r^c(S; D, E) \geq r(S; D, E)$. □

Remark 3.6. (i) Weiss has introduced the notion of a regular linear system for which the input and output functions are locally L^p, and on any finite interval, the final state and the output function depend continuously on the initial state and input function. These systems provide a unifying framework for a large class of phenomena usually described as linear, time-invariant, and well posed. If, with the notation in [29], we set

$$S(\cdot) = \mathbb{T}(\cdot), \qquad \overline{X} = X_{-1}, \qquad \underline{X} = D(C_L), \qquad E = C_L, \qquad D = B,$$

then for exponentially stable $\mathbb{T}(\cdot)$, the first part of the theorem applies to regular linear systems. In [28] Weiss considers time-invariant perturbations of regular linear systems without the stability constraint. He introduces a *well-posedness radius* with respect to output feedback and determines a lower bound that is exact when U and Y are finite-dimensional. Our lower bound (42) is, of course, a more conservative one for the well-posedness radius, but it does have the advantage that time-varying perturbations are allowed.

(ii) Even for semigroups $S(t) = e^{At}$ on finite-dimensional spaces it is well known [12] that in the *real* case ($\mathbb{K} = \mathbb{R}$) equality does not in general hold in (43).

(iii) A stability radius for complex time-invariant perturbations of infinite-dimensional systems was introduced in [23]. It was assumed that S restricts to an exponentially stable semigroup on $\underline{X}$; U, Y are Hilbert spaces; $p = 2$; and (b), (c), (e) hold. Instead of (d), the authors imposed the condition $\mathbb{M}_0 \in \mathcal{L}(L^2(0,\infty;U), L^2(0,\infty;\underline{X}))$. This clearly implies (d) because $\mathbb{L}_0 = E\mathbb{M}_0$, but splitting condition (d) in this way is in general much more restrictive. Under these stronger assumptions it was shown that the time-invariant complex stability radius equals $\|\mathbb{L}_0\|^{-1}$. So Theorem 3.5 considerably improves this result.

(iv) Of course, we cannot expect more than an inequality from Theorem 3.2 because it is concerned with time-varying systems and we know that equality does not hold even for scalar time-varying systems [10].

4. Perturbations of weak evolution operators. In this section we suppose that Φ is an exponentially stable *weak* evolution operator with generator A and we look for conditions under which the perturbed mild evolution operator $\Phi_\Delta(\cdot,\cdot)$ constructed in §3 is a weak evolution operator with generator $A(\cdot) + D(\cdot)\Delta(\cdot)E(\cdot)$. An immediate problem is that $D(t)\Delta(t)E(t)x$ may not be in X for any nonzero $x \in \underline{X}$, $t \geq 0$. Hence we cannot expect Φ_Δ to be a weak evolution operator on X. However, for $x \in \underline{X}$, $D(t)\Delta(t)E(t)x \in \overline{X}$, so it may be possible to impose conditions which ensure that Φ_Δ is a weak evolution operator on $\overline{X}$. If this is to be the case, an obvious first condition is that Φ_Δ is a mild evolution operator on $\overline{X}$. And because we want to apply Theorem 3.2, we have to set $\overline{X} = X$. So instead of the triple $(\underline{X}, X, \overline{X})$, in this

section we only consider the pair $(\underline{X}, X)$. As a consequence of setting $\overline{X} = X$, some of the hypotheses of Theorem 3.2 are automatically satisfied. For example, if Φ is a mild evolution operator on X, then necessarily Hypothesis 3 is satisfied (because $D(t) \in \mathcal{L}(U, X)$, $t \geq 0$). Also Hypothesis 4 is satisfied with the exception of the requirements that $k(t)$ be exponentially bounded and that $u(\cdot) \in L^p(s, t; U)$, $s \geq 0$ implies $(\mathbb{M}_s u)(t) \in \underline{X}$ for almost every $t \geq s$. We have, of course, paid the price for these advantages in that the unboundedness of the term $D(t)\Delta(t)E(t)$ has been severely reduced.

To prove that Φ_Δ is a weak evolution operator on X, we need to impose further conditions. Obviously, Φ must be a weak evolution operator on X. We recall that this means that for all $t \geq 0$, there is a linear operator $A(t)$ on X with a time-independent domain $D(A)$ dense in X, such that for all $x \in D(A)$, $A(\cdot)x$ is piecewise continuous in the sense of the Assumption and

$$\Phi(t, s)x - x = \int_s^t \Phi(t, \rho)A(\rho)x \, d\rho, \qquad (t, s) \in \Gamma.$$

The other additional hypotheses are as follows.

Hypothesis 9. For each $t > 0$ there exists $\gamma(t) \geq 0$ such that

$$\|(\mathbb{M}_s u)(t)\|_{\underline{X}} \leq \gamma(t)\|u(\cdot)\|_{L^p(s,t;U)}, \qquad u(\cdot) \in L^p(s, t; U). \tag{46}$$

Hypothesis 10. $D(A) \subset \underline{X}$ and for every $(t, s) \in \Gamma$, $x \in D(A)$, and $u(\cdot) \in L^p(s, t; U)$ we have

$$\int_s^t E(t)\Phi(t, \sigma)A(\sigma)x \, d\sigma = E(t) \int_s^t \Phi(t, \sigma)A(\sigma)x \, d\sigma \quad (= E(t)[\Phi(t, s)x - x]),$$

$$\int_s^t E(t)\Phi(t, \sigma)D(\sigma)u(\sigma)d\sigma = E(t) \int_s^t \Phi(t, \sigma)D(\sigma)u(\sigma)d\sigma \quad (= E(t)(\mathbb{M}_s u)(t)).$$

Remark 4.1. (i) If $\gamma(\cdot)$ is exponentially bounded, Hypothesis 9 can be viewed as a strengthening of Hypothesis 4. If, additionally, $\gamma(\cdot)$ is L^p-integrable on $[s, \infty)$, Hypothesis 9 then also implies Hypothesis 5.

(ii) It is not immediately clear that the integrals on the left-hand side (LHS) in Hypothesis 10 are well defined; however, we will see that this is a consequence of the ensuing lemmata.

Our main result is as follows.

THEOREM 4.2. *Suppose Hypotheses* 1–10 *hold with* $\overline{X} = X$, *and* Φ *is an exponentially stable weak evolution operator on* X *with generator* A. *If* $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y, U))$ *has norm* $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0\|^{-1}$, *then the perturbed mild evolution operator* Φ_Δ (*which exists and is exponentially stable by Theorem* 3.2) *is a weak evolution operator on* X *with generator* $A_\Delta(t) = A(t) + D(t)\Delta(t)E(t)$, $t \geq 0$ *and domain* $D(A_\Delta) = D(A)$.

To prove this theorem we use the following lemmata. The first shows that Hypothesis 7 is inherited by the perturbed evolution operator Φ_Δ.

LEMMA 4.3. *Under Hypotheses* 1–8, *for every* $x \in \underline{X}$, $t > 0$,

$$\lim_{s \to t} \|E(\cdot)\Phi_\Delta(\cdot, s)x\|_{L^p s,t;Y)} = 0, \tag{47}$$

and there exists a constant $K_\Delta \geq 0$ *such that*

$$\|E(\cdot)\Phi_\Delta(\cdot, s)x\|_{L^p(s,\infty;Y)} \leq K_\Delta \|x\|_X, \qquad x \in \underline{X}, \qquad s \geq 0. \tag{48}$$

Proof. Suppose $s \geq 0$, $x \in \underline{X}$ and let $y_\Delta(\cdot, s, x) = E(\cdot)\Phi_\Delta(\cdot, s)x$. It was shown in the proof of Theorem 3.2 that $y_\Delta(\cdot, s, x) \in L^p(s, \infty; Y)$ and

$$y_\Delta(\cdot, s, x) = (I - \mathbb{L}_s\Delta_s)^{-1}E(\cdot)\Phi(\cdot, s)x$$

(see (34)). Hence by Hypothesis 7,

$$\|y_\Delta(\cdot,s,x)\|_{L^p(s,\infty;Y)} \le (1-\|\mathbb{L}_0\|\|\Delta_0\|)^{-1}K\|x\|_X.$$

Thus, (48) follows with $K_\Delta = K/(1-\|\mathbb{L}_0\|\|\Delta_0\|)$. By (28) we have for all $t > s$,

$$y_\Delta(\cdot,s,x)|_{[s,t]} = y_0(\cdot,s,x)|_{[s,t]} + (\mathbb{L}_{s,t}\Delta_{s,t}y_\Delta(\cdot,s,x)|_{[s,t]})(\cdot)|_{[s,t]},$$

where $\mathbb{L}_{s,t}$, $\Delta_{s,t}$ denote the restrictions of the operators $\mathbb{L}_s$, Δ_s to $L^p(s,t;U) \subset L^p(s,\infty;U)$ respectively, $L^p(s,t;Y) \subset L^p(s,\infty;Y)$. Because $(1-\|\mathbb{L}_{s,t}\|\|\Delta_{s,t}\|)^{-1} \le (1-\|\mathbb{L}_0\|\|\Delta_0\|)^{-1}$, we obtain

$$\|y_\Delta(\cdot,s,x)\|_{L^p(s,t;Y)} \le (1-\|\mathbb{L}_0\|\|\Delta_0\|)^{-1}\|E(\cdot)\Phi(\cdot,s)x\|_{L^p(s,t;Y)}, \qquad t > s,$$

whence (47) follows. □

By (48), $E(\cdot)\Phi_\Delta(\cdot,s)x \in L^p(s,\infty;Y)$ is well defined for every $x \in X$, namely

$$E(\cdot)\Phi_\Delta(\cdot,s)x = \lim_{k\to\infty} E(\cdot)\Phi_\Delta(\cdot,s)x_k,$$

where (x_k) is a sequence in $\underline{X}$ converging toward x in X. For $s \ge 0$, $x \in X$ we denote by $E(t)\Phi_\Delta(t,s)x$, $t \ge s$ the values of a representative of the function class $E(\cdot)\Phi_\Delta(\cdot,s)x \in L^p(s,\infty;Y)$.

LEMMA 4.4. *Under Hypotheses* 1–8, *let* $(s,t) \in \Gamma$, $z(\cdot) \in PC(s,t;X)$, *and* $\Gamma_{s,t} := \{(\tau,\sigma);\ s \le \sigma \le \tau \le t\}$. *Then the maps* $z^t(\cdot,\cdot) : \Gamma_{s,t} \to X$ *and* $y(\cdot,\cdot) : \Gamma_{s,t} \to Y$ *defined by*

$$y(\tau,\sigma) = E(\tau)\Phi_\Delta(\tau,\sigma)z(\sigma), \qquad (\tau,\sigma) \in \Gamma_{s,t}, \tag{49}$$

$$z^t(\tau,\sigma) = \Phi(t,\tau)D(\tau)\Delta(\tau)y(\tau,\sigma), \qquad (\tau,\sigma) \in \Gamma_{s,t} \tag{50}$$

are (Bochner) integrable with respect to the Lebesgue measure on $\Gamma_{s,t}$, $\tau \mapsto \int_s^\tau y(\tau,\sigma)d\sigma$ *is p-integrable on* $[s,t]$, *and*

$$\int_s^t \int_\sigma^t z^t(\tau,\sigma)d\tau\, d\sigma = \int_s^t \int_s^\tau z^t(\tau,\sigma)d\sigma\, d\tau. \tag{51}$$

Proof. First we prove the statements concerning $y(\cdot,\cdot)$. For $\sigma \in [s,t]$, define $F(\sigma) \in L^p(s,t;Y)$ by

$$F(\sigma)(\tau) = \begin{cases} y(\tau,\sigma) & \text{if } (\tau,\sigma) \in \Gamma_{s,t}, \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

By Lemma 4.3, $F(\cdot) : [s,t] \to L^p(s,t;Y)$ is bounded because $z(\cdot)$ is bounded. We now show that $F(\cdot)$ is continuous at all $\sigma \in [s,t]$, where $z(\cdot)$ does not jump. Let $y_\Delta(\cdot,\sigma,x) = E(\cdot)\Phi_\Delta(\cdot,\sigma)x$ for all $x \in X$ so that $y(\tau,\sigma) = y_\Delta(\tau,\sigma,z(\sigma))$, $\tau \in [\sigma,t]$, $\sigma \in [s,t]$. For all $\hat\sigma \ge \sigma$ in $[s,t]$,

$$\begin{aligned}
\|F(\sigma) - F(\hat\sigma)\|^p_{L^p(s,t;Y)} &= \int_\sigma^{\hat\sigma} \|y(\tau,\sigma)\|_Y^p\, d\tau \\
&\quad + \int_{\hat\sigma}^t \|y(\tau,\sigma) - y(\tau,\hat\sigma)\|_Y^p\, d\tau \\
&= \int_\sigma^{\hat\sigma} \|y_\Delta(\tau,\sigma,z(\sigma))\|_Y^p\, d\tau \\
&\quad + \int_{\hat\sigma}^t \|y_\Delta(\tau,\sigma,z(\sigma)) - y_\Delta(\tau,\hat\sigma,z(\hat\sigma))\|_Y^p\, d\tau.
\end{aligned}$$

However,

$$\begin{aligned}\|y_\Delta(\cdot,\sigma,z(\sigma))\|_{L^p(\sigma,\hat\sigma;Y)} &\le \|y_\Delta(\cdot,\sigma,z(\hat\sigma))\|_{L^p(\sigma,\hat\sigma;Y)} + \|y_\Delta(\cdot,\sigma,z(\sigma)-z(\hat\sigma))\|_{L^p(\sigma,\hat\sigma;Y)}\\ &\le \|y_\Delta(\cdot,\sigma,z(\hat\sigma))\|_{L^p(\sigma,\hat\sigma;Y)} + K_\Delta\|z(\sigma)-z(\hat\sigma)\|_X\end{aligned}$$

by (48). The LHS clearly tends to zero as $\hat\sigma \downarrow \sigma$, whereas the first term on the RHS tends to zero as $\sigma \uparrow \hat\sigma$ by (47) and the second term tends to zero if $z(\cdot)$ is continuous at $\hat\sigma$. Now

$$\begin{aligned}\|y_\Delta(\cdot,\sigma,z(\sigma)) - y_\Delta(\cdot,\hat\sigma,z(\hat\sigma))\|_{L^p(\hat\sigma,t;Y)} &\le \|y_\Delta(\cdot,\sigma,z(\sigma)) - y_\Delta(\cdot,\hat\sigma,z(\sigma))\|_{L^p(\hat\sigma,t;Y)}\\ &\quad + \|y_\Delta(\cdot,\hat\sigma,(z(\sigma)-z(\hat\sigma)))\|_{L^p(\hat\sigma,t;Y)}.\end{aligned}$$

The first term on the RHS tends to zero as $\sigma \uparrow \hat\sigma$ or $\hat\sigma \downarrow \sigma$ by Remark 3.3 and the second term is majorized by $K_\Delta\|z(\sigma)-z(\hat\sigma)\|_X$. Therefore, $F(\cdot):[s,t]\to L^p(s,t;Y)$ is bounded and continuous at all continuity points of $z(\cdot)$, hence integrable on $[s,t]$. By Theorem III.11.17 in [7] there exists a strongly Lebesgue measurable function $f:[s,t]^2\to Y$ such that

$$F(\sigma) = f(\cdot,\sigma), \quad \text{a.e. } \sigma\in[s,t].$$

Hence $y(\cdot,\cdot):\Gamma_{s,t}\to Y$ is strongly Lebesgue measurable because it coincides almost everywhere on $\Gamma_{s,t}$ with $f(\cdot,\cdot)$. Moreover, by the same theorem in [7], $f(\tau,\cdot)$ is integrable on $[s,t]$ for almost all $\tau\in[s,t]$ and the integral $\int_s^t f(\tau,\sigma)d\sigma = \int_s^\tau y_\Delta(\tau,\sigma,z(\sigma))d\sigma$, as a function of τ, is a representative of the function class $\int_s^t F(\sigma)d\sigma$ in $L^p(s,t;Y)$. This proves the statements concerning $y(\cdot,\cdot)$.

The statements concerning $z^t(\cdot,\cdot)$ are proved similarly. For $\sigma\in[s,t]$ define $G(\sigma)\in L^p(s,t;X)$ by

$$G(\sigma)(\tau) = \begin{cases} z^t(\tau,\sigma) & \text{if } (\tau,\sigma)\in\Gamma_{s,t},\\ 0 & \text{otherwise.}\end{cases}$$

By assumption, there exist M, $\omega>0$ such that $\|\Phi(t,\tau)\|_{\mathcal{L}(X)}\le Me^{-\omega(t-\tau)}$. Let

$$c^p = M^p \sup_{\tau\in[s,t]}\|D(\tau)\|^p_{\mathcal{L}(U,\overline{X})}\|\Delta(\tau)\|^p_{\mathcal{L}(Y,U)}.$$

Then, for all $\hat\sigma\ge\sigma$ in $[s,t]$:

$$\begin{aligned}\|G(\sigma)-G(\hat\sigma)\|^p_{L^p(s,t;X)} &= \int_\sigma^{\hat\sigma}\|z^t(\tau,\sigma)\|_X^p\,d\tau + \int_{\hat\sigma}^t\|z^t(\tau,\sigma)-z^t(\tau,\hat\sigma)\|_X^p\,d\tau\\ &\le c^p\Bigg[\int_\sigma^{\hat\sigma}\|y_\Delta(\tau,\sigma,z(\sigma))\|_Y^p\,d\tau\\ &\quad + \int_{\hat\sigma}^t\|y_\Delta(\tau,\sigma,z(\sigma)) - y_\Delta(\tau,\hat\sigma,z(\hat\sigma))\|_Y^p\,d\tau\Bigg]\\ &= c^p\|F(\sigma)-F(\hat\sigma)\|^p_{L^p(s,t;Y)}.\end{aligned}$$

Hence $G(\cdot):[s,t]\to L^p(s,t;X)$ is bounded and continuous at all continuity points of $z(\cdot)$. By the same reasoning as above, we conclude that $z(\cdot,\cdot):\Gamma_{s,t}\to X$ is strongly Lebesgue measurable. Moreover, we have

$$\begin{aligned}\int_s^t\int_\sigma^t\|z^t(\tau,\sigma)\|_X\,d\tau\,d\sigma &\le \int_s^t\int_\sigma^t Me^{-\omega(t-\tau)}\|D(\tau)\|\|\Delta(\tau)\|\|y_\Delta(\tau,\sigma,z(\sigma))\|_Y\,d\tau\,d\sigma\\ &\le c\int_s^t(t-\sigma)^{1/q}\|y_\Delta(\cdot,\sigma,z(\sigma))\|_{L^p(\sigma,t;Y)}\,d\sigma\\ &\le c(t-s)^{1/q}\int_s^t K_\Delta\|z(\sigma)\|_X\,d\sigma<\infty,\end{aligned}$$

where $1/p + 1/q = 1$. By Tonelli's theorem [7, Cor. III.11.15] it follows that $z^t(\cdot,\cdot)$ is (Bochner) integrable on $\Gamma_{s,t}$ (with respect to the induced Lebesgue measure) and so Fubini's theorem yields (51). □

As a consequence of this lemma, the map $\sigma \mapsto y(t,\sigma) = E(t)\Phi_\Delta(t,\sigma)z(\sigma)$ is integrable on $[0,t]$ for almost all $t > 0$, if $z(\cdot) \in PC(s,t;X)$. Setting $\Delta = 0$ we see that the same is true for $E(t)\Phi(t,\sigma)z(\sigma)$ and thus the LHS integrals in Hypothesis 10 are well defined.

LEMMA 4.5. *Under the assumptions of Theorem* 4.2, *suppose that for* $(t,s) \in \Gamma$, $z(\cdot) \in PC(s,t;X)$. *If* $z^t(\tau,\sigma)$ *is defined by* (50), *then* $Z^t(\cdot) : \sigma \mapsto \int_\sigma^t z^t(\tau,\sigma)d\tau$ *is a piecewise continuous map from* $[s,t]$ *into* $\underline{X}$. *In particular,*

$$\int_s^t E(t)\int_\sigma^t z^t(\tau,\sigma)d\tau\,d\sigma = E(t)\int_s^t\int_\sigma^t z^t(\tau,\sigma)d\tau\,d\sigma. \tag{53}$$

Proof. Let $t \geq \hat\sigma \geq \sigma \geq s$. Then

$$Z^t(\sigma) - Z^t(\hat\sigma) = \int_\sigma^{\hat\sigma} z^t(\tau,\sigma)d\tau + \int_{\hat\sigma}^t \Phi(t,\tau)D(\tau)\Delta(\tau)[y_\Delta(\tau,\sigma,z(\sigma)) - y_\Delta(\tau,\hat\sigma,z(\hat\sigma))]d\tau.$$

Setting $s = \sigma$ and $u(\tau) = 0$, $\tau \geq \hat\sigma$ in (46) yields

$$\left\| \int_\sigma^{\hat\sigma} \Phi(t,\tau)D(\tau)u(\tau)d\tau \right\|_{\underline{X}} \leq \gamma(t)\|u(\cdot)\|_{L^p(\sigma,\hat\sigma;U)}, \qquad u(\cdot) \in L^p(\sigma,\hat\sigma;U).$$

Hence by Hypothesis 9 and (49),

$$\begin{aligned}
&\|Z^t(\sigma) - Z^t(\hat\sigma)\|_{\underline{X}} \\
&\quad \leq \gamma(t)\|\Delta_0\|\,[\|y_\Delta(\cdot,\sigma,z(\sigma))\|_{L^p(\sigma,\hat\sigma,Y)} + \|y_\Delta(\cdot,\sigma,z(\sigma)) - y_\Delta(\cdot,\hat\sigma,z(\hat\sigma))\|_{L^p(\hat\sigma,t;Y)}] \\
&\quad \leq C(p)\gamma(t)\|\Delta_0\|\,\|F(\sigma) - F(\hat\sigma)\|_{L^p(s,t;Y)},
\end{aligned}$$

where $F(\cdot)$ is given by (52) and $C(p)$ is some constant depending on p. This shows that $Z^t(\cdot) : [s,t] \to X$ is piecewise continuous and (53) is a consequence of this fact. □

We now have all the tools necessary to establish the theorem.

Proof of Theorem 4.2. Suppose $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$ has norm $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0\|^{-1}$ and let $s \geq 0$, $x \in D(A)$. Then $A_\Delta(\cdot)x \in PC(s,\infty;X)$ and from (21), we have

$$\Phi_\Delta(t,\sigma)A_\Delta(\sigma)x = \Phi(t,\sigma)A_\Delta(\sigma)x + \int_\sigma^t z^t(\tau,\sigma)d\tau, \tag{54}$$

where $z^t(\tau,\sigma)$ is defined as in (49) with $z(\sigma) = A_\Delta(\sigma)x$, namely

$$z^t(\tau,\sigma) = \Phi(t,\tau)D(\tau)\Delta(\tau)E(\tau)\Phi_\Delta(\tau,\sigma)A_\Delta(\sigma)x. \tag{55}$$

Hence, for every $t \geq s$,

$$\begin{aligned}
\int_s^t \Phi_\Delta(t,\sigma)A_\Delta(\sigma)x\,d\sigma = \int_s^t \Phi(t,\sigma)A(\sigma)x\,d\sigma + \int_s^t \Phi(t,\sigma)D(\sigma)\Delta(\sigma)E(\sigma)x\,d\sigma \\
+ \int_s^t\int_\sigma^t z^t(\tau,\sigma)d\tau\,d\sigma.
\end{aligned}$$

Using (51) and the fact that $\Phi(\cdot,\cdot)$ is a weak evolution operator in X with generator $A(\cdot)$, we have

$$\int_s^t \Phi_\Delta(t,\sigma)A_\Delta(\sigma)x\,d\sigma + x = \Phi(t,s)x + \int_s^t \Phi(t,\tau)D(\tau)\Delta(\tau)E(\tau)x\,d\tau + \int_s^t \int_s^\tau z^t(\tau,\sigma)d\sigma\,d\tau \qquad t \geq s. \tag{56}$$

By (55) this means

$$\int_s^t \Phi_\Delta(t,\sigma)A_\Delta(\sigma)x\,d\sigma + x = \Phi(t,s)x + \int_s^t \Phi(t,\tau)D(\tau)\Delta(\tau) \cdot \left[E(\tau)x + \int_s^\tau E(\tau)\Phi_\Delta(\tau,\sigma)A_\Delta(\sigma)x\,d\sigma\right] d\tau. \tag{57}$$

Now Hypothesis 10 implies

$$\int_s^\tau E(\tau)\Phi(\tau,\sigma)A_\Delta(\sigma)x\,d\sigma = \int_s^\tau E(\tau)\Phi(\tau,\sigma)[A(\sigma)x + D(\sigma)\Delta(\sigma)E(\sigma)x]d\sigma = E(\tau)\int_s^\tau \Phi(\tau,\sigma)A_\Delta(\sigma)x\,d\sigma.$$

Therefore, we obtain from Lemma 4.5 and (54), (55),

(58)

$$\int_s^\tau E(\tau)\Phi_\Delta(\tau,\sigma)A_\Delta(\sigma)x\,d\sigma = \int_s^\tau E(\tau)\left[\Phi(\tau,\sigma)A_\Delta(\sigma)x + \int_\sigma^\tau z^\tau(\rho,\sigma)d\rho\right] d\sigma = E(\tau)\int_s^\tau \Phi_\Delta(\tau,\sigma)A_\Delta(\sigma)x\,d\sigma.$$

From (57) we conclude

$$\int_s^t \Phi_\Delta(t,\sigma)A_\Delta(\sigma)x\,d\sigma + x = \Phi(t,s)x + \int_s^t \Phi(t,\tau)D(\tau)\Delta(\tau)E(\tau) \cdot \left[x + \int_s^\tau \Phi_\Delta(\tau,\sigma)A_\Delta(\sigma)x\,d\sigma\right] d\tau. \tag{59}$$

Hence $\Psi_\Delta(t,s)x := x + \int_s^t \Phi_\Delta(t,\sigma)A_\Delta(\sigma)x\,d\sigma$ satisfies the same equation (21) as $\Phi_\Delta(t,s)x$ on $[s,\infty)$. Now by (56),

$$\Psi_\Delta(t,s)x = \Phi(t,s)x + \int_s^t \Phi(t,\tau)D(\tau)\Delta(\tau)E(\tau)x\,d\tau + \int_s^t \int_\sigma^t z^t(\tau,\sigma)d\tau\,d\sigma.$$

For $x \in D(A)$, the RHS lies in $\underline{X}$. In fact, this is the case for the first term by Hypotheses 6 and 10, for the second term by Hypothesis 9, and it was proved to be so for the last term in Lemma 4.5. Thus,

$$\hat{y}(t,s,x) := E(t)\Psi_\Delta(t,s)x = E(t)x + E(t)\int_s^t \Phi_\Delta(t,\sigma)A_\Delta(\sigma)x\,d\sigma$$

is well defined for every $x \in D(A)$, $t \geq s$. It follows from Lemma 4.4 and (58) that

$$\hat{y}(\cdot,s,x)|_{[s,t]} = \left[E(\cdot)x + \int_s^\cdot E(\cdot)\Phi_\Delta(\cdot,\sigma)A_\Delta(\sigma)x\,d\sigma\right]\Big|_{[s,t]} \in L^p(s,t;Y)$$

for every $t > s$ and from (59) that

$$\hat{y}(\cdot, s, x)|_{[s,t]} = y_0(\cdot, s, x)|_{[s,t]} + (\mathbb{L}_{s,t}\Delta_{s,t}\hat{y}(\cdot, s, x)|_{[s,t]})(\cdot)|_{[s,t]}, \qquad t > s,$$

where $y_0(\cdot, s, x)$ is defined by (25). Hence,

$$E(\cdot)\Psi_\Delta(\cdot, s)x = y_\Delta(\cdot, s, x),$$

and (59) implies

$$\Psi_\Delta(t, s)x = \Phi(t, s)x + \int_s^t \Phi(t, \tau)D(\tau)\Delta(\tau)y_\Delta(\tau, s, x)d\tau = \Phi_\Delta(t, s)x,\ x \in D(A).$$

Or, equivalently,

$$\int_s^t \Phi_\Delta(t, \sigma)A_\Delta(\sigma)x\, d\sigma + x = \Phi_\Delta(t, s)x, \qquad x \in D(A).$$

This completes the proof. □

To illustrate the assumptions of the theorem, we discuss them in a time-invariant setting for the so called Pritchard–Salamon class of systems. This class was introduced in [22] to study linear quadratic optimal control for infinite-dimensional systems with unbounded input and output operators.

DEFINITION 4.6. *A system $\Sigma = (S(\cdot), D, E)$ is said to be in Pritchard–Salamon class ($\Sigma \in \mathbb{PS}$) if the following conditions hold.*

(a) *$\underline{X}, X, U, Y$ are Hilbert spaces; $\underline{X} \hookrightarrow X$ with a continuous dense injection; $D \in \mathcal{L}(U, X)$, $E \in \mathcal{L}(\underline{X}, Y)$.*

(b) *$S(\cdot)$ is a C_0-semigroup of operators on $\underline{X}$, which extends to a semigroup on X (denoted by the same symbol).*

(c) *There exists $t_1 > 0$ and $k_1 > 0$ such that $\int_0^{t_1} S(t_1 - \sigma)Du(\sigma)d\sigma \in \underline{X}$ for all $u(\cdot) \in L^2(0, t_1; U)$ and*

$$\left\| \int_0^{t_1} S(t_1 - \sigma)Du(\sigma)d\sigma \right\|_{\underline{X}} \le k_1\|u(\cdot)\|_{L^2(0,t_1;U)}, \qquad u(\cdot) \in L^2(0, t_1; U). \tag{60}$$

(d) *There exists $t_2 > 0$ and $k_2 > 0$ such that*

$$\|ES(\cdot)x\|_{L^2(0,t_2;Y)} \le k_2\|x\|_X, \qquad x \in \underline{X}. \tag{61}$$

Remark 4.7. (i) If (60) (respectively, (61)) hold for some $t_1 > 0\,(t_2 > 0)$, it can be shown that they hold for all $t_1\,(t_2)$, where k_1, (k_2) depend on t_1, (t_2). Moreover, if $S(t)$ is exponentially stable on $\underline{X}$ and X, then the $k_1(t_1)$ (respectively, $k_2(t_2)$) may be chosen to be independent of t_1 (respectively, t_2) [27].

(ii) Because $S(\cdot)$ is a strongly continuous semigroup on X, $\Phi(t, s) = S(t-s)$ is a strong (and therefore a weak) evolution operator.

The properties of the class $\mathbb{PS}$ are chosen to guarantee that (in spite of the unboundedness of the operators D, E) bounded linear output feedbacks generate a C_0-semigroup on X. More precisely, let $\Sigma = (S(\cdot), D, E) \in \mathbb{PS}$, let A be the generator of $S(\cdot)$ with domain $D(A)$, and consider, for any $\Delta \in \mathcal{L}(Y, U)$, the operator $A + D\Delta E$ with the same domain. If $D(A) \subset \underline{X}$, it is shown in [2] that A_Δ generates a strongly continuous semigroup $S_\Delta(\cdot)$ on X.

Using Remark 4.7 it is easily verified that if $(S(\cdot), D, E) \in \mathbb{PS}$ is exponentially stable, then all the assumptions of Theorem 3.2 (with $\overline{X} = X$, $p = 2$) are automatically satisfied

except Hypothesis 8. In fact, condition (a) of Definition 4.6 implies Hypothesis 1, (b) implies Hypotheses 2 and 6, (c) implies Hypotheses 4 and 5, and (d) implies Hypothesis 7 (for exponentially stable $S(\cdot)$). If Hypothesis 8 is not satisfied, that is, the operator

$$\mathbb{L}_0 : L^2(0,\infty;U) \to L^2(0,\infty;Y), \qquad u(\cdot) \mapsto E\int_0^{\cdot} S(\cdot-\sigma)Du(\sigma)d\sigma$$

is unbounded, then the statement of Theorem 3.2 is void ($\|\mathbb{L}_0\| = \infty$). If, however, $\mathbb{L}_0$ is bounded and $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$ is a perturbation with norm $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0\|^{-1}$, then there exists by Theorem 3.2 a unique perturbed mild evolution operator $\Phi_\Delta(\cdot,\cdot)$ on X satisfying (19)–(21) and $\Phi_\Delta(\cdot,\cdot)$ is exponentially stable on X.

To relate Φ_Δ to the perturbed differential equation

$$\dot{x}(t) = [A + D\Delta(t)E]x(t), \qquad t \geq 0, \tag{62}$$

we have to consider the additional assumptions required for Theorem 4.2, that is, Hypotheses 9 and 10. Hypothesis 9 is a direct consequence of (c). Now if we assume the additional property that

$$D(A) \subset \underline{X}, \tag{63}$$

then the closed graph theorem implies that the canonical injection $i : D(A) \to \underline{X}$ is continuous, when $D(A)$ is provided with the graph norm. It was proved in [22] that the second equation in Hypothesis 10 holds, and more recently it has been shown in [2] that this is in fact the case for all systems in $\mathbb{PS}$ *without* the property (63). The following lemma shows that for any system in $\mathbb{PS}$ possessing the property (63), the first equation in Hypothesis 10 also holds.

LEMMA 4.8. *If* $(S(\cdot), D, E) \in \mathbb{PS}$ *satisfies* $D(A) \subset \underline{X}$, *then*

$$E\int_s^t S(t-\sigma)Ax\,d\sigma = \int_s^t ES(t-\sigma)Ax\,d\sigma, \qquad x \in D(A). \tag{64}$$

Proof. An easy argument shows that for every $x \in D(A)$ there exists a sequence (x_k) in $D(A)$ such that $Ax_k \in \underline{X}$, $\lim_{k\to\infty}\|x_k - x\|_{\underline{X}} = 0$ and $\lim_{k\to\infty}\|Ax_k - Ax\|_X = 0$. Because the restriction $S(\cdot)|_{\underline{X}}$ is a semigroup on $\underline{X}$, we have for all $k \in \mathbb{N}$

$$E\int_s^t S(t-\sigma)Ax_k\,d\sigma = \int_s^t ES(t-\sigma)Ax_k\,d\sigma$$

and

$$\begin{aligned}\lim_{k\to\infty} E\int_s^t S(t-\sigma)Ax_k\,d\sigma &= \lim_{k\to\infty} E[S(t-s)x_k - x_k] \\ &= E[S(t-s)x - x] = E\int_s^t S(t-\sigma)Ax\,d\sigma.\end{aligned}$$

On the other hand, we have by (61)

$$\begin{aligned}&\lim_{k\to\infty}\left\| \int_s^t ES(t-\sigma)Ax\,d\sigma - \int_s^t ES(t-\sigma)Ax_k\,d\sigma \right\|_Y \\ &\quad\leq (t-s)^{1/2}\lim_{k\to\infty}\|ES(\cdot)(Ax - Ax_k)\|_{L^2(0,t-s;Y)} \\ &\quad\leq (t-s)^{1/2}k_2\lim_{k\to\infty}\|Ax - Ax_k\|_X = 0.\end{aligned}$$

This proves (64). □

As a consequence we obtain the following corollary of Theorem 4.2.

COROLLARY 4.9. *Suppose* $(S(\cdot), D, E) \in \mathbb{PS}$ *is exponentially stable,* $D(A) \subset \underline{X}$, $\mathbb{L}_0$ *is bounded, and* $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$ *is a time-varying perturbation with norm* $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0\|^{-1}$. *Then the unique mild evolution operator* $\Phi_\Delta(\cdot,\cdot)$ *satisfying* (19)–(21) *is an exponentially stable weak evolution operator on* X *with generator* $A_\Delta(\cdot)$. *Moreover, if* $\Delta \in \mathcal{L}(Y,U)$ *is constant, then the perturbed system operator* $A + D\Delta E$ *with domain* $D(A)$ *is the infinitesimal generator of a strongly continuous semigroup* $S_\Delta(\cdot)$ *on* X, *such that* $\Phi_\Delta(t,s) = S_\Delta(t-s)$ *and, for every* $x \in D(A)$, *the function* $x_\Delta(t) = S_\Delta(t)x$ *is a strong solution of the perturbed system equation* (62) *with* $x_\Delta(0) = x$. Φ_Δ *is a strong evolution operator and the Cauchy problem associated with* (62) *is well posed.*

Proof. It only remains to prove that the generator of $S_\Delta(\cdot)$ is $A + D\Delta E$ with domain $D(A)$ and this is carried out in [2]. □

Note that, by the counterexample of Phillips [21], we cannot, in general, expect to obtain a strong evolution operator Φ_Δ or strong solutions of the perturbed system equation if Δ is *time-varying*.

In [23], [24] the authors claim that if the conditions indicated in Remark 3.6 hold and if $\|\Delta\|_\infty < \|\mathbb{L}_0\|^{-1}$, then the perturbed semigroup $S_\Delta(\cdot)$ has the property that $x(t) = S_\Delta(t)x$, $x \in D(A)$ satisfies $\dot{x}(t) = (A + D\Delta E)x(t)$, $t \geq 0$. However, the proof is rather obscure and probably requires an additional condition like (60).

We now illustrate the restrictions imposed by the various hypotheses (in this and the previous section) with the following time-invariant example. In particular, we see that the assumptions that guarantee the perturbed evolution operator is exponentially stable and weak are far more restrictive than those that guarantee it is exponentially stable and mild.

Example 4.10. Suppose $-A$ generates an exponentially stable analytic semigroup $S(\cdot)$ on a Hilbert space H (with norm $\|\cdot\|$). Then for any $\gamma \in \mathbb{R}$, A^γ can be defined as in [20]. For α (respectively, β) $\in [0,1]$, we denote the domain of A^α by H_α and endow it with the associated graph norm, $\|\cdot\|_\alpha$ (respectively, the range of $A^{-\beta}$ by $H_{-\beta}$ and for $y = A^{-\beta}x$ we set $\|y\|_{-\beta} = \|x\|$). Then $S(\cdot)$ restricts (extends) to an exponentially stable strongly continuous semigroup on the Hilbert space H_α (respectively, $H_{-\beta}$). Moreover, there exist constants M, $\omega > 0$ such that for all $t > 0$ and $x \in H$ (respectively, $x \in H_{-\beta}$),

$$\|S(t)x\|_\alpha \leq \frac{Me^{-\omega t}}{t^\alpha}\|x\|, \qquad \left(\text{respectively, } \|S(t)x\| \leq \frac{Me^{-\omega t}}{t^\beta}\|x\|_{-\beta}\right),$$

where $\|\cdot\|$ is the norm on H [20]. (Note that for $t > 0$, $S(t)$ maps H into H_α and $H_{-\beta}$ into H.)

Now suppose that $E \in \mathcal{L}(H_{\alpha(E)}, H)$ and $D \in \mathcal{L}(H, H_{-\beta(D)})$ for some $\alpha(E), \beta(D) \in [0,1]$. Then, for arbitrary $\beta \in [0,1]$ and $x \in H_{\alpha(E)}$,

$$\begin{aligned}
\|ES(\cdot)x\|_{L^2(0,\infty;H)} &\leq \|E\|_{\mathcal{L}(H_{\alpha(E)},H)} \left[\int_0^\infty \|S(\rho/2)S(\rho/2)x\|^2_{\alpha(E)} d\rho\right]^{1/2} \\
&\leq \|E\|_{\mathcal{L}(H_{\alpha(E)},H)} \left[\int_0^\infty \frac{M^2 e^{-\omega\rho}}{(\rho/2)^{2\alpha(E)}}\|S(\rho/2)x\|^2 d\rho\right]^{1/2} \qquad (65) \\
&\leq \|E\|_{\mathcal{L}(H_{\alpha(E)},H)} M^2 \left[\int_0^\infty \frac{e^{-2\omega\rho}}{(\rho/2)^{2(\alpha(E)+\beta)}}\, d\rho\right]^{1/2} \|x\|_{-\beta}.
\end{aligned}$$

For arbitrary $\alpha \in [0,1]$, $u(\cdot) \in L^2(0,t;H)$,

$$\begin{aligned}\|(\mathbb{M}_0 u)(t)\|_\alpha &\le \int_0^t \|S((t-\rho)/2)S((t-\rho)/2)Du(\rho)\|_\alpha d\rho \\ &\le \int_0^t \frac{M^2 e^{-\omega(t-\rho)}}{((t-\rho)/2)^{\alpha+\beta(D)}}\|D\|_{\mathcal{L}(H,H_{-\beta(D)})}\|u(\rho)\|d\rho \qquad (66)\\ &\le M^2\|D\|_{\mathcal{L}(H,H_{-\beta(D)})}\left[\int_0^t \frac{e^{-2\omega\rho}}{(\rho/2)^{2(\alpha+\beta(D))}}\,d\rho\right]^{1/2}\|u(\cdot)\|_{L^2}.\end{aligned}$$

Finally, for arbitrary $\alpha \in [0,1]$, $u(\cdot) \in L^2(0,\infty;H)$,

$$\begin{aligned}\|(\mathbb{M}_0 u)(\cdot)\|^2_{L^2(0,\infty;H_\alpha)} &\le \int_0^\infty \left(\int_0^t \|S(t-\rho)Du(\rho)\|_\alpha\,d\rho\right)^2 dt \qquad (67)\\ &\le M^4\|D\|^2_{\mathcal{L}(H,H_{-\beta(D)})}\int_0^\infty\left(\int_0^t \frac{e^{-\omega(t-\rho)}\|u(\rho)\|d\rho}{((t-\rho)/2)^{\alpha+\beta(D)}}\right)^2 dt.\end{aligned}$$

To apply Theorem 3.5 we set

$$X = H, \quad \underline{X} = H_{\alpha(E)}, \quad \overline{X} = H_{-\beta(D)}, \quad U = Y = H.$$

Then (a) of Theorem 3.5 is automatically satisfied. If $\alpha(E) < \frac{1}{2}$, $\beta(D) < \frac{1}{2}$ and $p = 2$, we see from (66) (with $\alpha = 0$) that (b) is satisfied and from (65) (with $\beta = 0$) that (c) is satisfied. Using the convolution estimate $\|f*g\|_{L^2(0,\infty)} \le \|f\|_{L^1(0,\infty)}\|g\|_{L^2(0,\infty)}$ in (67) (with $\alpha = \alpha(E)$), we see that (d) holds provided $\alpha(E)+\beta(D) < 1$. Finally, $D(A^{H_{-\beta(D)}}) \subset H_{\alpha(E)}$ and hence (e) holds if $\alpha(E) + \beta(D) \le 1$ [20]. In summary, the conditions of Theorem 3.5 are satisfied with $p = 2$ provided $\alpha(E) < \frac{1}{2}$, $\beta(D) < \frac{1}{2}$. Hence, under these conditions, $r(S(\cdot); D, E) \ge \|\mathbb{L}_0\|^{-1}$ with equality in the case of a complex perturbation.

Now suppose

$$X = H_{-\beta(D)}, \quad \underline{X} = H_{\alpha(E)}, \quad U = Y = H.$$

Then (a) and (b) of Definition 4.6 are automatically satisfied. From (66) and (65) we see that (c) and (d) are satisfied if $\alpha(E) + \beta(D) < \frac{1}{2}$. Hence for $\alpha(E) + \beta(D) < \frac{1}{2}$ we may apply Corollary 4.9 to conclude that if $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0\|^{-1}$, then there exists an exponentially stable weak evolution operator $\Phi_\Delta(\cdot,\cdot)$ on $H_{-\beta(D)}$ with generator $-A + D\Delta(\cdot)E$. Moreover, if $\Delta \in \mathcal{L}(H)$ is constant, $-A + D\Delta E$ with domain $D(A)$ is the infinitesimal generator of the C_0-semigroup $S_\Delta(\cdot) = \Phi_\Delta(\cdot, 0)$ on $H_{-\beta(D)}$.

5. Multiperturbations. In this section we extend the results of the previous sections from the single to the multiperturbation case. But before introducing the formal definitions, we briefly review the effect of time-varying state transformations $\hat{x}(t) = T(t)^{-1}x(t)$ on mild evolution operators and stability radii.

To derive a formula for the transformed evolution operator, let us assume for a moment that Φ is generated by a family $A(\cdot)$ of bounded linear operators on X and $T(\cdot) \in PC^1(\mathbb{R}_+, \mathcal{U}(X))$. Then the above transformation converts the system (1) into

$$\dot{\hat{x}}(t) = \hat{A}(t)\hat{x}(t), \qquad t \ge 0, \qquad (68)$$

where

$$\hat{A}(t) = T(t)^{-1}A(t)T(t) - T(t)^{-1}\dot{T}(t), \qquad t \ge 0.$$

The evolution operator of the system (68) is

$$\hat{\Phi}(t,s) = T(t)^{-1}\Phi(t,s)T(s), \qquad t \geq s \geq 0. \tag{69}$$

The RHS of (69) makes sense for arbitrary mild evolution operators $\Phi(t,s)$ on X and arbitrary time-varying transformations $T(\cdot) \in C(\mathbb{R}_+, \mathcal{U}(X))$. In this way (69) defines a right group action of $C(\mathbb{R}_+, \mathcal{U}(X))$ on the set of mild evolution operators on X. We say that the mild evolution operator $\hat{\Phi}(t,s)$ defined by (69) is obtained from the mild evolution operator $\Phi(t,s)$ by application of the time-varying similarity transformation $T = (T(t))_{t\in\mathbb{R}_+}$.

Because these transformations will not, in general, preserve stability properties, additional assumptions must be imposed if we want to use them in stability analysis. The following class of transformations preserves exponential stability.

DEFINITION 5.1 (Bohl transformation). *$T(\cdot) \in C(\mathbb{R}_+, \mathcal{U}(X))$ is said to be a* Bohl transformation *if*

$$\inf\{\varepsilon \in \mathbb{R};\ \exists M_\varepsilon > 0\ \forall t,s \geq 0 : \|T(t)^{-1}\|\|T(s)\| \leq M_\varepsilon e^{\varepsilon|t-s|}\} = 0.$$

The next example characterizes scalar Bohl transformations and shows that every time-varying scalar system can be made *time-invariant* via a Bohl transformation.

Example 5.2. Suppose $\theta(\cdot) \in PC^1(\mathbb{R}_+, \mathbb{C}^*)$, $\mathbb{C}^* = \mathbb{C}\backslash\{0\}$ and let $a(t) = \dot{\theta}(t)/\theta(t)$, $t \geq 0$, so that $\dot{\theta}(t) = a(t)\theta(t)$. The fundamental solution of this differential equation is

$$\Phi(t,s) = \theta(t)\theta(s)^{-1}, \qquad s,t \in \mathbb{R}_+.$$

By Definition 5.1, $\theta(\cdot)$ is a Bohl transformation if and only if for every $\varepsilon > 0$ there exists $M_\varepsilon > 0$ such that, for $t \geq s \geq 0$,

$$M_\varepsilon^{-1} e^{-\varepsilon(t-s)} \leq |\Phi(s,t)^{-1}| = |\Phi(t,s)| \leq M_\varepsilon e^{\varepsilon(t-s)}.$$

This condition holds if and only if $a(\cdot)$ has *strict* Bohl exponent 0, that is, $\lim_{s,t-s\to\infty} \ln|\theta(t)\theta(s)^{-1}|/(t-s) = 0$.

It is easily verified [10] that every *scalar* system $\dot{x}(t) = A(t)x(t)$ that has a strict Bohl exponent β can be transformed via the Bohl transformation

$$\theta(t) = \exp\left(\int_0^t [A(\rho) - \beta]d\rho\right), \qquad t \geq 0$$

into the *time-invariant* linear system $\dot{\hat{x}}(t) = \beta\hat{x}(t)$, $t \geq 0$.

The following proposition summarizes some elementary properties of Bohl transformations (compare [10]).

PROPOSITION 5.3. (i) *The Bohl transformations form a group with respect to (pointwise) multiplication.*

(ii) *The Bohl exponent is invariant with respect to Bohl transformations.*

(iii) *Let Φ be a mild evolution operator on X. If $T(\cdot) \in C(\mathbb{R}_+, \mathcal{U}(X))$ is a Bohl transformation and $\hat{\Phi}$ is defined by* (69), *then*

$$r(\hat{\Phi}; T^{-1}D,\ ET) = r(\Phi; D, E). \tag{70}$$

(iv) *If $\theta(\cdot) \in C(\mathbb{R}_+, \mathbb{C}^*)$ is a scalar Bohl transformation and Φ^θ is defined by $\Phi^\theta(t,s) = \theta(t)^{-1}\Phi(t,s)\theta(s)$, then*

$$r(\Phi^\theta; D, E) = r(\Phi; D, E). \tag{71}$$

We omit the proof, which is straightforward. Note that, contrary to (70), the structure operators D, E are not transformed in (71). In general, the stability radius may vary dramatically if a *non-scalar* Bohl transformation $T(\cdot)$ (or even a constant transformation $T \in \mathcal{U}(X)$) is applied to Φ alone and not to (D, E). Indeed, if $\Phi(t,s) = e^{A(t-s)}$, where $A \in \mathbb{C}^{n\times n}$ is exponentially stable, it has been shown in [11] that $r(T^{-1}\Phi T; I_n, I_n)$ varies from 0 to $\beta(\Phi) = -\sup_{\lambda\in\sigma(A)} \operatorname{Re}\lambda$ as T varies through $Gl_n(\mathbb{C})$.

Let us now turn to *multiperturbations*. On the level of system equations these are additive perturbations of the generator $A(\cdot) \rightsquigarrow A(\cdot) + \sum_{i=1}^N D_i(\cdot)\Delta_i(\cdot)E_i(\cdot)$, leading to the perturbed system equation

$$\dot{x}(t) = A(t)x(t) + \sum_{i=1}^{N} D_i(t)\Delta_i(t)E_i(t)x(t), \qquad t \geq 0. \tag{72}$$

Here $\Delta_i(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y_i, U_i))$, $i \in \underline{N}$ are unknown *bounded* time-varying disturbance operators; Y_i, U_i are Banach spaces over $\mathbb{K}$; and $D_i(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(U_i, \overline{X}))$, $E_i(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(\underline{X}, Y_i))$, $i \in \underline{N}$ are given operator-valued functions that describe the structure and unboundedness of the perturbation. In the previous sections, we only considered the *single perturbation* case $N = 1$. Setting $Y = \oplus_1^N Y_i$, $U = \oplus_1^N U_i$ (with norms $\|(y)\|_Y = (\sum_1^N \|y_i\|_{Y_i}^2)^{1/2}$, $\|(u)\|_U = (\sum_1^N \|u_i\|_{U_i}^2)^{1/2}$), the size of the overall disturbance operator

$$\Delta(\cdot) = \bigoplus_1^N \Delta_i(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U)) \tag{73}$$

is measured by

$$\|\Delta(\cdot)\|_\infty = \sup_{t\geq 0} \max_{i\in\underline{N}} \|\Delta_i(t)\|_{\mathcal{L}(Y_i,U_i)}. \tag{74}$$

The effect of multiperturbations on mild evolution operators is most easily described by writing multiperturbations in the form of a single perturbation. For this let

$$D(t) = [D_1(t), \dots, D_N(t)] : U \to \overline{X}; \quad (u_i)_{i=1}^N \mapsto \sum_{i=1}^N D_i(t)u_i, \quad (u_i) \in U,\ t \in \mathbb{R}_+;$$

$$E(t) = \begin{bmatrix} E_1(t) \\ \vdots \\ E_N(t) \end{bmatrix} : \underline{X} \to Y; \quad x \mapsto \begin{bmatrix} E_1(t)x \\ \vdots \\ E_N(t)x \end{bmatrix}, \quad x \in \underline{X}, \quad t \in \mathbb{R}_+. \tag{75}$$

Then $D(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(U, \overline{X}))$, $E(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(\underline{X}, Y))$, and the perturbed equation (72) takes the same form as (3), namely

$$\dot{x}(t) = [A(t) + D(t)\Delta(t)E(t)]x(t), \qquad t \geq 0.$$

As a consequence we can apply all the concepts introduced in §2 to multiperturbations. An adequate definition of stability radius for multiperturbations is obtained by restricting the perturbations in (23) to be of the form (73).

DEFINITION 5.4. *Suppose Φ is exponentially stable, $D(\cdot), E(\cdot)$ are given by* (75) *and Hypotheses* 1–4 *hold. The* stability radius *of $\Phi(\cdot,\cdot)$ with respect to the perturbation structure $(D_i, E_i)_{i\in\underline{N}}$ is defined by*

$$r(\Phi; (D_i, E_i)) = \sup\{r \in \mathbb{R}_+; \forall \Delta_i(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y_i, U_i)) :$$
$$\left\| \bigoplus_1^N \Delta_i \right\|_\infty \leq r \Rightarrow \Delta = \bigoplus_1^N \Delta_i \text{ is admissible and } \Phi_\Delta(\cdot,\cdot) \text{ is exponentially stable}\}. \tag{76}$$

In (76) only perturbations $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$ are considered, which have the block structure $\Delta(\cdot) = \oplus_1^N \Delta_i(\cdot)$ given by the families of Banach spaces $(U_i)_{i\in\underline{N}}$, $(Y_i)_{i\in\underline{N}}$. But the norm $\|\Delta(\cdot)\|_\infty$ defined in (74) is just the L^∞-norm of $\Delta(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y,U))$, where $\mathcal{L}(Y,U)$ is provided with the operator norm corresponding to the norms $\|(u)\|_U$, $\|(y)\|_Y$ defined above. Hence the perturbation norms applied in Definitions 5.4 and 2.17 are the same so that we obtain

$$r(\Phi; (D_i, E_i)) \geq r(\Phi; D, E), \tag{77}$$

where D, E are defined by (75).

We now examine the following effect of rescaling the structure operators D_i, E_i, $i \in \underline{N}$:

$$D_i(t) \rightsquigarrow D_i^\alpha(t) = \alpha_i^{-1}(t) D_i(t) \quad \text{and} \quad E_i(t) \rightsquigarrow E_i^\alpha(t) = \alpha_i(t) E_i(t), \qquad i \in \underline{N}, \tag{78}$$

where $\alpha_i(t) \in PC(\mathbb{R}_+, \mathbb{C}^*)$, $i \in \underline{N}$. Interpreting D_i, E_i, $i \in \underline{N}$ as input and output operators, the transformations (78) represent scalar input and output transformations of the system. We set

$$D^\alpha(t) = [D_1(t)\alpha_1^{-1}(t), \ldots, D_N(t)\alpha_N^{-1}(t)], \qquad E^\alpha(t) = \begin{bmatrix} \alpha_1(t)E_1(t) \\ \vdots \\ \alpha_N(t)E_N(t) \end{bmatrix}, \qquad t \geq 0, \tag{79}$$

where

$$\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_N(\cdot)) \in \mathcal{D}_N := PC(\mathbb{R}_+, \mathbb{C}^*)^N.$$

Because

$$D(t)\Delta(t)E(t) = \sum_1^N D_i(t)\Delta_i(t)E_i(t) = \sum_1^N D_i^\alpha(t)\Delta_i(t)E_i^\alpha(t) = D^\alpha(t)\Delta(t)E^\alpha(t),$$

(72) and (21) have the same solution if D, E are replaced by D^α, E^α. As a consequence we obtain

$$r(\Phi, (D_i, E_i)) = r(\Phi, (D_i^\alpha, E_i^\alpha)), \qquad \alpha(\cdot) \in \mathcal{D}_N. \tag{80}$$

Now suppose that $\theta \in C(\mathbb{R}_+, \mathbb{C}^*)$ is a Bohl transformation, $\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_N(\cdot)) \in \mathcal{D}_N$, D^α, E^α are defined by (79), and Φ^θ is as in Proposition 5.3. We denote by $\mathbb{L}_s^{\alpha,\theta}$, $\mathbb{M}_s^{\alpha,\theta}$ the operators given by (25), (16) with (Φ, D, E) replaced by $(\Phi^\theta, D^\alpha, E^\alpha)$, assuming that these operators are well defined. For example,

$$\begin{aligned} \mathbb{L}_s^{\alpha,\theta} &= \big((\mathbb{L}_s^{\alpha,\theta})_{ij}\big)_{i,j\in\underline{N}}, \\ \big((\mathbb{L}_s^{\alpha,\theta})_{ij}u\big)(t) &= \alpha_i(t)E_i(t)\int_s^t \theta(t)^{-1}\Phi(t,\rho)\theta(\rho)D_j(\rho)\alpha_j^{-1}(\rho)u(\rho)d\rho, \\ & \qquad\qquad t \geq s, \qquad i,j \in \underline{N}. \end{aligned} \tag{81}$$

As a consequence of Theorems 3.2, 4.2, and (71), (80), (77) we have the following corollaries.

COROLLARY 5.5. *Let $\theta \in C(\mathbb{R}_+, \mathbb{C}^*)$ be a Bohl transformation, $\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_N(\cdot)) \in \mathcal{D}_N$. If Hypotheses 1–8 hold for $(\Phi^\theta, D^\alpha, E^\alpha)$ and Φ is exponentially stable, then*

$$r(\Phi; (D_i, E_i)) \geq \|\mathbb{L}_0^{\alpha,\theta}\|^{-1}, \tag{82}$$

where $\mathbb{L}_0^{\alpha,\theta}$ *is defined by* (81) *and* Φ^θ *is as in Proposition* 5.3.

COROLLARY 5.6. *Let* $\theta \in C(\mathbb{R}_+, \mathbb{C}^*)$ *be a Bohl transformation,* $\alpha(\cdot) = (\alpha_1(\cdot), \dots, \alpha_N(\cdot)) \in \mathcal{D}_N$. *If Hypotheses* 1–10 *hold for* $(\Phi^\theta, D^\alpha, E^\alpha)$ *with* $\overline{X} = X$, Φ *is an exponentially stable weak evolution operator on* X *with generator* A, *and* $\|\Delta(\cdot)\|_\infty < \|\mathbb{L}_0^{\alpha,\theta}\|^{-1}$, *then the perturbed mild evolution operator* Φ_Δ *is an exponentially stable weak evolution operator on* X *with generator* A_Δ.

Corollary 5.5 tightens the lower bound in (27) for the single perturbation case. Indeed, it has been shown in [10] for the scalar case that the lower bound (27) may not equal the stability radius, whereas if $a(\cdot)$ has a strict Bohl exponent it is always possible to find a Bohl transformation that increases the lower bound (82) to the stability radius. In the time-invariant finite-dimensional multiperturbation case, it is known that if $N \leq 3$, equality holds in (82); however, this may not be so if $N > 3$.

Because in (81) the effect of θ can be subsumed into the α_i's by means of the transformation $\alpha_i \mapsto \alpha_i\theta^{-1}$, $i \in \underline{N}$, it would seem that there is no loss in generality in setting $\theta(t) \equiv 1$, $t \geq 0$. But this is not the case because the transformation $\alpha_i \mapsto \alpha_i\theta^{-1}$ changes Hypotheses 5–7. For example, Hypothesis 5 for the operator $\mathbb{M}_s^{\alpha,\theta}$ is not equivalent to Hypothesis 5 for the operator $\mathbb{M}_s^{\alpha\theta^{-1},1}$. Indeed the scaling of the evolution operator Φ and the structure operators (D_i, E_i) does not only open up the possibility of tightening the lower bound in (27), but may also extend the applicability of Theorem 3.2 and Theorem 4.2. This is because if Hypotheses 1–10 do not hold for the original data, it may be their validity can be enforced by scaling (see Example 6.1). On the other hand, supposing that Hypotheses 1–10 are satisfied for (Φ, D, E), then they will also be satisfied for $(\Phi^\theta, D^\alpha, E^\alpha)$, if the scaling functions $\alpha_i(\cdot), \alpha_i(\cdot)^{-1}, \theta(\cdot), \theta(\cdot)^{-1}$ are all bounded on $\mathbb{R}_+$.

6. Examples. In this final section we consider three examples. The first one is time-varying and illustrates that scaling transformations can be used to extend the applicability of the results as well as (possibly) improving the lower bound.

Example 6.1. Consider the time-varying system

$$\dot{x}(t) = -tAx(t), \qquad t \geq 0, \tag{83}$$

where A is a bounded linear operator on a Hilbert space H. We assume that $-A$ is exponentially stable, that is, there exist constants $M, \omega_0 > 0$ such that $\|e^{-At}\| \leq Me^{-\omega_0 t}$, $t \geq 0$. The linear operators $A(t) = -tA \in \mathcal{L}(H)$ are bounded for each $t \geq 0$, but their norm $\|A(t)\|$ is unbounded in time. The strong evolution operator generated by $A(\cdot)$,

$$\Phi(t, s) = e^{-A(t^2-s^2)/2}, \qquad t \geq s \geq 0,$$

satisfies $\|\Phi(t, s)\| \leq Me^{-\omega_0(t^2-s^2)/2}$, $t \geq s \geq 0$.

We assume that (83) is a nominal model that is subjected to perturbations of the form $A(t) \rightsquigarrow A_\Delta(t) = -t(A + \Delta)$, where $\Delta \in \mathcal{L}(H)$. This can be catered for by the perturbed system (72) in a number of different ways, for example, setting $\underline{X} = X = \overline{X} = U = Y = H$ and

(i) $D(t) = t^{1/2}I$, $E(t) = t^{1/2}I$ or (ii) $D(t) = I$, $E(t) = tI$.

Clearly, Hypotheses 1–4 are satisfied for both of the above perturbation structures. Now in case (i),

$$\begin{aligned} \|E(\cdot)\Phi(\cdot, s)x\|_{L^2}^2 &= \int_s^\infty \rho\|e^{-A(\rho^2-s^2)/2}x\|_H^2 \, d\rho \\ &\leq M^2 \int_s^\infty \rho e^{-\omega_0(\rho^2-s^2)} \, d\rho \|x\|_H^2 = \frac{M^2\|x\|_H^2}{(2\omega_0)} \end{aligned} \tag{84}$$

and hence Hypothesis 7 holds with $p = 2$. For the second case,

$$\begin{aligned}\|E(\cdot)\Phi(\cdot,s)x\|_{L^2}^2 &= \int_s^\infty \rho^2 \|e^{-A(\rho^2-s^2)/2}x\|_H^2 \, d\rho \\ &\geq s\int_s^\infty \rho \|e^{-A(\rho^2-s^2)/2}x\|_H^2 \, d\rho = \frac{s}{2}\int_0^\infty \|e^{-A\sigma}x\|_H^2 \, d\sigma.\end{aligned} \tag{85}$$

So it is not possible to find a constant k such that Hypothesis 7 holds with $p = 2$. However, note that the scaling factor $\alpha(t) = t^{-1/2}$, $t > 0$ transforms the second structure into the first. This demonstrates the point made after Corollary 5.5—scaling can extend the applicability of Theorem 3.2. Another possibility would be to take $p = 1$. An easy calculation shows that Hypothesis 7 holds for $p = 1$ in case (ii). We do not pursue this but continue our development for the first case with $p = 2$. The associated input–output operator is given by

$$(\mathbb{L}_0 u)(t) = t^{1/2}\int_0^t e^{-A(t^2-s^2)/2} s^{1/2} u(s) ds. \tag{86}$$

It is easy to see that Hypothesis 5 is satisfied. To get a good estimate for $\|\mathbb{L}_0\|^{-1}$, we may use the Riccati equation (40) as described in Remark 3.4(iii). A short calculation yields $\|\mathbb{L}_0\| \leq \max_{\omega\in\mathbb{R}} \|(i\omega I + A)^{-1}\|$. Hence (83) will be exponentially stable if

$$\|\Delta\|_{\mathcal{L}(H)} < [\max_{\omega\in\mathbb{R}} \|(i\omega I + A)^{-1}\|]^{-1}.$$

Because $\dot{x}(t) = -t(A+\Delta)x(t)$ is exponentially stable if and only if the time-invariant system $\dot{x}(t) == -(A + \Delta)x(t)$ is exponentially stable, it follows easily that $[\max_{\omega\in\mathbb{R}} \|(i\omega I + A)^{-1}\|]^{-1}$ is in fact the stability radius of (83) with respect to time-varying perturbations with the structure $(t^{1/2}I, t^{1/2}I)$.

In concrete applications the general results contained in Example 4.10 can be sharpened. This is illustrated by the following example.

Example 6.2. Consider a thermal process modeled by an equation of the form

$$\begin{aligned}&\frac{\partial T}{\partial t}(\xi,t) = \sum_{i=1}^3 \frac{\partial}{\partial \xi_i}\left[k(\xi)\frac{\partial T}{\partial \xi_i}(\xi,t)\right], \qquad \xi \in \Omega,\ t \geq 0, \\ &T(\xi,t) = 0, \quad \xi \in \partial\Omega, \quad t \geq 0, \quad T(\xi,0) = T_0(\xi), \quad \xi \in \Omega, \quad T_0(\cdot) \in L^2(\Omega),\end{aligned} \tag{87}$$

where $\emptyset \neq \Omega \subset \mathbb{R}^3$ is a bounded connected open set with smooth boundary $\partial\Omega$. $T(\xi,t)$ is the temperature at point $\xi \in \Omega$ and time t, $k(\xi)$ is the conductivity. Lions [18] has developed a general existence theory for partial differential equations of the type given by (87). A bilinear form on the Hilbert space $V = H_0^1(\Omega)$ is associated with (87) and if $k(\cdot) \in L^\infty(\Omega)$, $k(\xi) \geq \underline{k} > 0$ almost everywhere, the form is used to obtain an operator $A \in \mathcal{L}(V, V^*)$, where $V^* = H^{-1}(\Omega)$. The operator A with domain $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$ generates a strongly continuous semigroup $S(t)$ on $L^2(\Omega)$ and for sufficiently smooth initial data $T_0(\cdot)$, $T(\cdot,t) = S(t)T_0$ is a classical solution of (87). We will use Lions' framework throughout this example without further reference.

Now suppose that the conductivity $k(\xi)$ is uncertain with nominal value k_0 (a space-independent positive constant). The nominal system can be written in the form $\dot{x}(t) = k_0 A_0 x(t)$, $t \geq 0$, where $A_0 = \nabla^2$ (the Laplacian with zero boundary condition). For simplicity, we assume that on $X = L^2(\Omega)$, ∇^2 with zero Dirichlet boundary conditions has simple eigenvalues $-\lambda_1 > -\lambda_2 > \cdots$ with $\lambda_1 > 0$ and $\Psi_n \in H^2(\Omega) \cap H_0^1(\Omega)$, $n \in \mathbb{N}$ is an

orthonormal basis of eigenfunctions satisfying $\nabla^2\Psi_n = -\lambda_n\Psi_n$ in Ω. Then

$$A_0x = -\sum_{n=1}^{\infty}\lambda_n\Psi_n\langle\Psi_n, x\rangle,$$

$$D(A_0) = \left\{x \in L^2(\Omega); \sum_{n=1}^{\infty}\lambda_n^2\langle\Psi_n, x\rangle^2 < \infty\right\} = H^2(\Omega)\cap H_0^1(\Omega).$$

k_0A_0 generates an analytic semigroup $S_0(t)$ on $L^2(\Omega)$ where

$$S_0(t)x = \sum_{n=1}^{\infty}e^{-k_0\lambda_n t}\Psi_n\langle\Psi_n, x\rangle,$$

and hence

$$\|S_0(t)\| \le e^{-k_0\lambda_1 t}, \qquad t \ge 0.$$

The perturbations of the nominal system take the form

$$(B_kT)(\xi) = \sum_{i=1}^{3}\frac{\partial}{\partial\xi_i}\left[(k(\xi) - k_0)\frac{\partial T}{\partial\xi_i}(\xi)\right], \qquad \xi \in H^2(\Omega)\cap H_0^1(\Omega). \tag{88}$$

However, this specific perturbation structure cannot be catered for by our model (4). We have to consider a more general structure. A natural one is $D(t) = A_0^{1/2}$, $E(t) = A_0^{1/2}$, $Y = U = L^2(\Omega)$,

$$A_0^{1/2}x = \sum_{n=1}^{\infty}\lambda_n^{1/2}\Psi_n\langle\Psi_n, x\rangle,$$

$$\underline{X} = V = \left\{x \in L^2(\Omega); \sum_{n=1}^{\infty}\lambda_n\langle\Psi_n, x\rangle^2 < \infty\right\} = H_0^1(\Omega),$$

$$\overline{X} = V^* = \left\{x \in L^2(\Omega); \sum_{n=1}^{\infty}\lambda_n^{-1}\langle\Psi_n, x\rangle^2 < \infty\right\} = H^{-1}(\Omega).$$

Note that this structure corresponds, in the context of Example 4.10, to the case $\alpha(D) = \frac{1}{2}$, $\beta(E) = \frac{1}{2}$, which is not (quite) allowed in the result that establishes conditions $\alpha(D) < \frac{1}{2}$, $\beta(E) < \frac{1}{2}$ for the applicability of Theorem 3.5. Nevertheless, we see that the assumptions of this theorem are satisfied. Clearly, (a) holds. If $x \in L^2(\Omega)$, $x = \sum_{n=1}^{\infty}x_n\Psi_n$, $\|x\|_{L^2(\Omega)} = 1$,

$$\|ES_0(\cdot)x\|^2_{L^2(0,\infty;L^2(\Omega))} = \int_0^{\infty}\left\|\sum_{n=1}^{\infty}\lambda_n^{1/2}e^{-k_0\lambda_n t}x_n\Psi_n\right\|^2_{L^2(\Omega)}dt,$$

$$= \int_0^{\infty}\sum_{n=1}^{\infty}\lambda_n e^{-2k_0\lambda_n t}x_n^2\,dt = \frac{1}{2k_0}\sum_{n=1}^{\infty}x_n^2 = \frac{1}{2k_0}.$$

For $u(\cdot) \in L^2(0,\infty;L^2(\Omega))$, $\|u(\cdot)\|_{L^2} = 1$, let $u(t) = \sum_{n=1}^{\infty}u_n(t)\Psi_n$. Then

$$\left\|\int_0^t S_0(t-\rho)Du(\rho)d\rho\right\|^2_{L^2(\Omega)} = \sum_{n=1}^{\infty}\left(\int_0^t\lambda_n^{1/2}e^{-k_0\lambda_n(t-\rho)}u_n(\rho)d\rho\right)^2$$

$$\le \frac{1}{2k_0}\sum_{n=1}^{\infty}(1 - e^{-2k_0\lambda_n t})\int_0^t u_n^2(\rho)d\rho \le \frac{1}{2k_0}.$$

Also

$$\left\| \int_0^{\cdot} S_0(\cdot - \rho) Du(\rho) d\rho \right\|^2_{L^2(0,\infty;V)} = \sum_{n=1}^{\infty} \lambda_n^2 \int_0^{\infty} \left(\int_0^t e^{-k_0 \lambda_n (t-\rho)} u_n(\rho) d\rho \right)^2 dt.$$

But by a well-known convolution inequality [6],

$$\int_0^{\infty} \left(\int_0^t e^{-k_0 \lambda_n (t-\rho)} u_n(\rho) d\rho \right)^2 dt \le \frac{1}{k_0^2 \lambda_n^2} \int_0^{\infty} u_n^2(\rho) d\rho.$$

Hence,

$$\|\mathbb{L}_0 u(\cdot)\|^2_{L^2(0,\infty;L^2(\Omega))} = \|(\mathbb{M}_0 u)(\cdot)\|^2_{L^2(0,\infty;V)} \le \frac{1}{k_0^2} \sum_{n=1}^{\infty} \int_0^{\infty} u_n^2(\rho) d\rho = \frac{1}{k_0^2}.$$

The above estimates prove that (b)–(d) of Theorem 3.5 hold with $p = 2$ and $\|\mathbb{L}_0\| \le k_0^{-1}$. We are therefore able to conclude that $r(S_0(\cdot), D, E) \ge k_0$. For the time-invariant case we have proved that if $\Delta \in \mathcal{L}(L^2(\Omega))$, $\|\Delta\| < k_0$, then the perturbed semigroup $S_\Delta(\cdot)$ is exponentially stable. This is quite a strong result because for arbitrary $B \in \mathcal{L}(V, V^*)$, it is possible to find a $\Delta \in \mathcal{L}(L^2(\Omega))$ such that $B = A_0^{1/2} \Delta A_0^{1/2}$. Moreover, the result compares well with others in the literature. For example, in [19] it was shown that if A generates a contraction semigroup on a reflexive Banach space X, B is accretive, $D(B) \subset D(A)$ and for some $a \ge 0$,

$$\|Bx\|_X \le a\|x\|_X + \|Ax\|_X, \qquad x \in D(A). \tag{89}$$

Then the closure of $A - B$ generates a contraction semigroup. Equation (89) is slightly weaker than our requirement that $\|\Delta\| < k_0$. However, we do not assume that the perturbation is accretive, and our conclusion is different in that $S_\Delta(\cdot)$ is not necessarily a contraction semigroup but it is exponentially stable. Also our result is valid for time-varying $\Delta(\cdot)$.

It is easy to see that the operator in (88) satisfies $B_k \in \mathcal{L}(V, V^*)$ and $\|B_k\|_{\mathcal{L}(V,V^*)} \le \|k(\cdot) - k_0\|_{L^\infty(\Omega)}$. So the mild solution of (87) will be exponentially stable if we have $\|k(\cdot) - k_0\|_{L^\infty(\Omega)} < k_0$. This estimate is tight because if $k(\cdot) \equiv 0$, (87) is not exponentially stable. Hence, $r(S_0(\cdot), D, E) = r^c(S_0(\cdot), D, E) = \|\mathbb{L}_0\|^{-1} = k_0$. Note that the first two equalities could have been inferred directly from Theorem 3.5 because assumption (e) is also satisfied.

In the above examples we carefully matched the unboundedness in the structure operators with the unboundedness in $A(\cdot)$. Robust stability cannot be expected for structured perturbations whose unboundedness surpasses that of the nominal system generator. For example, it would not be sensible to introduce structure operators $D(t)$, $E(t)$ in Example 6.1, which are unbounded operators on H. Similarly, we should not introduce operators $D(t) \in \mathcal{L}(U, \overline{X})$, $E(t) \in \mathcal{L}(\underline{X}, Y)$ in Example 6.2, which are unbounded in time.

We conclude with an example of an interconnected system with uncertain couplings illustrating the effect of scaling in a multiperturbation problem.

Example 6.3. Suppose that we have two control systems both of which are uncertain, where the uncertainty can be modeled by a single perturbation structure. We write them formally as

$$\begin{aligned} \dot{x}_i(t) &= A_i(t)x_i(t) + D_i(t)\Delta_i(t)E_i(t)x_i(t) + B_i v_i(t), \qquad t \ge 0,\ i = 1, 2, \\ z_i(t) &= C_i(t)x_i(t), \end{aligned}$$

where $v_i(\cdot)$, $z_i(\cdot)$, $i = 1, 2$ represent the inputs and outputs. Now let us assume that the two systems are coupled via uncertain couplings $v_1(t) = K_1(t)z_2(t)$, $v_2(t) = K_2(t)z_1(t)$ so that the overall system takes the form

$$\text{(90)} \quad \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} A_1(t) + D_1(t)\Delta_1(t)E_1(t) & B_1(t)K_1(t)C_2(t) \\ B_2(t)K_2(t)C_1(t) & A_2(t) + D_2(t)\Delta_2(t)E_2(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}.$$

Or

$$\text{(91)} \quad \dot{x}(t) = A(t)x(t) + D(t)\Delta(t)E(t)x(t), \qquad t \geq 0,$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \quad A(t) = \begin{bmatrix} A_1(t) & 0 \\ 0 & A_2(t) \end{bmatrix},$$

$$D(t) = \begin{bmatrix} D_1(t) & B_1(t) & 0 & 0 \\ 0 & 0 & B_2(t) & D_2(t) \end{bmatrix}, \quad E(t) = \begin{bmatrix} E_1(t) & 0 \\ 0 & C_2(t) \\ C_1(t) & 0 \\ 0 & E_2(t) \end{bmatrix},$$

$$\Delta(t) = \begin{bmatrix} \Delta_1(t) & 0 & 0 & 0 \\ 0 & K_1(t) & 0 & 0 \\ 0 & 0 & K_2(t) & 0 \\ 0 & 0 & 0 & \Delta_2(t) \end{bmatrix}.$$

The nominal models are assumed to be given by exponentially stable mild evolution operators $\Phi_i(\cdot, \cdot)$ on Banach spaces X_i, $i = 1, 2$. $D_i(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(U_i, \overline{X}_i))$, $E_i(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(\underline{X}_i, Y_i))$, $B_i(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(V_i, \overline{X}_i))$, $C_i(\cdot) \in PC(\mathbb{R}_+, \mathcal{L}(\underline{X}_i, Z_i))$, $\Delta_i(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Y_i, U_i))$, $i = 1, 2$, $K_1(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Z_2, V_1))$, $K_2(\cdot) \in PC_b(\mathbb{R}_+, \mathcal{L}(Z_1, V_2))$, where $\underline{X}_i$, $\overline{X}_i$, Y_i, U_i, V_i, Z_i, $i = 1, 2$ are Banach spaces satisfying $\underline{X}_i \subset X_i \subset \overline{X}_i$, with continuous dense injections.

The problem is to obtain joint bounds on each of the unknowns $\Delta_i(\cdot)$, $K_i(\cdot)$, $i = 1, 2$, which guarantee the overall coupled uncertain system is exponentially stable. The mild perturbed system associated with (91) is described by the equation

$$x(t) = \Phi(t, s)x + \int_s^t \Phi(t, \rho)D(\rho)\Delta(\rho)E(\rho)x(\rho)d\rho, \qquad t \geq s \geq 0,$$

where

$$\Phi(t, s) = \begin{bmatrix} \Phi_1(t, s) & 0 \\ 0 & \Phi_2(t, s) \end{bmatrix}, \qquad t \geq s \geq 0.$$

Suppose that $(\Phi(\cdot, \cdot), D(\cdot), E(\cdot))$ satisfy Hypotheses 1–8 for $p = 2$, then $r(\Phi, D, E) \geq \|\mathbb{L}_0\|^{-1}$, and we seek an estimate for this lower bound. Now

$$(\mathbb{L}_0 u)(t) = \begin{bmatrix} (L_{11}u_1)(t) & (L_{12}v_1)(t) & 0 & 0 \\ 0 & 0 & (L_{23}v_2)(t) & (L_{24}u_2)(t) \\ (L_{31}u_1)(t) & (L_{32}v_1)(t) & 0 & 0 \\ 0 & 0 & (L_{43}v_2)(t) & (L_{44}u_2)(t) \end{bmatrix},$$

where $u(t) = [u_1(t),\, v_1(t),\, v_2(t),\, u_2(t)]^\top$, $u_i(\cdot) \in L^2(0,\infty;U_i)$, $v_i(\cdot) \in L^2(0,\infty;V_i)$, $i = 1,2$, and

(92)

$$(L_{11}u_1)(t) = E_1(t)\int_0^t \Phi_1(t,\rho)D_1(\rho)u_1(\rho)d\rho, (L_{12}v_1)(t) = E_1(t)\int_0^t \Phi_1(t,\rho)B_1(\rho)v_1(\rho)d\rho,$$

$$(L_{23}v_2)(t) = C_2(t)\int_0^t \Phi_2(t,\rho)B_2(\rho)v_2(\rho)d\rho, (L_{24}u_2)(t) = C_2(t)\int_0^t \Phi_2(t,\rho)D_2(\rho)u_2(\rho)d\rho,$$

$$(L_{31}u_1)(t) = C_1(t)\int_0^t \Phi_1(t,\rho)D_1(\rho)u_1(\rho)d\rho, (L_{32}v_1)(t) = C_1(t)\int_0^t \Phi_1(t,\rho)B_1(\rho)v_1(\rho)d\rho,$$

$$(L_{43}v_2)(t) = E_2(t)\int_0^t \Phi_2(t,\rho)B_2(\rho)v_2(\rho)d\rho, (L_{44}u_2)(t) = E_2(t)\int_0^t \Phi_2(t,\rho)D_2(\rho)u_2(\rho)d\rho.$$

Hence

$$\|\mathbb{L}_0\| = \max\{\|{}^1\mathbb{L}_0\|, \|{}^2\mathbb{L}_0\|\} \quad \text{with } {}^1\mathbb{L}_0 = \begin{bmatrix} L_{11} & L_{12} \\ L_{31} & L_{32} \end{bmatrix}, \quad {}^2\mathbb{L}_0 = \begin{bmatrix} L_{23} & L_{24} \\ L_{43} & L_{44} \end{bmatrix}. \tag{93}$$

Introducing time-invariant positive scaling $\alpha_1, \ldots, \alpha_4$, we have

$$\|\mathbb{L}_0^\alpha\| = \max\{\|{}^1\mathbb{L}_0^\alpha\|, \|{}^2\mathbb{L}_0^\alpha\|\}, \tag{94}$$

where

$${}^1\mathbb{L}_0^\alpha = \begin{bmatrix} L_{11} & \frac{\alpha_1}{\alpha_2}L_{12} \\ \frac{\alpha_3}{\alpha_1}L_{31} & \frac{\alpha_3}{\alpha_2}L_{32} \end{bmatrix}, \qquad {}^2\mathbb{L}_0^\alpha = \begin{bmatrix} \frac{\alpha_2}{\alpha_3}L_{23} & \frac{\alpha_2}{\alpha_4}L_{24} \\ \frac{\alpha_4}{\alpha_3}L_{43} & L_{44} \end{bmatrix}.$$

Suppose l_{ij} are upper bounds for $\|\mathbb{L}_{ij}\|$, which we assume are non-zero $(i,j = 1,2)$, then

$$\|{}^1\mathbb{L}_0^\alpha\| \le \left\| \begin{bmatrix} l_{11} & \frac{\alpha_1}{\alpha_2}l_{12} \\ \frac{\alpha_3}{\alpha_1}l_{31} & \frac{\alpha_3}{\alpha_2}l_{32} \end{bmatrix} \right\|, \qquad \|{}^2\mathbb{L}_0^\alpha\| \le \left\| \begin{bmatrix} \frac{\alpha_2}{\alpha_3}l_{23} & \frac{\alpha_2}{\alpha_4}l_{24} \\ \frac{\alpha_4}{\alpha_3}l_{43} & l_{44} \end{bmatrix} \right\|. \tag{95}$$

LEMMA 6.4. *Given* $\beta_i > 0$, $i \in \underline{4}$, *then*

$$\left\| \begin{bmatrix} \beta_1 & \sqrt{\beta_2\beta_3} \\ \sqrt{\beta_2\beta_3} & \beta_4 \end{bmatrix} \right\| \le \left\| \begin{bmatrix} \beta_1 & \gamma\beta_2 \\ \beta_3/\gamma & \beta_4 \end{bmatrix} \right\| \quad \textit{for all } \gamma > 0. \tag{96}$$

Proof. The two matrices in (96) have the same eigenvalues. Because the matrix on the LHS is symmetric, its norm is equal to the spectral radius, whereas the norm of the matrix on the RHS is not less than this spectral radius. □

Setting $\alpha_3/\alpha_2 = \rho$ and choosing α_1, α_4 to optimise the estimates in (95), we see

$$\|{}^1\mathbb{L}_0^\alpha\| \le \left\| \begin{bmatrix} l_{11} & \sqrt{\rho l_{12}l_{31}} \\ \sqrt{\rho l_{12}l_{31}} & \rho l_{32} \end{bmatrix} \right\|, \qquad \|{}^2\mathbb{L}_0^\alpha\| \le \left\| \begin{bmatrix} l_{23}/\rho & \sqrt{l_{24}l_{43}/\rho} \\ \sqrt{l_{24}l_{43}/\rho} & l_{44} \end{bmatrix} \right\|. \tag{97}$$

PROPOSITION 6.5. *With the notations introduced above,* $\|\mathbb{L}_0^\alpha\| \leq \lambda_m$, *where* λ_m *is the spectral radius of the nonnegative matrix*

$$L = \begin{bmatrix} l_{11} & l_{12} & 0 & 0 \\ 0 & 0 & l_{23} & l_{24} \\ l_{31} & l_{32} & 0 & 0 \\ 0 & 0 & l_{43} & l_{44} \end{bmatrix}.$$

Proof. By the Frobenius theorem, λ_m is an eigenvalue of L; hence it satisfies the associated characteristic equation. A simple calculation yields

$$(\lambda_m^2 - l_{44}\lambda_m)(\lambda_m^2 - l_{11}\lambda_m) + (l_{23}l_{44} - l_{24}l_{43} - l_{23}\lambda_m)(l_{12}l_{31} - l_{11}l_{32} + l_{32}\lambda_m) = 0. \tag{98}$$

From (97) $\|\mathbb{L}_0^\alpha\| \leq \max\{\mu_1(\rho), \mu_2(\rho)\}$, where $\mu_1(\rho), \mu_2(\rho)$ are the maximum solutions of

$$\lambda^2 - (l_{11} + \rho l_{32})\lambda + \rho(l_{11}l_{32} - l_{12}l_{31}) = 0, \tag{99}$$

$$\lambda^2 - (l_{44} + l_{23}/\rho)\lambda + (l_{23}l_{44} - l_{24}l_{43})/\rho = 0, \tag{100}$$

respectively. It is not difficult to show that the continuous function $\mu_1(\cdot)$ is strictly increasing from l_{11} to ∞, whilst $\mu_2(\cdot)$ is strictly decreasing from ∞ to l_{44} as ρ varies from 0 to ∞. Hence there is a unique $\hat{\rho} \in (0, \infty)$ for which $\mu_1(\hat{\rho}) = \mu_2(\hat{\rho})$ and by (94), (97), $\|\mathbb{L}_0^\alpha\| \leq \mu_2(\hat{\rho})$. All that remains to be proved is $\mu_2(\hat{\rho}) = \lambda_m$. Now $\mu_2(\hat{\rho}) > l_{44}$ and $l_{23}\mu_2(\hat{\rho}) - l_{23}l_{44} + l_{24}l_{43} > 0$. Hence from (100),

$$\hat{\rho} = \frac{l_{23}\mu_2(\hat{\rho}) - l_{23}l_{44} + l_{24}l_{43}}{(\mu_2(\hat{\rho}))^2 - l_{44}\mu_2(\hat{\rho})},$$

and substituting in (99) we see that $\mu_2(\hat{\rho})$ satisfies (98). Suppose there exists a root $\tilde{\lambda}$ of (98) with $\tilde{\lambda} > \mu_2(\hat{\rho})$, then $\tilde{\lambda} > l_{44}$ and $l_{23}\tilde{\lambda} - l_{23}l_{44} + l_{24}l_{43} > 0$. Let

$$\rho = \frac{l_{23}\tilde{\lambda} - l_{23}l_{44} + l_{24}l_{43}}{\tilde{\lambda}^2 - l_{44}\tilde{\lambda}},$$

then by (98), $\tilde{\lambda}^2 - l_{11}\tilde{\lambda} - \rho(l_{11}l_{31} - l_{11}l_{32} + l_{32}\tilde{\lambda}) = 0$. Hence the pair $(\tilde{\lambda}, \rho)$ satisfies (99), (100), which is a contradiction. This completes the proof. □

Summarizing the above results we have the following.

PROPOSITION 6.6. *Suppose* $(\Phi(\cdot), D, E)$ *satisfies Hypotheses* 1–8 *and* λ_m *is the spectral radius of the matrix*

$$M_L = \begin{bmatrix} \|L_{11}\| & \|L_{12}\| & 0 & 0 \\ 0 & 0 & \|L_{23}\| & \|L_{24}\| \\ \|L_{31}\| & \|L_{32}\| & 0 & 0 \\ 0 & 0 & \|L_{43}\| & \|L_{44}\| \end{bmatrix},$$

where the operators L_{ij} *are defined by* (92). *If*

$$\|\Delta_i(\cdot)\|_\infty, \|K_i(\cdot)\|_\infty < \lambda_m^{-1}, \qquad i = 1, 2,$$

then the uncertain coupled system (90) *is exponentially stable. Moreover, the same conclusion holds if* λ_m *is the spectral radius of the matrix* M_l, *where the norms of the* L *operators are replaced by upper estimates.*

As an application of this result, let us assume that the two basic models and uncertainty structures are those considered in Examples 6.1 and 6.2. To illustrate the proposition rather than the (difficult) estimation of the norm of various operators, we assume that $B_1(t) = C_1(t) = D_1(t) = E_1(t) = t^{1/2}I$ and $B_2(t) = C_2(t) = D_2(t) = E_2(t) = A_0^{1/2}$, and $A_1(t) = tA$, $A_2(t) = A_0$, where A_1A_0 are as in Examples 6.1 and 6.2. Then each of the norms $\|L_{11}\|$, $\|L_{12}\|$, $\|L_{31}\|$, $\|L_{32}\|$ is bounded by $[\max_{\omega\in\mathbb{R}} \|(i\omega I + A)^{-1}\|]^{-1} =: \sigma$ and each of the norms $\|L_{23}\|$, $\|L_{24}\|$, $\|L_{43}\|$, $\|L_{44}\|$ is bounded by k_0^{-1}. So the matrix M_l is

$$M_l = \begin{bmatrix} \sigma & \sigma & 0 & 0 \\ 0 & 0 & k_0^{-1} & k_0^{-1} \\ \sigma & \sigma & 0 & 0 \\ 0 & 0 & k_0^{-1} & k_0^{-1} \end{bmatrix}$$

and the maximum eigenvalue is $k_0^{-1} + \sigma$. Note that the estimate (93) obtained without scaling gives $\|\mathbb{L}_0\| \le 2\max\{k_0^{-1}; \sigma\}$, which is inferior to $k_0^{-1} + \sigma$ for all $\sigma \ne k_0^{-1}$. In this very special case the perturbation can be written in the form

$$\begin{bmatrix} t^{1/2}I & 0 \\ 0 & A_0^{1/2} \end{bmatrix} \begin{bmatrix} \Delta_1 & K_1 \\ K_2 & \Delta_2 \end{bmatrix} \begin{bmatrix} t^{1/2}I & 0 \\ 0 & A_0^{1/2} \end{bmatrix}.$$

Applying Theorem 3.2 to this single perturbation structure, we see that the perturbed system is exponentially stable if

$$\left\| \begin{matrix} \Delta_1 & K_1 \\ K_2 & \Delta_2 \end{matrix} \right\| < \min\{k_0; 1/\sigma\}.$$

In the particular case that $H = L^2(\Omega)$ and $\Delta_1 = \Delta_2 = K_1 = K_2 = \delta I$, the above estimate requires $\delta < \min\{k_0/2;\ 1/(2\sigma)\}$, which is again inferior to the estimate $\delta < 1/(k_0^{-1} + \sigma)$ obtained from Proposition 6.6.

Acknowledgment. We thank the anonymous referee whose detailed comments helped us to improve the organization and content of the paper.

REFERENCES

[1] P. Bohl, *Über Differentialgleichungen*, *Journal für Reine und Angewandte Mathematik*, 144 (1913), pp. 284–313.

[2] R. F. Curtain, H. Logemann, S. Townley, and H. Zwart, *Well-posedness, stabilizability and admissibility for Pritchard Salamon systems*, Report 260, Institut für Dynamische Systeme, Universität Bremen, 1992.

[3] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Information Sciences, Vol. 8, Springer-Verlag, Berlin, Heidelberg, New York, 1978.

[4] Ju. L. Daleckii and M. G. Krein, *Stability of Solutions of Differential Equations in Banach Spaces*, American Mathematical Society, Providence RI, 1974.

[5] R. Datko, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–445.

[6] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input–Output Properties*, Academic Press, New York, 1975.

[7] N. Dunford and J. T. Schwartz, *Linear Operators, Part* I: *General Theory*, Wiley-Interscience, New York, 1958.

[8] H. O. Fattorini, *The Cauchy Problem*, Addison–Wesley, Reading, MA, 1983.

[9] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, American Mathematical Society, Providence, RI, 1957.

[10] D. Hinrichsen, A. Ilchmann, and A. J. Pritchard, *Robustness of stability of time-varying linear systems*, J. Differential Equations, 82 (1989), pp. 219–250.

[11] D. Hinrichsen and A. J. Pritchard, *Stability radius for structured perturbations and the algebraic Riccati equation*, Systems Control Lett. 8 (1986), pp. 105–113.

[12] D. HINRICHSEN AND A. J. PRITCHARD, *Real and complex stability radii: A survey*, in Proc. Workshop Control of Uncertain Systems, Bremen, 1989; Progress in System and Control Theory, Vol. 6, Birkhäuser-Verlag, Basel, Switzerland, 1990, pp. 119–162.

[13] ———, *Robust stability of linear time-varying systems with respect to multi-perturbations*, in Proc. European Control Conference, Grenoble, 1991, pp. 1366–1371.

[14] T. KATO, *Integration of the equation of evolution in a Banach space*, J. Japan Math. Soc., 5 (1953), pp. 208–234.

[15] ———, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[16] M. G. KREIN, *A generalization of some investigations of A.M. Liapunov on linear differential equations with periodic coefficients*, Dokl. Akad. Nauk SSR, 73 (1950) pp. 445–448.

[17] S. G. KREIN, *Linear Differential Equations in Banach Space*, Translations of Mathematical Monographs, Vol. 29, American Mathematical Society, Providence, RI, 1971.

[18] J. L. LIONS, *Optimal Control of Systems described by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[19] N. OKAZAWA, *A perturbation theorem for linear contraction semigroups on reflexive Banach spaces*, Proc. Japan Acad., 47 (1971), pp. 947–949.

[20] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, 1983.

[21] R. S. PHILLIPS, *Perturbation theory for semigroups of linear operators*, Trans. Amer. Math. Soc., 74 (1953), pp. 199–221.

[22] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic control problem for infinite-dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.

[23] A. J. PRITCHARD AND S. TOWNLEY, *A stability radius for infinite dimensional systems*, in Proc. Conf. of Distributed Parameter Systems, Vorau, 1986; Lecture Notes Control and Information Sciences, Vol. 102, Springer-Verlag, Berlin, Heidelberg, New York, 1987.

[24] ———, *Robustness of linear systems*, J. Differential Equations, 77 (1989), pp. 254–286.

[25] H. TANABE, *Evolution equations of parabolic type*, Proc. Japan Acad., 37 (1961), pp. 610–613.

[26] ———, *Equations of Evolution*, Pitman, London, 1979.

[27] G. WEISS, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

[28] ———, *Regular linear systems with feedback*, Report, The Weizmann Institute of Science, 1993.

[29] ———, *Transfer functions of regular linear systems, Part* I: *Characterizations of regularity*, Trans. Amer. Math. Soc., to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1542–1554, November 1994

002

SOLUTION DIFFERENTIABILITY FOR NONLINEAR PARAMETRIC CONTROL PROBLEMS*

HELMUT MAURER† AND HANS JOSEF PESCH‡

Abstract. Perturbed nonlinear control problems with data depending on a vector parameter are considered. Using second-order sufficient optimality conditions, it is shown that the optimal solution and the adjoint multipliers are differentiable functions of the parameter. The proof exploits the close connections between solutions of a Riccati differential equation and shooting methods for solving the associated boundary value problem. Solution differentiability provides a firm theoretical basis for numerical feedback schemes that have been developed for computing neighbouring extremals. The results are illustrated by an example that admits two extremal solutions. Second-order sufficient conditions single out one optimal solution for which a sensitivity analysis is carried out.

Key words. parametric control problems, second-order sufficient conditions, solution differentiability, shooting methods, feedback controls

AMS subject classifications. 49K15, 49K40

1. Introduction.

This paper is concerned with parametric nonlinear control problems where all data depend on a vector parameter $p \in \mathbb{R}^k$. To make the main ideas more transparent, we restrict the discussion to the following basic control problem:

$$\text{(P(p))} \qquad \begin{aligned} &\text{Minimize} && J(x,u,p) = \int_a^b L(t,x,u,p)dt \\ &\text{subject to} && \dot{x} = f(t,x,u,p), \qquad a \le t \le b, \\ & && x(a) = \varphi(p), \qquad x(b) = \psi(p). \end{aligned}$$

Problem (P(p_0)) corresponding to a fixed parameter $p_0 \in \mathbb{R}^k$ is considered the *unperturbed* problem. It is assumed that a local minimum (optimal solution) x_0, u_0 exists for (P(p_0)). An important problem in sensitivity analysis is the following: Find conditions for the unperturbed optimal solution x_0, u_0 such that the perturbed problem (P(p)) admits an optimal solution $x(p)$, $u(p)$ near x_0, u_0 that is a differentiable function of the parameter p near p_0. Comparing sensitivity approaches in optimization and optimal control it is apparent that *second-order sufficient optimality conditions* (SSCs) are a crucial assumption for this type of sensitivity result. Let us briefly review some papers in this regard.

A survey of basic results on stability and sensitivity for *finite-dimensional* optimization problems can be found in Fiacco [15]. A direct generalization of the second-order sensitivity result to equality constrained *optimization problems in Hilbert spaces* is given in Wierzbicki and Kurcyusz [46, Thm. 8.6]. For *Hilbert-space* optimization problems with general cone constraints, Alt [1]–[3] and Malanowski [24]–[28] have shown that the optimal solution is directionally differentiable with respect to the parameter. These results have recently been extended by Colonius and Kunisch [10]. The setting in Hilbert spaces allows for applications to *convex* control problems. A direct treatment of convex control problems with control appearing linearly has been performed earlier by Dontchev [13] and Malanowski [24]–[26].

There is a second stream of papers dealing with the second-order sensitivity analysis for *nonlinear* control problems. These papers are mainly concerned with developing neighbouring feedback schemes for perturbed solutions. Here the main ideas go back as far as the ingenious papers of Breakwell and Ho [6], Breakwell et al. [7], and Kelley [17], [18]. This approach is

* Received by the editors June 10, 1992; accepted for publication (in revised form) June 2, 1993.
† Institut für Numerische und Instrumentelle Mathematik, Westfälische Wilhelms-Universität Münster, Einsteinstrasse 62, 48149 Münster, Germany.
‡ Mathematisches Institut, Technische Universität München, Arcisstrasse 21, 80333 München, Germany.

summarized in Bryson and Ho [8]. Similar ideas may be found in [11], [14], [22], [23], [36], [47]. Extensions to control problems with inequality constraints are treated in Bock [4], Bock and Krämer-Eis [5], Dillon and Tun [12], Krämer-Eis [19]. Kugelmann and Pesch [20], [21], and Pesch [37]–[41]. All these papers suffer from the fact that the theory developed therein is rather formal and nonrigorous. The work of these authors supports the fact that theory lags behind numerical implementation.

We may conclude that a second-order sensitivity result for *nonlinear* control problems is still lacking. It is the main purpose of this paper to provide such a result using SSCs. SSCs depend on the existence of a finite solution of a *Riccati* differential equation that is associated with the variational system for the underlying boundary value problem. In addition to SSCs, we require that the iteration matrix of the shooting method is nonsingular. The nonsingularity of this matrix allows for an application of the implicit function theorem that yields a family of neighbouring extremals. The implicit function theorem has been used by many authors to establish the existence of a family of extremals; cf. [4], [5], [19], [31], [37]–[41]. However, the proof of optimality remains incomplete unless one superimposes sufficient conditions.

Up to now, the two streams of papers described above have flown separately. We hope that the ideas of this paper provide some help for merging these two streams into one that carries theoretical and numerical methods as close partners. This paper is a slightly shortened version of the report [34]. After completing this report we became aware that Malanowski [29], [30] obtained results on directional differentiability of solutions to nonlinear parametric optimization and control problems. Malanowski develops a fully infinite-dimensional implicit function theorem based on the two-norm approach. Our approach uses more elementary methods and tries to interweave theory and numerical techniques.

2. Second-order sufficient conditions and shooting methods. We begin with the *unperturbed* problem (P(p_0)) and suppress the parameter p_0 in this section. We summarize the second-order sufficient conditions (SSCs) derived in [16], [32], [35], [44], [48]–[50] and establish the close connections between SSCs and shooting methods for solving the associated boundary value problem (BVP).

Let the following data be given: a fixed interval $[a, b] \subset \mathbb{R}$; end-points x_a, $x_b \in \mathbb{R}^n$; an open, convex, and bounded set $U \subset \mathbb{R}^m$; and functions $L : \mathbb{R} \times \mathbb{R}^n \times U \to \mathbb{R}$ and $f : \mathbb{R} \times \mathbb{R}^n \times U \to \mathbb{R}^n$. The control problem (P) is defined to be

$$\text{(P)} \qquad \text{minimize} \quad J(x,u) = \int_a^b L(t,x,u)dt$$

over all feasible pairs (x, u) of piecewise continuous functions $u : [a, b] \to \mathbb{R}^m$ and absolutely continuous functions $x : [a, b] \to \mathbb{R}^n$ such that

$$\dot{x} = f(t,x,u) \quad \text{for almost every } t \in [a,b], \tag{2.1}$$

$$x(a) = x_a, \qquad x(b) = x_b, \tag{2.2}$$

$$u(t) \in U. \tag{2.3}$$

The control constraint (2.3) with U *open* and bounded has been introduced for technical reasons; mainly it should allow for a practical verification of the regularity condition in Definition 2.1.

Let $C^n[a, b]$ denote the space of continuous functions $x : [a, b] \to \mathbb{R}^n$ equipped with the usual topology. For $x \in C^n[a, b]$ and $\varepsilon > 0$ we denote by $B(x; \varepsilon)$ the open ε-ball around x in $C^n[a, b]$. Similarly, the *tube* about $x \in C^n[a, b]$ in $\mathbb{R}^{n+1}$ is the set

$$T(x;\varepsilon) = \{(t,y) \in \mathbb{R}^{n+1} \mid t \in [a,b], \|y - x(t)\| < \varepsilon\}.$$

A feasible pair (x_0, u_0) is called a (*strong*) *local minimum* if for some $\varepsilon > 0$,

$$J(x, u) \geq J(x_0, u_0)$$

for all feasible pairs (x, u) with $x \in B(x_0; \varepsilon)$ and u satisfying (2.3).

We use hereafter the terminology $\varphi(t) = \varphi(t, x_0(t), u_0(t))$ for any function φ. Given a pair (x_0, u_0) we assume the following hypotheses.

Hypothesis 1. The functions L and f are of class C^k with $k \geq 2$ on $T(x_0; \varepsilon) \times U$.

Hypothesis 2. The linearized system $\dot{y} = f_x(t)y + f_u(t)v$ is completely controllable in $[a, b]$.

The controllability assumption (Hypothesis 2) is usually referred to as the *normality* condition.

The *Hamiltonian* of (P) is defined by

$$H(t, x, \lambda, u) = L(t, x, u) + \lambda^T f(t, x, u), \qquad \lambda \in \mathbb{R}^n, \tag{2.4}$$

where T denotes the transpose. Assuming normality (Hypothesis 2), the first-order necessary conditions for a strong local minimum (minimum principle of *Pontryagin*) are as follows. There exists an absolutely continuous function $\lambda_0 : [a, b] \to \mathbb{R}^n$ such that

$$\dot{\lambda}_0 = -H_x(t)^T \quad \text{for almost every } t \in [a, b], \tag{2.5}$$

$$u_0(t) \in \arg\min\{H(t, x_0(t), \lambda_0(t), u) \mid u \in U\} \quad \text{for all } t \in [a, b]. \tag{2.6}$$

The latter minimum condition yields because U is open

$$H_u(t) = 0, \qquad H_{uu}(t) \geq 0 \quad \text{(positive semidefinite)}.$$

One basic assumption for SSCs is that the following strengthened *Legendre* condition holds:

$$H_{uu}(t) > 0 \quad \text{positive definite for } t \in [a, b]. \tag{2.7}$$

This condition is not sufficient to guarantee the *continuity* of the control $u_0(t)$. The continuity and, in fact, the smoothness of $u_0(t)$ follows from the *regularity* of the *Hamiltonian*. In the following definition, $T(x_0, \lambda_0; \varepsilon)$ denotes the tube about (x_0, λ_0) which is defined in an obvious way.

DEFINITION 2.1. *Let $k \geq 1$. The Hamiltonian H is called C^k-regular (about (x_0, λ_0, u_0)) if there exists $\varepsilon > 0$ and a C^k-function*

$$u^* : T(x_0, \lambda_0; \varepsilon) \to U$$

such that

$$u^*(t, x, \lambda) = \arg\min\{H(t, x, \lambda, u) \mid u \in U\}$$

is the unique minimum for all $(t, x, \lambda) \in T(x_0, \lambda_0; \varepsilon)$.

This condition strengthens the regularity condition (1.2)‴ of Zeidan [49, p. 22]. Also, this regularity condition is tacitly underlying numerical methods for solving the BVP defined by (2.1), (2.2), (2.5), and (2.6). This can be seen as follows. The optimal solution (x_0, u_0) is obtained by solving the BVP

$$\begin{aligned} \dot{x} &= f(t, x, u^*(t, x, \lambda)), \\ \dot{\lambda} &= -H_x(t, x, \lambda, u^*(t, x, \lambda))^T \end{aligned} \tag{2.8}$$

with boundary values $x(a) = x_a$, $x(b) = x_b$. The solutions $x_0(t)$, $\lambda_0(t)$ of this BVP are C^k-functions because the right-hand side of (2.8) is a C^k-function. Hence the optimal control

$$u_0(t) = u^*(t, x_0(t), \lambda_0(t)) \tag{2.9}$$

is also a C^k-function.

Remark. The C^k-regularity of the Hamiltonian is not as restrictive as it may appear. The C^k-regularity holds for most practical examples where the control variable appears *nonlinearly*; compare, for example, the famous reentry problem in Stoer and Bulirsch [45]. Observe that the open and bounded control set U has been introduced to allow for a check of the uniqueness of the minimizing function $u^*(t, x, \lambda)$. Note also that the strict Legendre–Clebsch condition (2.7) alone does not guarantee the continuity or even the differentiability of the control $u(t)$ that is indispensable for the sensitivity result in Theorem 3.1. This can be seen by choosing $L(t, x, u) = (u^2 - 1)^2$, $U = \mathbb{R}$, $f(t, x, u) = u$. Here any control $u(t) = \pm 1$ is optimal with $H_{uu}(t) \equiv 8$, but H is not C^k-regular.

Next, we introduce the *variational system* corresponding to (2.8). The continuity of $u_0(t)$ and (2.7) imply that there exists $\varepsilon > 0$ such that the C^k-function $u^*(t, x, \lambda)$ in Definition 2.1 satisfies

$$\begin{aligned} H_u(t, x, \lambda, u^*(t, x, \lambda)) &= 0 \\ H_{uu}(t, x, \lambda, u^*(t, x, \lambda)) &> 0 \end{aligned} \qquad \text{for } (t, x, \lambda) \in T(x_0, \lambda_0; \varepsilon). \tag{2.10}$$

By differentiation of the first equation we obtain the identities

$$H_{ux} + H_{uu} u_x^* \equiv 0, \qquad H_{u\lambda} + H_{uu} u_\lambda^* \equiv 0$$

and hence in view of the second relation in (2.10) and $H_{u\lambda} = f_u^T$ we obtain

$$u_x^* = -H_{uu}^{-1} H_{ux}, \qquad u_\lambda^* = -H_{uu}^{-1} f_u^T. \tag{2.11}$$

Then the variational system for (2.8) about (x_0, u_0) becomes (see [8, (6.1.21)–(6.1.25)], [38], [51], [52])

$$\dot{y} = A(t)y + B(t)\eta, \qquad \dot{\eta} = C(t)y - A(t)^T \eta, \tag{2.12}$$

where

$$\begin{aligned} A(t) &= f_x(t) - f_u(t) H_{uu}(t)^{-1} H_{ux}(t), \\ B(t) &= -f_u(t) H_{uu}(t)^{-1} f_u(t)^T, \\ C(t) &= -H_{xx}(t) + H_{xu}(t) H_{uu}(t)^{-1} H_{ux}(t). \end{aligned} \tag{2.13}$$

We use the system with vector solutions $y(t), \eta(t)$ as well as with (n, n) matrix solutions $y(t), \eta(t)$.

Our aim now is to establish the connection between the variational system (2.12) and shooting methods for solving the BVP (2.2), (2.8). Consider solutions of the differential equation (2.8) with initial values depending on a shooting parameter $s \in \mathbb{R}^n$ (compare [9], [45]):

$$x(a) = x_a, \qquad \lambda(a) = s. \tag{2.14}$$

These solutions denoted by $x(t, s)$ and $\lambda(t, s)$ are C^k-functions for s near $s_0 := \lambda_0(a)$. We have to solve the nonlinear equation

$$F(s) := x(b, s) - x_b = 0, \tag{2.15}$$

where F is a C^k function for s near s_0. Newton's method for solving this equation requires the nonsingularity of the matrix

$$y(b) := \frac{\partial F}{\partial s}(s_0) = \frac{\partial x}{\partial s}(b, s_0). \tag{2.16}$$

The matrix $y(b)$ is computed by noting that the matrices

$$y(t) = \frac{\partial x}{\partial s}(t, s_0), \qquad \eta(t) = \frac{\partial \lambda}{\partial s}(t, s_0) \tag{2.17}$$

are solutions of the variational system (2.12) with initial conditions

$$y(a) = O_n, \qquad \eta(a) = I_n. \tag{2.18}$$

Now we consider the following matrix Riccati equation associated with the variational system (2.12) (compare Reid [43, Chap. III]):

$$\begin{aligned} \dot{Q} &= -QA(t) - A(t)^T Q - QB(t)Q + C(t) \\ &= -Qf_x(t) - f_x(t)^T Q - H_{xx}(t) \\ &\quad +(H_{xu}(t) + Qf_u(t))H_{uu}(t)^{-1}(H_{ux}(t) + f_u(t)^T Q). \end{aligned} \tag{2.19}$$

Here, $Q(t)$ is a symmetric $n \times n$ matrix. Based on the Riccati equation (2.19), the following SSCs are obtained in [32, Thm. 5.2], [35, Thm. 2.2], [44, Thm. 5.3], and [49, Thm. 2.2].

THEOREM 2.1. *Let (x_0, u_0) be a feasible pair for* (P) *such that Hypotheses* 1 *and* 2 *hold. Assume that there exists an absolutely continuous function $\lambda_0 : [a, b] \to \mathbb{R}^n$ such that the necessary conditions* (2.5), (2.6) *are satisfied and assume further that the following conditions hold*:

(a) $H_{uu}(t) > 0$, $\forall\, t \in [a, b]$;

(b) *the Hamiltonian H is C^k regular*;

(c) *there exists a symmetric C^1 solution $Q(t)$ of the Riccati equation* (2.19).

Then (x_0, u_0) provides a local minimum for (P) *and, moreover, u_0 is a C^k function.*

Note that conditions (a)–(c) are stable with respect to small C^k perturbations of the data. This property is crucial for the second-order sensitivity result in the next section.

3. Solution differentiability. The problem (P) considered in §2 is embedded into the following parametric control problem (P(p)) depending on a parameter $p \in \mathbb{R}^k$:

$$\text{Minimize} \quad J(x, u, p) = \int_a^b L(t, x, u, p)dt$$

subject to

$$\dot{x} = f(t, x, u, p), \qquad a \le t \le b, \tag{3.1}$$

$$x(a) = \varphi(p), \qquad x(b) = \psi(p), \tag{3.2}$$

$$u(t) \in U. \tag{3.3}$$

The unperturbed problem corresponding to $p = p_0 \in \mathbb{R}^k$ is identified with problem (P) of §2. Let (x_0, u_0) be a feasible pair for (P(p_0)). Hypothesis 1 is replaced by Hypothesis 1′.

Hypothesis 1′. The functions $L : \mathbb{R}\times\mathbb{R}^n\times U\times\mathbb{R}^k \to \mathbb{R}$ and $f : \mathbb{R}\times\mathbb{R}^n\times U\times\mathbb{R}^k \to \mathbb{R}^n$ are of class C^k ($k \geq 2$) on $T(x_0, u_0; \varepsilon) \times U \times B(p_0; \varepsilon)$ and the functions $\varphi, \psi : \mathbb{R}^k \to \mathbb{R}^n$ are of class C^k on $B(p_0; \varepsilon)$ for some $\varepsilon > 0$.

The Hamiltonian for problem (P(p)) is

$$H(t, x, \lambda, u, p) = L(t, x, u, p) + \lambda^T f(t, x, u, p), \qquad \lambda \in \mathbb{R}^n. \tag{3.4}$$

We assume that (x_0, u_0) satisfies the second-order sufficient conditions of Theorem 2.1 with a C^k function λ_0. The C^k regularity of the unperturbed *Hamiltonian* (2.4) carries over to the perturbed *Hamiltonian* (3.4): there exists $\varepsilon > 0$ and a C^k function

$$u^* : T(x_0, \lambda_0; \varepsilon) \times B(p_0; \varepsilon) \to U$$

such that the minimum of H is uniquely attained at

$$u^*(t, x, \lambda, p) = \arg\min\{H(t, x, \lambda, u, p) \mid u \in U\} \tag{3.5}$$

for all $(t, x, \lambda, p) \in T(x_0, \lambda_0; \varepsilon) \times B(p_0; \varepsilon)$. The uniqueness follows from the compactness of $\bar{U}$ combined with arguments used in Proposition 3.1 of Zeidan [49]. The smoothness property of u^* is a consequence of the implicit function theorem because u^* satisfies

$$H_u(t, x, \lambda, u^*(t, x, \lambda, p), p) = 0$$

and the strict *Legendre* condition (2.7) holds.

Now we can state the main result of this paper. A preliminary version for more general control problems has been announced in [40].

THEOREM 3.1 (Solution differentiability). *Let (x_0, u_0) be feasible for the unperturbed problem* (P(p_0)) *such that Hypothesis* 1′ *holds. Assume that (x_0, u_0) satisfies the second-order sufficient conditions of Theorem* 2.1 *and that the shooting matrix $y(b)$ in* (2.16) *is regular.*

Then there exists a neighbourhood $V \subset \mathbb{R}^k$ of $p = p_0$ and C^k functions

$$x, \lambda : [a, b] \times V \to \mathbb{R}^n, \qquad u : [a, b] \times V \to U$$

such that the following statements hold:

(1) $x(t, p_0) = x_0(t)$, $u(t, p_0) = u_0(t)$, $\lambda(t, p_0) = \lambda_0(t)$ *for all* $t \in [a, b]$;

(2) *The triple $x(\cdot, p)$, $u(\cdot, p)$, $\lambda(\cdot, p)$ satisfies the second-order sufficient conditions in Theorem* 2.1 *for all $p \in V$ and $x(\cdot, p)$, $u(\cdot, p)$ provide a strong local minimum for* (P(p)).

Proof. In a first step we construct functions $x(t, p)$, $u(t, p)$, $\lambda(t, p)$ satisfying the first order conditions (2.5), (2.6) for p near p_0. Using the minimizing function $u^*(t, x, \lambda, p)$ in (3.5), this amounts to solving the BVP

$$\begin{aligned} \dot{x} &= f(t, x, u^*(t, x, \lambda, p), p), \\ \dot{\lambda} &= -H_x(t, x, \lambda, u^*(t, x, \lambda, p), p)^T, \end{aligned} \tag{3.6}$$

$$x(a) = \varphi(p), \qquad x(b) = \psi(p). \tag{3.7}$$

The shooting procedure is a parametric version of (2.14)–(2.18). We consider the differential equation (3.6) with initial values depending on a shooting parameter $s \in \mathbb{R}^n$,

$$x(a) = \varphi(p), \qquad \lambda(a) = \lambda_0(a) + s.$$

Solutions $\tilde{x}(t,p,s)$, $\tilde{\lambda}(t,p,s)$ exist in $[a,b]$ for $\|s\|$ and $\|p-p_0\|$ small and are C^k functions with respect to all arguments (t,p,s). Then the mapping $F:\mathbb{R}^k\times\mathbb{R}^n\to\mathbb{R}^n$ defined by

$$F(p,s)=\tilde{x}(b,p,s)-\psi(p)$$

is a C^k function with $F(p_0,0)=0$. Solving the BVP (3.6), (3.7) is then equivalent to solving the nonlinear equation $F(p,s)=0$ for $s=s(p)$. To apply the implicit function theorem, we have to check the nonsingularity of the (n,n)-matrix

$$\frac{\partial F}{\partial s}(p_0,0)=\frac{\partial\tilde{x}}{\partial s}(b,p_0,0).$$

As we have already seen in (2.16) this matrix agrees with the matrix $y(b)$ where

$$y(t)=\frac{\partial\tilde{x}}{\partial s}(t,p_0,0),\qquad \eta(t)=\frac{\partial\tilde{\lambda}}{\partial s}(t,p_0,0)$$

are solutions of the variational system (2.12) with initial conditions

$$y(a)=O_n,\qquad \eta(a)=I_n.$$

Because $y(b)$ is assumed to be regular, the implicit function theorem yields a neighbourhood $V\subset\mathbb{R}^k$ of $p=p_0$ and a C^k function $s:V\to\mathbb{R}^n$ such that $s(p_0)=0$ and

$$F(p,s(p))=\tilde{x}(b,p,s(p))-\psi(p)=0\qquad\forall\, p\in V.$$

The conclusion to this point is that the functions

$$x(t,p):=\tilde{x}(t,p,s(p)),\qquad \lambda(t,p):=\tilde{\lambda}(t,p,s(p))$$

are C^k functions that solve the BVP (3.6) and (3.7) for $p\in V$. The associated control function

$$u(t,p):=u^*(t,x(t,p),\lambda(t,p),p)$$

is also of class C^k and satisfies the minimum principle in view of (3.5). Claim (1) of the theorem is immediate.

In a second step we have to show that, indeed, $x(t,p)$ and $u(t,p)$ are optimal for problem (P(p)). We can choose the neighbourhood V so small that the following two statements are true for all $p\in V$:

(a) The strict *Legendre* condition holds

$$H_{uu}(t,x(t,p),\lambda(t,p),u(t,p),p)>0\qquad\forall\, t\in[a,b];$$

(b) The *Riccati* equation

$$\dot{Q}=-Q\,A(t,p)-A(t,p)^TQ-Q\,B(t,p)Q+C(t,p)$$

has a symmetric C^1 solution $Q(t,p)$ on $[a,b]$ where $A(t,p)$, $B(t,p)$, $C(t,p)$ are the matrices (2.13) evaluated at $x(t,p)$, $\lambda(t,p)$, $u(t,p)$.

Statement (b) follows from the standard embedding theorem for differential equations. Applying Theorem 2.1 for each $p\in V$, we arrive at the desired conclusion that the pair $(x(\cdot,p),u(\cdot,p))$ is a local minimum for every $p\in V$. □

We now briefly illustrate the use of this sensitivity result when devising efficient numerical feedback schemes for neighbouring extremals. Because the functions $x(t,p)$, $\lambda(t,p)$, and $u(t,p)$ are of class C^k on $[a,b] \times V$ $(k \geq 2)$, the following *Taylor*-expansions exist:

$$x(t,p) = x_0(t) + \frac{\partial x}{\partial p}(t,p_0)(p-p_0) + 0(\|p-p_0\|^2),$$

$$\lambda(t,p) = \lambda_0(t) + \frac{\partial \lambda}{\partial p}(t,p_0)(p-p_0) + 0(\|p-p_0\|^2),$$

$$u(t,p) = u_0(t) + \frac{\partial u}{\partial p}(t,p_0)(p-p_0) + 0(\|p-p_0\|^2).$$

The variations

$$z(t) := \frac{\partial x}{\partial p}(t,p_0), \quad \mu(t) := \frac{\partial \lambda}{\partial p}(t,p_0), \quad v(t) := \frac{\partial u}{\partial p}(t,p_0)$$

are (n,p), respectively, (m,p), matrices of class C^1 that satisfy the *linear* inhomogeneous BVP

$$\begin{aligned} \dot{z} &= A^0(t)z + B^0(t)\mu + f_p^0(t) - f_u^0(t)H_{uu}^0(t)^{-1}H_{up}^0(t), \\ \dot{\mu} &= C^0(t)z - A^0(t)^T\mu + H_{xx}^0 H_{uu}^0(t)^{-1}H_p^0(t) - H_{xp}^0(t), \end{aligned} \tag{3.8}$$

$$z(a) = \varphi_p(p_0), \qquad z(b) = \psi_p(p_0). \tag{3.9}$$

Here the upper index zero denotes arguments evaluated at $p = p_0$. This result follows by differentiating (3.6) and (3.7) with respect to p. Moreover, differentiation of the identity

$$H_u(t, x(t,p), \lambda(t,p), u(t,p), p) = 0$$

yields

$$v(t) = -H_{uu}^0(t)^{-1}(H_{ux}^0(t)z(t) + f_u^0(t)^T\mu(t) + H_{up}^0(t)). \tag{3.10}$$

The linear BVP (3.8) and (3.9) can be solved by stable shooting techniques (see [4], [5], [19]–[21], [37]–[41] for special perturbations).

4. An illustrative example. We present an example that admits two kinds of extremal solutions, both with a nonsingular shooting matrix. The sufficient conditions single out only one solution as optimal. For this solution a sensitivity analysis is performed according to Theorem 3.1. Consider the following variational problem depending on a parameter $p \in \mathbb{R}$:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\int_0^1 (px(t)^3 + \dot{x}(t)^2)dt \\ &\text{subject to} \quad x(0) = 4, \qquad x(1) = 1. \end{aligned}$$

The unperturbed problem corresponds to $p_0 = 1$. Defining as usual the control variable by $u := \dot{x}$, the *Hamiltonian* becomes

$$H(x,\lambda,u,p) = \frac{1}{2}(px^3 + u^2) + \lambda u. \tag{4.1}$$

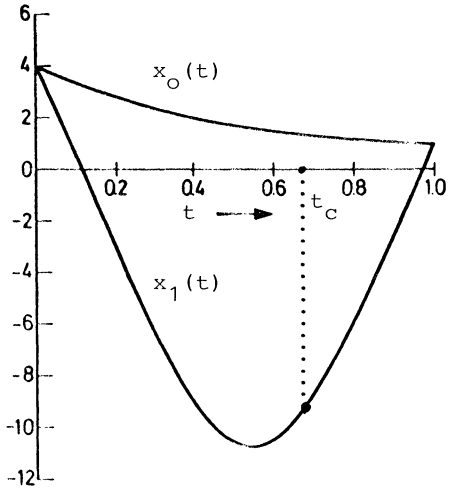


FIG. 1. *Solutions* $x_0(t)$ *and* $x_1(t)$ *of* BVP (4.2 with $p = 1$); *conjugate point* $t_c = 0.674437$ *for* $x_1(t)$.

The strict *Legendre* condition $H_{uu} = 1 > 0$ holds throughout. The function u^* in (3.5) minimizing the *Hamiltonian* is $u^*(x, \lambda, p) = -\lambda$. The *Hamiltonian* (4.1) is C^∞-regular. The BVP (3.6) and (3.7) leads to

$$\ddot{x} = \frac{3}{2} p x^2, \qquad x(0) = 4, \qquad x(1) = 1. \tag{4.2}$$

Unperturbed solution for $p_0 = 1$. Using shooting methods, Stoer and Bulirsch [45, p. 471], have shown that the BVP (4.2) with $p = 1$ has two solutions $x_0(t) = 4/(1 + t)^2$ and $x_1(t)$ characterized by

$$\dot{x}_0(0) = -8 \quad \text{and} \quad \dot{x}_1(0) = -35.85849. \tag{4.3}$$

The two solutions are shown in Fig. 1. To test $x_0(t)$ and $x_1(t)$ for optimality, we consider the following variational equation for (4.2) with respect to $x_0(t)$ or $x_1(t)$ (compare also (2.12) with initial conditions (2.18)):

$$\begin{aligned} \ddot{x}_i &= \frac{3}{2} x_i^2, \quad x_i(0) = 4, \quad \dot{x}_i(0) \quad \text{as in (4.3)}, \\ \ddot{y}_i &= 3x_i(t) y_i, \qquad y_i(0) = 0, \quad \dot{y}_i(0) = 1 \quad (i = 0, 1). \end{aligned} \tag{4.4}$$

It can be verified by numerical integration that the following classical *Jacobi* condition holds:

$$y_0(t) \neq 0 \quad \text{for } 0 < t \leq 1.$$

Hence $x_0(t)$ is optimal. Alternatively, optimality of $x_0(t)$ can be verified by invoking Theorem 2.1. The Riccati equation (2.19) becomes $\dot{Q} = -3x_0(t) + Q^2$. It is straightforward to compute a bounded solution Q in [0,1].

On the other hand, for the solution x_1 we find that

$$y_1(t_c) = 0 \quad \text{for } t_c = 0.674437.$$

Thus, there exists a point $t_c \in (0, 1)$ that is conjugate to $t = 0$. This violates the necessary condition of optimality in [51, Theorem 3.1]. Hence $x_1(t)$ is *nonoptimal*. We note that the

TABLE 1
First- and second-order Taylor approximation in (4.5), $p = 1 + \Delta p$.

$\lvert\Delta p\rvert$	$e_1(\Delta p)$	$e_1(\Delta p)/\lvert\Delta p\rvert^2$	$e_2(\Delta p)$	$e_2(\Delta p)/\lvert\Delta p\rvert^3$
0.01	$0.16 \cdot 10^{-4}$	0.16	$0.67 \cdot 10^{-7}$	0.067
0.05	0.00041	0.17	$0.85 \cdot 10^{-5}$	0.068
0.1	0.0017	0.17	$0.69 \cdot 10^{-4}$	0.069
0.3	0.017	0.18	0.0021	0.077
0.5	0.051	0.20	0.011	0.086

exact value of the conjugate point t_c can be computed via the BVP (4.4) and $y_1(t_c) = 0$ treating t_c as a *free* variable.

Perturbed solutions and neighbouring extremals. By Theorem 3.1 there exists a neighbourhood $V \subset \mathbb{R}$ of $p_0 = 1$ and a C^∞ function $x(t,p)$ for $(t,p) \in [0,1] \times V$ such that $x(\cdot,p)$ is optimal for the variational problem and satisfies $x(t,p_0) = x_0(t) = 4/(1+t)^2$.

The function $x(t,p)$ solves the BVP (4.2) and admits a *Taylor* expansion

$$x(t,p) = x_0(t) + \frac{\partial x}{\partial p}(t,p_0)(p-p_0) + \frac{1}{2}\frac{\partial^2 x}{\partial p^2}(t,p_0)(p-p_0)^2 + O(|p-p_0|^3) \tag{4.5}$$

on $[0,1] \times V$. The variations

$$z_1(t) := \frac{\partial x}{\partial p}(t,p_0), \qquad z_2(t) = \frac{\partial^2 x}{\partial p^2}(t,p_0)$$

are solutions of the linear inhomogeneous BVP

$$\ddot{z}_1 = 3x_0(t)z_1 + \frac{3}{2}x_0(t)^2, \qquad z_1(0) = z_1(1) = 0,$$

respectively,

$$\ddot{z}_2 = 3x_0(t)z_2 + 3z_1(t)(2x_0(t) + z_1(t)), \qquad z_2(0) = z_2(1) = 0,$$

which can be obtained from (4.2) by formal differentiation; compare also (3.8) and (3.9). The solutions $z_1(t)$ and $z_2(t)$ are given by

$$\dot{z}_1(0) = -3.779528, \qquad \dot{z}_2(0) = 1.483277.$$

Table 1 presents some numerical results reflecting the error of the first- and second-order Taylor expansion in (4.5) where $p = 1 + \Delta p$, $k = 1, 2$, as follows:

$$e_k(\Delta p) := \max_{0 \le t \le 1} \left| x(t,p) - \sum_{i=0}^{k} \frac{1}{i!}\frac{\partial^i x}{\partial p^i}(t,p_0)(\Delta p)^i \right|.$$

5. Conclusion and extensions. The second-order sensitivity result derived in this paper states that the optimal solution of a nonlinear control problem is differentiable with respect to parameters provided that (1) second-order sufficient conditions (SSCs) hold for the unperturbed (nominal) problem and (2) the shooting matrix associated with the boundary value problem is nonsingular. Many authors have used the nonsingularity of the shooting matrix as the only tool to obtain a differentiable family of extremals. The example in §4 demonstrates that this alone does not suffice to find an optimal solution to which perturbation analysis can be applied.

Thus, the solution differentiability result in this paper gives a firm theoretical basis to existing numerical feedback schemes for computing neighbouring extremals.

It is desirable to extend the solution differentiability to perturbed nonlinear control problems with inequality constraints of the type

$$\begin{aligned} \text{mixed state-control constraints:} \quad & C(x,u,p) \le 0, \qquad C : \mathbb{R}^{n+1+k} \to \mathbb{R}, \\ \text{state constraints:} \quad & S(x,p) \le 0, \qquad S : \mathbb{R}^{n+k} \to \mathbb{R}. \end{aligned}$$

These general control constraints comprise the special case of a *closed* control set U that has not been treated in this paper. Boundary value problems for inequality constrained problems have been successfully solved by multiple shooting techniques. These techniques usually require that the structure of the solution, that is, the number and the type of junction points with the boundary, is known. Then one main obstacle to extending the techniques of this paper to inequality constraints is the fact that SSCs in [32], [44] are too strong and are not directly related to the variational system of the associated boundary value problem. SSCs that use a type of Riccati ODE related to this variational system have first been obtained by Orrell and Zeidan [35] for the state independent constraint $C(u) \le 0$. Recently, extensions of these SSCs to mixed constraints $C(x,u) \le 0$ have been derived in Maurer [33] and Pickenhain [42]. The SSCs in [33] are developed directly on the basis of the variational system to provide the ingredients for a sensitivity result along the lines of this paper. Results and examples will be reported in a future paper.

REFERENCES

[1] W. Alt, *Stability of solutions to control constrained nonlinear control problems*, Appl. Math. Optim., 21 (1980), pp. 53–68.

[2] ———, *Stability of solutions for a class of nonlinear cone constrained optimization problems, Part* I: *Basis theory*, Numer. Funct. Anal. Optim., 10 (1989), pp. 1053–1064.

[3] ———, *Parametric optimization with applications to optimal control and sequential quadratic programming*, Bayreuther Math. Schriften, Heft 35 (1991), pp. 1–37.

[4] H. G. Bock, *Zur numerischen Behandlung zustandsbeschränkter Steuerungsprobleme mit Mehrzielmethode und Homotopieverfahren*, ZAMM, 57 (1977), pp. T266–T268.

[5] H. G. Bock and P. Krämer-eis, *An efficient algorithm for approximate computation of feedback control laws in nonlinear processes*, ZAMM, 61 (1981), pp. T330–T332.

[6] J. V. Breakwell and Y. C. Ho, *On the conjugate point condition for the control problem*, Internat. J. Engrg. Sci., 2 (1965), pp. 565–579.

[7] J. V. Breakwell, J. L. Speyer, and A. E. Bryson, *Optimization and control of nonlinear systems using the second variation*, SIAM J. Control, 1 (1963), pp. 193–223.

[8] A. E. Bryson and Y. C. Ho, *Applied Optimal Control*, Ginn, Waltham, MA, 1969.

[9] R. Bulirsch, *Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung*, Report of the Carl–Cranz Gesellschaft, Oberpfaffenhofen, 1971.

[10] F. Colonius and K. Kunisch, *Sensitivity analysis for optimization problems in Hilbert spaces with bilateral constraints*, Schwerpunkt programm der Deutschen Forschungsgemeinschaft "Anwendungsbezogene Optimierung und Steuerung," Report No. 267, 1991.

[11] I. B. Cruz, Jr. and W. R. Perkins, *A new approach to the sensitivity problem in multivariable feedback system design*, IEEE Trans. Automat. Control, AC-9 (1964), pp. 216–233.

[12] T. S. Dillon and T. Tun, *Application of sensitivity methods to the problem of optimal control of hydrothermal power systems*, Optim. Control Appl. Meth., 2 (1981), pp. 117–143.

[13] A. L. Dontchev, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Information Sciences, Vol. 52, Springer-Verlag, Berlin, 1983.

[14] P. Dorato, *On sensitivity in optimal control systems*, IEEE Trans. Automat. Control, AC-8 (1963), pp. 256–257.

[15] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.

[16] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum: Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.

[17] H. J. Kelley, *Guidance theory and extremal fields*, IEEE Trans. on Automat. Control, AC-7 (1962), pp. 75–82.

[18] ———, *An optimal guidance approximation theory*, IEEE Trans. Automat. Control, AC-7 (1964), pp. 375–380.

[19] P. KRÄMER-EIS, *Ein Mehrzielverfahren zur numerischen Berechnung optimaler Feedback-Steuerungen bei beschränkten nichtlinearen Steuerungsproblemen*, Bonner Math. Schriften, 164 (1985).

[20] B. KUGELMANN AND H. J. PESCH, *A new general guidance method in constrained optimal control, part 1: Numerical method*, J. Optim. Theory Appl., 67 (1990), pp. 421–435.

[21] ———, *A new general guidance method in constrained optimal control, part 2: Application to space shuttle guidance*, J. Optim. Theory Appl., 67 (1990), pp. 437–446.

[22] I. LEE, *Optimal trajectory, guidance, and conjugate points*, Inform. and Control, 8 (1965), pp. 589–606.

[23] A. Y. LEE AND A. E. BRYSON, JR., *Neighbouring extremals of dynamic optimization problems with parameter variations*, Optim. Control Appl. Meth., 10 (1989), pp. 39–52.

[24] K. MALANOWSKI, *Differential stability of solutions to convex, control constrained optimal control problems*, Appl. Math. Optim., 12 (1984), pp. 1–14.

[25] ———, *On differentiability with respect to a parameter of solutions to convex optimal control problems subject to state space constraints*, Appl. Math. Optim., 12 (1984), pp. 231–245.

[26] ———, *Stability and sensitivity of solutions to optimal control problems for systems with control appearing linearly*, Appl. Math. Optim., 16 (1987), pp. 73–91.

[27] ———, *Sensitivity analysis of optimization problems in Hilbert space with application to optimal control*, Appl. Math. Optim., 21 (1990), pp. 1–20.

[28] ———, *Second-order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.

[29] ———, *Two norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.

[30] ———, *Stability and sensitivity of solutions to nonlinear optimal control problems*, Systems Research Institute, Polish Academy of Sciences, Warszawa, 1991, preprint.

[31] H. MAURER, *Numerical solution of singular control problems using multiple shooting techniques*, J. Optim. Theory Appl. 18 (1976), pp. 235–257.

[32] H. MAURER, *First- and second-order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Stud., 14 (1981), pp. 163–177.

[33] H. MAURER, *The two-norm approach for second-order sufficient conditions in mathematical programming and optimal control*, "Angewandte Mathematik und Informatik" der Universität Münster, Report No. 6/92-N, 1992, preprint.

[34] H. MAURER AND H. J. PESCH, *Solution differentiability for nonlinear parametric control problems*, Schwerpunktprogramm der Deutschen Forschungsgemeinschaft "Anwendungsbezogene Optimierung und Steuerung", Report No. 316, 1991.

[35] D. ORRELL AND V. ZEIDAN, *Another Jacobi sufficiency criterion for optimal control with smooth constraints*, J. Optim. Theory Appl., 58 (1988), pp. 283–300.

[36] B. PAGUREK, *Sensitivity of the performance of optimal control systems to plant parameter variations*, IEEE Trans. Automat. Control, AC-10 (1965), pp. 178–180.

[37] H. J. PESCH, *Echtzeitberechnung fastoptimaler Rückkopplungssteue-rungen bei Steuerungsproblemen mit Beschränkungen*, Munich University of Technology, Habilitationsschrift, 1986.

[38] ———, *Real-time computation of feedback controls for constrained optimal control problems, Part 1: Neighbouring extremals*, Optim. Control Appl. Meth., 10 (1989), pp. 129–145.

[39] ———, *Real-time computation of feedback controls for constrained optimal control problems, Part 2: A correction method based on multiple shooting*, Optim. Control Appl. Meth., 10 (1989), pp. 147–171.

[40] ———, *Optimal control problems under disturbances*, in Proc. 14th IFIP Conference on System Modelling and Optimization, H.-J. Sebastian and K. Tammer, eds., Leipzig, 1989 Springer, Berlin; Lecture Notes in Control and Information Science, 143 (1990), pp. 377–386.

[41] ———, *Optimal and nearly optimal guidance by multiple shooting*, in Proc. Internat. Symp. Mécanique Spatiale-Space Dynamics, Centre National d'Etudes Spatiales, ed., Toulouse, 1989, Cepadus Editions, Toulouse, 1990, pp. 761–771.

[42] S. PICKENHAIN, *Sufficiency conditions for weak local minima in multidimensional optimal control problems with mixed control-state restrictions*, Z. Anal. Anwendungen, 11 (1992), pp. 559–568.

[43] W. T. REID, *Riccati Differential Equations*, Mathematics in Science and Engineering, Vol. 86, Academic Press, New York, 1972.

[44] G. SORGER, *Sufficient optimality conditions for nonconvex problems with state constraints*, J. Optim. Theory Appl., 62 (1989), pp. 289–310.

[45] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer, New York, 1980.

[46] A. P. WIERZBICKI AND S. KURCYUSZ, *Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space*, SIAM J. Control Optim., 15 (1977), pp. 25–56.

[47] H. S. WITSENHAUSEN, *On the sensitivity of optimal control systems*, IEEE Trans. Automat. Control, AC-10 (1965), pp. 495–496.

[48] V. ZEIDAN, *Sufficient conditions for the generalized problem of Bolza*, Trans. Amer. Math. Soc., 275 (1983), pp. 561–586.

[49] ———, *Sufficiency conditions with minimal regularity assumption*, Appl. Math. Optim., 20 (1989), pp. 19–31.
[50] ———, *Sufficiency criteria via focal points and via coupled points*, SIAM J. Control Optim., 30 (1992), pp. 82–98.
[51] V. ZEIDAN AND P. ZEZZA, *Necessary conditions for optimal control problems: Conjugate points*, SIAM J. Control Optim., 20 (1988), pp 592–608.
[52] ———, *The conjugate point condition for smooth control sets*, J. Math. Anal. Appl., 132 (1988), pp. 572–589.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1555–1558, November 1994

003

COMMENTS ON A STRUCTURAL APPROACH TO THE NONLINEAR MODEL MATCHING PROBLEM*

Ü. KOTTA†

Abstract. It is shown that, under certain regularity assumptions, the sufficient conditions for solvability of the model matching problem (MMP) in terms of structural invariants presented by Moog, Perdon, and Conte [*SIAM J. Control Optim.*, 29 (1991), pp. 769–785] are also necessary. The seeming controversy involving the example of Huijberts is resolved.

Key words. nonlinear system, model matching, structure algorithm, regularity, dynamic precompensation

AMS subject classifications. 93B50, 93C10, 93B52

1. Introduction.

In [3] the model matching problem (MMP) for affine nonlinear systems has been investigated using an approach based on the structure (inversion) algorithm [4]. Sufficient conditions for the solvability of the problem have been found in terms of the structural invariants: the MMP admits a solution if the structure at infinity of the system and that of the so-called extended system (i.e., a suitable composition of the system and the model) are equal. It has been claimed that the proposed conditions are not necessary and an example of Huijberts [1] has been presented to confirm the argument. Although not stated explicitly, to obtain the equations of the compensator, the authors of [3] had to work under certain regularity assumptions about the triplet (state, disturbance, output).

The purpose of the present paper is to show that under slightly stronger regularity assumptions, these conditions are also necessary and to resolve the seeming controversy with the example of Huijberts.

2. Main result.

Consider a nonlinear plant G, described by the equations

$$\begin{cases} \dot{z} = f_G(z) + g_G(z)v, & z(0) = z_0, \\ y_G = h_G(z), \end{cases} \tag{2.1}$$

where the state $z \in R^n$, the input $v \in R^m$, the output $y_G \in R^p$, $f_G(\cdot)$, and the columns of g_G and h_G are meromorphic functions of z.

Furthermore, let a nonlinear model T be given, which is described by the equations

$$\begin{cases} \dot{x} = f(x) + g(x)u, & x(0) = x_0, \\ y_T = h(x), \end{cases} \tag{2.2}$$

where the state $x \in R^{n_T}$, the input $u \in R^{m_T}$, the output $y_T \in R^p$, $f(\cdot)$, and the columns of $g(\cdot)$ and $h(\cdot)$ are meromorphic functions of x.

An extended system GT can be associated with the plant G and the model T as follows:

$$\begin{cases} \dot{x} = f(x) + g(x)u, \\ \dot{z} = f_G(z) + g_G(z)v, \\ y_{GT} = h(x) - h_G(z), \end{cases} \tag{2.3}$$

with the state $x_{GT} = (x^T, z^T)^T$, the control v, the measurable input disturbance u, and the output y_{GT}.

* Received by the editors May 28, 1992; accepted for publication (in revised form) June 21, 1993.
† Institute of Cybernetics, Akadeemia tee 21, Tallinn EE0026, Estonia.

The compensator H used to control G is a nonlinear system described by the equations of the form

$$\begin{cases} \dot{\xi} = f_H(\xi, z, u), & \xi(0) = \xi_0, \\ v = h_H(\xi, z, u), \end{cases} \tag{2.4}$$

with state $\xi \in R^q$.

The MMP in [3] has been formulated in the following way.

DEFINITION 1. *Given the plant* (2.1) *and model* (2.2), *find a compensator of the form* (2.4) *and a map* φ: $R^n \to R^q$ *such that the difference* $y_T(u, x_0) - y_{GH}(u, \varphi(x_0), z_0)$ *between the output of the model, viewed as a function of* u *and of the initial state* x_0, *and the output* y_{GT} *of the closed-loop system* (2.1), (2.4), *viewed as a function of* u *and of the initial states* z_0 *and* $\xi_0 = \varphi(x_0)$, *does not depend on* u.

The result of [3] on the MMP is based on the inversion (structure) algorithm. Actually, two versions of the inversion algorithm for systems with two types of inputs—controls and disturbances—are considered. In the first version (denoted by Singh), inversion is accomplished with regard to both types of inputs, controls and disturbances, whereas in the second version (denoted by Singh$_v$) the disturbances are considered as system parameters and inversion is accomplished with regard to control inputs only. We do not present the inversion algorithm here, and in the following, we borrow the notations for vectors, functions, etc., appearing in the inversion algorithm from [3].

Moog, Perdon, and Conte [3] presented sufficient conditions for the solvability of the MMP in terms of the structure at infinity ρ_k, $k \geq 1$ of the original system G and the extended system GT. They stated that the problem is solvable if the following structure at infinity of G and GT are equal:

$$\rho_k(GT) = \rho_k(G), \qquad k \geq 1.$$

In the sequel we need the two notions of regularity for GT associated with the two versions of the inversion algorithm, respectively.

Denote by T^nU_{GT} the nth-order tangent bundle of the input manifold $U_{GT} = R^{m_T}$ of the extended system GT, and denote by T^nY_{GT} the nth-order tangent bundle of the output manifold $Y_{GT} = R^P$ of the extended system GT.

DEFINITION 2. *Let a point* $x_{GT,0} = (x_0, z_0) \in R^{n_T} \times R^n$, *a point* $\tau_{GT,0} = (u_0, u_0^{(1)}, \ldots, u_0^{(n)}) \in T^nU_{GT}$, *and a point* $\gamma_{GT,0} = (y_{GT,0}, y_{GT,0}^{(1)}, \ldots, y_{GT,0}^{(n)}) \in T^nY_{GT}$ *be given. We call* $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$ *a locally strongly regular triplet for* GT *with respect to Singh (Singh$_v$) if, for each application of the algorithm,* rank $G_k(\cdot)$ $1 \leq k \leq n + n_T$ *is constant around a triplet* $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$.

Moog, Perdon, and Conte did not state explicitly that they work around a locally strongly regular triplet for the extended system GT with respect to Singh$_v$. However, implicitly they assumed it because otherwise it is not possible to solve the equations, obtained at the last step of the inversion algorithm, to find the compensator. Taking this observation into account, we can reformulate the result of [3] in the following way.

THEOREM 1. *Consider a nonlinear plant* G *and a nonlinear model* T. *Assume that* $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$ *is a strongly regular triplet for the extended system* GT *with respect to Singh$_v$. The* MMP *is locally solvable around* $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$ if

$$\rho_k(GT) = \rho_k(G), \qquad 1 \leq k \leq n + n_T. \tag{2.5}$$

Of course, if we do not work around a locally strongly regular triplet with respect to

Singh_v, then the conditions (2.5) are no longer sufficient. Take the extremely simple example

$$G = \begin{cases} \dot{x} = vz, \\ y_G = z, \end{cases} \qquad T = \begin{cases} \dot{x} = u, \\ y_T = x. \end{cases}$$

Condition (2.5) for G and T is clearly verified since $y_{GT} = vz - u$, and hence $\rho_i(GT) = \rho_i(G)$, $i = 1, 2$. Nevertheless, around the nonregular triplet defined by $z_0 = 0$, the MMP is not solvable.

Now we are going to prove that under slightly stronger regularity assumptions about the triplet, the conditions of (2.5) are also necessary for the solvability of the MMP.

THEOREM 2. *Consider a nonlinear plant G and a nonlinear model T. Assume that $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$ is a strongly regular triplet for the extended system GT with respect to both versions of the inversion algorithm. The* MMP *is locally solvable around $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$ only if*

$$\rho_k(GT) = \rho_k(G), \qquad 1 \le k \le n + n_T.$$

Proof. Let us assume that there exists a precompensator H of the form (2.4) for G and T that, around the strongly regular triplet $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$, locally solves the MMP. Apply the first step of the inversion algorithm to GT with respect to control v only, considering disturbances u as parameters, to obtain

$$\begin{cases} \tilde{y}_1 = \tilde{f}_1(x, z, u) + \tilde{g}_1(x, z)v, \\ \hat{y}_1 = \hat{y}_1(x, z, u, \tilde{y}_1), \end{cases} \tag{2.6}$$

where $\operatorname{rank} \tilde{G}_1(x, z) := \operatorname{rank} \tilde{g}_1(x, z) = \rho_{1v}$.

If we plug the output of H in (2.6), the equations no longer depend on u, since H solves the MMP for G and T. In particular, this means that either

$$\frac{\partial \hat{y}_1}{\partial u} = \gamma(x, z, \tilde{y}_1) \equiv \dot{0} \tag{2.7}$$

everywhere around the triplet $(x_{GT,0}, \tau_{GT,0}, \gamma_{GT,0})$, or the compensator H will guarantee the equality (2.7). Note that around the regular triplet with respect to Singh, $\partial \hat{y}_1 / \partial u$ is everywhere either equal to zero or different from zero. This means that if $\partial \hat{y}_1 / \partial u \not\equiv 0$, we can never make it equal to zero by a suitable choice of the compensator. This implies that $\partial \hat{y}_1 / \partial u \equiv 0$, which gives us

$$\rho_1 = \operatorname{rank} \begin{bmatrix} \tilde{G}_1(\cdot) & \partial \tilde{f}_1 / \partial u \\ 0 & \partial \hat{y}_1 / \partial u \end{bmatrix} = \operatorname{rank} \tilde{G}_1(\cdot) = \rho_{1v}.$$

Applying this argument repeatedly, we finally obtain $\rho_k(GT) = \rho_{kv}(GT)$, $k \ge 1$. The conclusion of Theorem 2 follows using the fact that $\rho_k(G) = \rho_{kv}(GT)$, $k \ge 1$ [3].

Of course, around a nonregular triplet for the system GT with respect to Singh_v it is sometimes possible to make $\partial \hat{y}_1 / \partial u$ equal to zero by the proper choice of H. Such a choice will exploit the nonregularity of the triplet around which we work; namely, with this choice, we fall into a nonregular triplet. Note that $\rho_k(GT)$, evaluated at this point, is less than optimal, and exactly equal to $\rho_{kv}(GT)$.

Now we are ready to solve the seeming controversy of Theorem 2 with the example due to Huijberts [1] (and also presented in [3]),

$$G = \begin{cases} \dot{z} = (z_2 + z_2 v_2, v_1, v_2)^T, \\ y_G = (z_2, z_3, z_1)^T, \end{cases} \qquad T = \begin{cases} \dot{x} = (x_2, x_3 + u_1, u_2, x_4)^T, \\ y_T = (x_2, x_4, x_1)^T. \end{cases}$$

By applying the inversion algorithm to G, we obtain $\rho_1(G) = 2$, $\rho_2(G) = \rho_3(G) = 0$. The same procedure applied to GT gives

$$\left\{ \begin{array}{l} \dot{y}_{GT,1} = v_1 - x_3 - u_1, \\ \dot{y}_{GT,2} = v_2 - x_4, \\ \dot{y}_{GT,3} = z_2(1 + \dot{y}_{GT,2} + x_4) - x_2, \end{array} \right. \tag{2.8}$$

and thus

$$\ddot{y}_{GT,3} = u_1(\dot{y}_{GT,2} + x_4) + (\dot{y}_{GT,1} + x_3)(\dot{y}_{GT,2} + x_4) + z_2(\dot{y}_{GT,2} + x_4) + \dot{y}_{GT,1}.$$

So $\rho_1(GT) = s_1(GT) = 2$,

$$s_2(GT) = \operatorname{rank} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & \dot{y}_{GT,2} + x_4 \end{pmatrix} = 3,$$

and $\rho_2(GT) = s_2(GT) - s_1(GT) = 1$.

In [1], [3] it has been claimed that the condition $\rho_2(GT) = \rho_2(G)$ is not satisfied, but the compensator

$$\left\{ \begin{array}{l} \dot{\xi} = (\xi_2, \xi_3 + u_1, u_2, -\xi_4)^T, \\ v = (\xi_3 + u_1, 0)^T \end{array} \right. \tag{2.9}$$

still solves the MMP. What actually happens with the choice of the compensator (2.9) is that we fall into the nonregular triplet (for the extended system GT with respect to Singh), defined by $\dot{y}_{GT,2} + x_4 = 0$ (see (2.8)). This triplet is nonregular, because $\rho_2(GT) = 1$ everywhere except for the case where $\dot{y}_{GT,2} + x_4 = 0$; in this case, $\rho_2(GT) = 0$. So Theorem 2 does not state anything about the solvability of the MMP around this triplet.

3. Conclusions. Around a regular triplet for the extended system with respect to both versions of the inversion algorithm, the conditions (2.5) are sufficient as well as necessary. Hence around a regular triplet, the necessary and sufficient conditions for solvability of the nonlinear MMP are in accordance with the corresponding conditions for linear systems [2], and a solution of the nonlinear MMP naturally extends the solution of the MMP for linear systems.

Of course, this similarity no longer holds around a nonregular triplet, where the conditions (2.5) are neither necessary (see the example by Huijberts) nor sufficient. In other words, around a nonregular triplet with respect to Singh_v, the nonlinear MMP is not solvable even if the condition $\rho_k(GT) = \rho_k(G)$, $1 \leq k \leq n + n_T$ holds. And in spite of this, that for some j, $\rho_j(GT) \neq \rho_j(G)$, the problem may still be solvable if we work around a nonregular triplet with respect to Singh.

REFERENCES

[1] H. J. C. Huijberts, *A nonregular solution of the nonlinear dynamic disturbance decoupling problem with an application to a complete solution of the nonlinear model matching problem*, SIAM J. Control Optim., 30 (1992), pp. 350–366.

[2] M. Malabre, *Structure a l'infini des triplets invariants. Application a la poursuite parfaite de modele*, Lecture Notes Control Inf. Sci., 44 (1982), pp. 43–53.

[3] C. H. Moog, A. M. Perdon, and G. Conte, *Model matching and factorization for nonlinear systems: A structural approach*, SIAM J. Control Optim., 29 (1991), pp. 769–785.

[4] S. N. Singh, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 595–598.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1559–1576, November 1994

004

RECURSIVE ALGORITHMS FOR SOLVING A CLASS OF NONLINEAR MATRIX EQUATIONS WITH APPLICATIONS TO CERTAIN SENSITIVITY OPTIMIZATION PROBLEMS*

WEI-YONG YAN†, JOHN B. MOORE‡, AND UWE HELMKE§

Abstract. This paper is concerned with solving a class of nonlinear algebraic matrix equations. Two recursive algorithms are proposed in terms of matrix *difference* equations and are studied. A set of initial values is characterized, from which the convergence of the algorithms can be guaranteed.

Based on the general results, several effective algorithms are presented to compute L^2-sensitivity optimal realizations, as well as Euclidean norm balancing realizations, of a given linear system. A locally exponential convergence property is proved for one of them. As is shown by simulation in this paper, these algorithms prove to be far more practical for digital computer implementation than the gradient flows previously proposed.

Key words. matrix equations, difference equations, linear systems, sensitivity, minimal realizations

AMS subject classifications. 93B40, 15A24, 39A10

1. Introduction.

Consider the algebraic matrix equation of the form

$$\mathcal{F}(X) - X\mathcal{G}(X)X = 0, \qquad X \in \mathcal{P}(n), \tag{1.1}$$

with $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ continuous operators from $\mathcal{P}(n)$ to itself, where $\mathcal{P}(n)$ denotes the set of all positive definite symmetric $n \times n$ matrices. In this paper we are interested in finding the solution to (1.1) under the following basic assumptions:

Assumption 1. $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)^{-1}$ are nondecreasing, that is, for any $X_1, X_2 \in \mathcal{P}(n)$ with $X_2 \geq X_1$, there hold $\mathcal{F}(X_2) \geq \mathcal{F}(X_1)$ and $\mathcal{G}(X_2) \leq \mathcal{G}(X_1)$, where the notation $X_2 \geq X_1$ ($X_2 > X_1$) means that $X_2 - X_1$ is positive semidefinite (definite).

Assumption 2. Equation (1.1) has a unique solution $\bar{X}$ in $\mathcal{P}(n)$.

This type of matrix equation often arises in systems and control. For example, it has been recently found [1], [3] that solving the problems of L^2-sensitivity minimization and Euclidean norm balancing can be reduced to solving certain highly nonlinear equations of the form (1.1). Unfortunately, there is no explicit formula for their unique solution to these algebraic equations. The only computation method available to date is to solve certain related nonlinear *differential* equations. For high-order problems, this method may be impractical or inefficient using conventional digital computers. Therefore, it is desirable to develop suitable iterative algorithms in terms of *difference* equations whose solution can converge to the required solution from appropriate initial conditions.

Of course, it is not always possible to relate (1.1) to some optimization problem or differential equation so that certain numerical methods such as Euler approximation and Newton–Raphson can be applied. Moreover, even if possible, the existing numerical methods may not always work well and may be inefficient because their success heavily depends on intricate step size adjustment, which is sometimes time consuming. In contrast, the method to be proposed in this paper for solving (1.1) does not require any step size adjustment, as will be seen soon.

* Received by the editors March 2, 1992; accepted for publication (in revised form) June 23, 1993. This work was partially supported by the Boeing Commercial Aircraft Corporation (BCAC).

† Department of Mathematics, University of Western Australia, Nedlands, Western Australia 6009, Australia.

‡ Department of Systems Engineering, Research School of Physical Sciences and Engineering, Australian National University, GPO Box 4, Canberra, Australian Capital Territory 2601, Australia.

§ Department of Mathematics, University of Regensburg, 8400 Regensburg, Germany.

To suggest a workable algorithm for (1.1), let us consider the trivial case where $\mathcal{F}(P) = F \in \mathcal{P}(n)$ and $\mathcal{G}(P) = G \in \mathcal{P}(n)$, that is, the operators $\mathcal{F}$ and $\mathcal{G}$ are constant. Quite obviously, Assumptions 1 and 2 are fulfilled in this case. Equation (1.1) then reduces to

$$F - XGX = 0, \tag{1.2}$$

which is apparently a very special form of the algebraic Riccati equation in continuous time. Thus, by the bilinear transformation given in [2], (1.2) is equivalent to the following algebraic Riccati equation in discrete time:

$$X = X - 2(X + F)(2X + F + G^{-1})^{-1}(X + F) + 2F. \tag{1.3}$$

It is known from [2] that the solution of the Riccati difference equation

$$X_{i+1} = X_i - 2(X_i + F)(2X_i + F + G^{-1})^{-1}(X_i + F) + 2F \tag{1.4}$$

converges to the solution of (1.2) or (1.3) from any initial condition $X_0 \in \mathcal{P}(n)$.

The above-mentioned fact inspires us to come up with the following difference equation for the general purpose:

$$X_{i+1} = X_i - 2[X_i + \mathcal{F}(X_i)][2X_i + \mathcal{F}(X_i) + \mathcal{G}(X_i)^{-1}]^{-1}[X_i + \mathcal{F}(X_i)] + 2\mathcal{F}(X_i), \tag{1.5}$$

which is obtained by respective substitution of the operators $\mathcal{F}$ and $\mathcal{G}$ for F and G into (1.4). A natural question arises as to whether the solution of (1.5) can still converge to the solution of (1.1) from any initial condition $X_0 \in \mathcal{P}(n)$ in the general case. In particular, can (1.5) serve as an iterative algorithm in the practical cases of interest?

Remark 1.1. It is worth emphasizing that the operators $\mathcal{F}$ and $\mathcal{G}$ will not be required to be smooth in the development to follow. We hope that this would widen the potential applications of the algorithms to be developed in the paper.

Remark 1.2. Note that computing the value of the operators $\mathcal{F}$ and $\mathcal{G}$ is required at each iteration of the algorithm (1.5), which may be undesirable or difficult in the situation where $\mathcal{F}$ and $\mathcal{G}$ are complicated or there are even no explicit expressions for them. As will be seen in the sequel, this difficulty can be overcome by way of incorporating two additional difference equations with (1.5).

Remark 1.3. If the operators $\mathcal{F}$ and $\mathcal{G}$ satisfy differentiability conditions, homotopy methods might be used to find the solution of (1.1). However, this kind of method eventually rests in solving a differential equation [4]. Moreover, its success crucially depends on the construction of a homotopy map satisfying certain requirements; otherwise, there is no guarantee that the method is globally convergent (see, e.g., [4]).

In the next section we prove some auxiliary results. Section 3 is devoted to studying the convergence properties of two types of general nonlinear difference equations including (1.5). Section 4 discusses specific iterative schemes for solving (1.1) in the cases of L^2-sensitivity minimization and Euclidean norm balancing. Section 5 proves that the convergence of the proposed algorithm for the L^2-sensitivity minimization problem is locally exponential, and §6 presents some simulation results. Conclusions appear in §7.

2. Preliminary results. Define

$$\mathcal{R}(X) \triangleq X - 2[X + \mathcal{F}(X)][2X + \mathcal{F}(X) + \mathcal{G}(X)^{-1}]^{-1}[X + \mathcal{F}(X)] + 2\mathcal{F}(X) \tag{2.1}$$

for $X \in \mathcal{P}(n)$, where $\mathcal{F}(\cdot), \mathcal{G}(\cdot) : \mathcal{P}(n) \mapsto \mathcal{P}(n)$. The following fundamental properties of $\mathcal{R}(\cdot)$ are instrumental in discussing the convergence of the algorithm $X_{i+1} = \mathcal{R}(X_i)$ subsequently.

LEMMA 2.1. *Suppose $\mathcal{F}(X)$ and $\mathcal{G}(X)^{-1}$ are nondecreasing with respect to $X \in \mathcal{P}(n)$. Then $\mathcal{R}(\cdot)$ maps $\mathcal{P}(n)$ to itself; moreover, for any $X, Y \in \mathcal{P}(n)$ there hold*

$$X \geq Y \Longrightarrow \mathcal{R}(X) \geq \mathcal{R}(Y), \tag{2.2}$$

$$\mathcal{R}(X) \geq X \Longleftrightarrow \mathcal{F}(X) \geq X\mathcal{G}(X)X, \tag{2.3}$$

$$\mathcal{R}(X) \leq X \Longleftrightarrow \mathcal{F}(X) \leq X\mathcal{G}(X)X, \tag{2.4}$$

$$\mathcal{R}(X) = X \Longleftrightarrow \mathcal{F}(X) = X\mathcal{G}(X)X. \tag{2.5}$$

Proof. Upon noting from the matrix inversion formula that

$$\mathcal{R}(X) = 2\{[X + \mathcal{F}(X)]^{-1} + [X + \mathcal{G}(X)^{-1}]^{-1}\}^{-1} - X, \tag{2.6}$$

it follows that $X \in \mathcal{P}(n)$ implies $\mathcal{R}(X) \in \mathcal{P}(n)$. Now assume $X \geq Y$. Because $\mathcal{F}(X)$ and $\mathcal{G}(X)^{-1}$ are nondecreasing, we obtain

$$\begin{aligned}
\mathcal{R}(X) &\geq 2\{[X + \mathcal{F}(Y)]^{-1} + [X + \mathcal{G}(Y)^{-1}]^{-1}\}^{-1} - X \\
&= X + 2\mathcal{F}(Y) - 2[X + \mathcal{F}(Y)][2X + \mathcal{F}(Y) + \mathcal{G}(Y)^{-1}]^{-1}[X + \mathcal{F}(Y)] \\
&= \mathcal{F}(Y) + [\mathcal{G}(Y)^{-1} - \mathcal{F}(Y)][2X + \mathcal{F}(Y) + \mathcal{G}(Y)^{-1}]^{-1}[X + \mathcal{F}(Y)] \\
&= \frac{1}{2}[\mathcal{F}(Y) + \mathcal{G}(Y)^{-1}] - \frac{1}{2}[\mathcal{G}(Y)^{-1} - \mathcal{F}(Y)] \\
&\quad \cdot [2X + \mathcal{F}(Y) + \mathcal{G}(Y)^{-1}]^{-1}[\mathcal{G}(Y)^{-1} - \mathcal{F}(Y)] \\
&\geq \frac{1}{2}[\mathcal{F}(Y) + \mathcal{G}(Y)^{-1}] - \frac{1}{2}[\mathcal{G}(Y)^{-1} - \mathcal{F}(Y)] \\
&\quad \cdot [2Y + \mathcal{F}(Y) + \mathcal{G}(Y)^{-1}]^{-1}[\mathcal{G}(Y)^{-1} - \mathcal{F}(Y)] \\
&= \mathcal{R}(Y),
\end{aligned}$$

implying that (2.2) holds. Further, from the identity

$$\mathcal{R}(X) = X + 2\mathcal{F}(X) - 2\{[X + \mathcal{F}(X)]^{-1} + [X + \mathcal{F}(X)]^{-1}[X + \mathcal{G}(X)^{-1}][X + \mathcal{F}(X)]^{-1}\}^{-1} \tag{2.7}$$

it can be seen that

$$\begin{aligned}
&\mathcal{R}(X) \geq X \\
&\quad \Longleftrightarrow \mathcal{F}(X)^{-1} \leq [X + \mathcal{F}(X)]^{-1} + [X + \mathcal{F}(X)]^{-1}[X + \mathcal{G}(X)^{-1}][X + \mathcal{F}(X)]^{-1} \\
&\quad \Longleftrightarrow [X + \mathcal{F}(X)]\mathcal{F}(X)^{-1}[X + \mathcal{F}(X)] \leq [X + \mathcal{F}(X)] + [X + \mathcal{G}(X)^{-1}] \\
&\quad \Longleftrightarrow X\mathcal{F}(X)^{-1}X \leq \mathcal{G}(X)^{-1} \\
&\quad \Longleftrightarrow \mathcal{F}(X) \geq X\mathcal{G}(X)X,
\end{aligned}$$

that is, (2.3) holds. In the same spirit, (2.4) can be proved and (2.5) is obtained as a direct result of (2.3) and (2.4). □

COROLLARY 2.1. *Let*

$$f(X, Y, Z) = X - 2(X + Y)(2X + Y + Z)^{-1}(X + Y) + 2Y \tag{2.8}$$

for $(X, Y, Z) \in \mathcal{P}(n) \times \mathcal{Q}(n) \times \mathcal{Q}(n)$, where $\mathcal{Q}(n)$ denotes the set of all $n \times n$ nonnegative definite symmetric matrices. Then

$$0 < X_1 \leq X_2, \quad 0 \leq Y_1 \leq Y_2, \quad 0 \leq Z_1 \leq Z_2 \tag{2.9}$$

imply

$$f(X_1, Y_1, Z_1) \leq f(X_2, Y_2, Z_2). \tag{2.10}$$

In particular, if $Y, Z \geq A$ for some $A \in \mathcal{P}(n)$, there holds

$$f(X, Y, Z) \geq A, \quad \forall X \in \mathcal{P}(n). \tag{2.11}$$

Proof. From (2.2) of Lemma 2.1 and (2.6) it follows that

$$\begin{aligned} f(X_1, Y_1, Z_1) &\leq f(X_2, Y_1, Z_1) \\ &= 2\{(X_2 + Y_1)^{-1} + (X_2 + Z_1)^{-1}\}^{-1} - X_2 \\ &\leq 2\{(X_2 + Y_2)^{-1} + (X_2 + Z_2)^{-1}\}^{-1} - X_2 \\ &= f(X_2, Y_2, Z_2), \end{aligned}$$

which gives (2.10). □

Remark 2.1. From (2.5), we can see that under Assumption 1 in §1 the equilibrium point of the difference equation $\mathcal{R}(X_{i+1}) = X_i$ exactly equals the solution of (1.1).

The next two results are only used in developing an efficient iterative scheme for finding L^2-sensitivity optimal realizations.

LEMMA 2.2. *Given a minimal realization (A, b, c) with the eigenvalues of A being in the open unit disk, let $U(\mu)$ be the solution of the following Lyapunov equation:*

$$Q = \begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix} Q \begin{bmatrix} A & 0 \\ bc & A \end{bmatrix} + \begin{bmatrix} c'c & 0 \\ 0 & \mu I \end{bmatrix}. \tag{2.12}$$

Then there holds

$$\lim_{\mu \to \infty} U(\mu)^{-1} = 0. \tag{2.13}$$

Proof. Let $V(\mu)$ be the solution of the Lyapunov equation

$$Q = \begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix} Q \begin{bmatrix} A & 0 \\ bc & A \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mu I \end{bmatrix}. \tag{2.14}$$

Then it is quite evident that $U(\mu) \geq V(\mu) = \mu V(1)$. Because the realization (A, b, c) is minimal, it follows that the pair

$$\left(\begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix}, \begin{bmatrix} 0 \\ I \end{bmatrix} \right)$$

is controllable. Thus, $V(1)$ is positive definite, leading to $\lim_{\mu \to \infty} U(\mu)^{-1} = 0$. □

LEMMA 2.3. *Given a minimal realization (A, b, c) with the eigenvalues of A being in the open unit disk, consider the difference equation*

(2.15)

$$Q_{i+1} \triangleq \begin{bmatrix} Q_{i+1}^{11} & Q_{i+1}^{12} \\ Q_{i+1}^{21} & Q_{i+1}^{22} \end{bmatrix} = \begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix} \begin{bmatrix} Q_i^{11} & Q_i^{12} \\ Q_i^{21} & Q_i^{22} \end{bmatrix} \begin{bmatrix} A & 0 \\ bc & A \end{bmatrix} + \begin{bmatrix} c'c & 0 \\ 0 & P_i \end{bmatrix}$$

with an initial symmetric matrix Q_0.

(i) *If* P_i *is nonnegative definite for all* $i = 0, 1, \ldots$, *then there exists a constant* $\beta > 0$ *and an integer* k *such that*

$$Q_i^{11} > \beta I, \quad \forall\, i \geq k. \tag{2.16}$$

(ii) *If there hold* $P_i \geq \mu I$, $i = 0, 1, \ldots$, *for some constant* $\mu > 0$, *then there exists an integer* k *such that*

$$Q_i > U(\mu/2), \quad \forall\, i \geq k, \tag{2.17}$$

where $U(\cdot)$ *is defined as in Lemma* 2.2.

Proof. By close inspection, it can be seen that $\lim_{i\to\infty} Q_i^{11} = \bar{Q}$, where $\bar{Q}$ is the unique positive definite solution to the Lyapunov equation

$$A'QA - c'c = Q. \tag{2.18}$$

From this, (i) immediately follows.

As for (ii), note that $Q_i \geq \tilde{Q}_i$ for all $i \geq 0$, where $\tilde{Q}_i$ is the solution to

$$\tilde{Q}_{i+1} = \begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix} \tilde{Q}_i \begin{bmatrix} A & 0 \\ bc & A \end{bmatrix} + \begin{bmatrix} c'c & 0 \\ 0 & \mu I \end{bmatrix},$$

with $\tilde{Q}_0 = Q_0$. Because

$$\lim_{i\to\infty} \tilde{Q}_i = U(\mu) > U(\mu/2),$$

(ii) is concluded. □

3. General convergence analysis. Recall that the main purpose of this paper is to solve the matrix equation

$$\mathcal{F}(X) - X\mathcal{G}(X)X = 0, \qquad X \in \mathcal{P}(n), \tag{3.1}$$

with $\mathcal{F}(\cdot), \mathcal{G}(\cdot) : \mathcal{P}(n) \mapsto \mathcal{P}(n)$ continuous operators. For this purpose, we propose the following difference equation:

$$\Phi_{i+1} = \mathcal{R}(\Phi_i), \tag{3.2}$$

with

$$\mathcal{R}(X) \triangleq X - 2[X + \mathcal{F}(X)][2X + \mathcal{F}(X) + \mathcal{G}(X)^{-1}]^{-1}[X + \mathcal{F}(X)] + 2\mathcal{F}(X). \tag{3.3}$$

Our first convergence result characterizes a set of initial values for which the solution of (3.2) will converge to the unique solution of (3.1).

THEOREM 3.1. *Adopt Assumptions* 1 *and* 2 *in* §1. *Suppose there exist* $X_1, X_2 \in \mathcal{P}(n)$ *with* $X_1 \leq X_2$ *such that*

$$\mathcal{F}(X_1) \geq X_1\mathcal{G}(X_1)X_1 \quad \textit{and} \quad \mathcal{F}(X_2) \leq X_2\mathcal{G}(X_2)X_2. \tag{3.4}$$

Then the solution Φ_i *of* (3.2) *converges to the solution* $\bar{X}$ *of* (3.1) *from any initial condition* Φ_0 *satisfying*

$$X_1 \le \Phi_0 \le X_2. \tag{3.5}$$

Proof. Let $\Phi_0^{(1)} = X_1$, $\Phi_0^{(2)} = X_2$, and Φ_0 satisfy (3.5). Then it follows from (2.3)–(2.4) in Lemma 2.1 that

$$\Phi_0^{(1)} \le \mathcal{R}(\Phi_0^{(1)}) = \Phi_1^{(1)} \quad \text{and} \quad \Phi_0^{(2)} \ge \mathcal{R}(\Phi_0^{(2)}) = \Phi_1^{(2)}.$$

Thus, making use of (2.2) in Lemma 2.1 leads to

$$\Phi_0^{(1)} \le \Phi_1^{(1)} \le \Phi_1^{(2)} \le \Phi_0^{(2)}.$$

By induction, the following relation is established

$$\Phi_0^{(1)} \le \Phi_1^{(1)} \le \cdots \le \Phi_k^{(1)} \le \Phi_k^{(2)} \le \cdots \le \Phi_1^{(2)} \le \Phi_0^{(2)} \quad \forall\, k \ge 1. \tag{3.6}$$

Therefore, $\lim_{k\to\infty} \Phi_k^{(1)}$ and $\lim_{k\to\infty} \Phi_k^{(2)}$ exist and are in $\mathcal{P}(n)$ because $\{\Phi_k^{(1)}\}$ and $\{\Phi_k^{(2)}\}$ are bounded above by $\Phi_0^{(2)}$ and below by $\Phi_0^{(1)}$, respectively. Consequently, these two limits satisfy $\mathcal{R}(X) = X$ and therefore are solutions of (3.1) because of (2.5) in Lemma 2.1. By Assumption 2, it turns out that

$$\lim_{k\to\infty} \Phi_k^{(1)} = \lim_{k\to\infty} \Phi_k^{(2)} = \bar{X}. \tag{3.7}$$

Further, note that $\Phi_0^{(1)} \le \Phi_0 \le \Phi_0^{(2)}$. Hence, again by induction and using (2.2) in Lemma 2.1, there results

$$\Phi_k^{(1)} \le \Phi_k \le \Phi_k^{(2)}, \quad \forall\, k \ge 0.$$

This together with (3.7) yields that $\lim_{k\to\infty} \Phi_k = \bar{X}$. □

Remark 3.1. It is worth mentioning that Theorem 3.1 is still valid if (3.5) is replaced with the existence of an integer $N \ge 0$ such that

$$X_1 \le \Phi_N \le X_2. \tag{3.8}$$

Remark 3.2. Observe that (3.1) is equivalent to

$$\mathcal{F}_\alpha(X) - X\mathcal{G}_\alpha(X)X = 0, \qquad X \in \mathcal{P}(n), \tag{3.9}$$

where $\mathcal{F}_\alpha(X) \triangleq \mathcal{F}(X)/\alpha$ and $\mathcal{G}_\alpha(X) \triangleq \mathcal{G}(X)/\alpha$ for any $\alpha > 0$. This suggests that an infinite number of algorithms can be generated for (3.1) by substituting $\mathcal{F}_\alpha$ and $\mathcal{G}_\alpha$ for $\mathcal{F}$ and $\mathcal{G}$ and setting different values to α. Naturally, we might expect α to play some role in achieving a satisfactory convergence rate. Later simulation results do demonstrate that a suitable choice of α can significantly speed up the convergence of the algorithm.

Because the equation

$$\mathcal{G}(Y^{-1}) - Y\mathcal{F}(Y^{-1})Y = 0, \qquad Y \in \mathcal{P}(n) \tag{3.10}$$

is equivalent to (3.1) with $Y = X^{-1}$, it follows from Theorem 3.1 that the solution of the difference equation

$$\begin{aligned} \Sigma_{i+1} = {} & \Sigma_i - 2[\Sigma_i + \mathcal{G}(\Sigma_i^{-1})] \\ & \times [2\Sigma_i + \mathcal{G}(\Sigma_i^{-1}) + \mathcal{F}(\Sigma_i^{-1})^{-1}]^{-1}[\Sigma_i + \mathcal{G}(\Sigma_i^{-1})] + 2\mathcal{G}(\Sigma_i^{-1}) \end{aligned} \tag{3.11}$$

converges to $\bar{X}$ from the initial condition $\Sigma_0 = \Phi_0^{-1}$, where Φ_0 satisfies (3.5). What is more interesting is the following result.

THEOREM 3.2 *With the same hypotheses and notation as in Theorem* 3.1, *let* Σ_i *denote the solution of the difference equation* (3.11) *with the initial condition* $\Sigma_0 = \Phi_0^{-1}$. *Then there holds*

$$\Phi_i \Sigma_i = I, \qquad i = 0, 1, \ldots. \tag{3.12}$$

Proof. Simple matrix manipulations yield

$$\begin{aligned} \Phi_{i+1} &= 2\{[\Phi_i + \mathcal{F}(\Phi_i)]^{-1} + [\Phi_i + \mathcal{G}(\Phi_i)^{-1}]^{-1}\}^{-1} - \Phi_i \\ &= \{2\{[I + \mathcal{F}(\Phi_i)\Phi_i^{-1}]^{-1} + [I + \mathcal{G}(\Phi_i)^{-1}\Phi_i^{-1}]^{-1}\}^{-1} - I\}\Phi_i \\ &= \{[I + \Phi_i \mathcal{F}(\Phi_i)^{-1}]^{-1} + [I + \Phi_i \mathcal{G}(\Phi_i)]^{-1}\} \end{aligned} \tag{3.13}$$

$$\times\{[I + \mathcal{F}(\Phi_i)\Phi_i^{-1}]^{-1} + [I + \mathcal{G}(\Phi_i)^{-1}\Phi_i^{-1}]^{-1}\}^{-1}\Phi_i. \tag{3.14}$$

Similarly, we have

$$\begin{aligned} \Sigma_{i+1} = \Sigma_i\{&[I + \mathcal{G}(\Sigma_i^{-1})^{-1}\Sigma_i]^{-1} + [I + \mathcal{F}(\Sigma_i^{-1})\Sigma_i]^{-1}\} \\ &\times\{[I + \Sigma_i^{-1}\mathcal{G}(\Sigma_i^{-1})]^{-1} + [I + \Sigma_i^{-1}\mathcal{F}(\Sigma_i^{-1})^{-1}]^{-1}\}^{-1}. \end{aligned} \tag{3.15}$$

Quite obviously, $\Phi_k \Sigma_k = I$ implies $\Phi_{k+1}\Sigma_{k+1} = I$. Thus, by induction, (3.12) is proved. □

Note that implementing (3.2) requires computation of the values of the operators $\mathcal{F}$ and $\mathcal{G}$ at each iteration. However, in some situations, the operators $\mathcal{F}$ and $\mathcal{G}$ may be so complicated that evaluating them is too time consuming if not impossible, which severely diminishes the efficiency of the algorithm. On the other hand, it is sometimes possible to approximate $\mathcal{F}$ and $\mathcal{G}$ by simpler operators somehow. In this case, it does not seem unreasonable to calculate the approximate values of $\mathcal{F}$ and $\mathcal{G}$ instead at each iteration. Let us now consider this strategy and address the related convergence issue in detail.

Assume that there exist two continuous operators

$$\mathcal{S} : \mathcal{Q}(k) \times \mathcal{P}(n) \mapsto \mathcal{Q}(k) \quad \text{and} \quad \mathcal{T} : \mathcal{Q}(l) \times \mathcal{P}(n) \mapsto \mathcal{Q}(l),$$

and two constant matrices $K \in \mathbb{R}^{k\times n}$, $L \in \mathbb{R}^{l\times n}$ such that we have the following assumptions.

Assumption 3. $\mathcal{S}$ is nondecreasing with respect to each of its arguments, and $\mathcal{T}$ is nondecreasing with respect to its first argument and nonincreasing with respect to its second argument.

Assumption 4. For any given $X \in \mathcal{P}(n)$, the equations

$$U = \mathcal{S}(U, X) \quad \text{and} \quad V = \mathcal{T}(V, X) \tag{3.16}$$

have unique solutions $U(X) \in \mathcal{Q}(k)$ and $V(X) \in \mathcal{Q}(l)$, which satisfy

$$\mathcal{F}(X) = K'U(X)K \quad \text{and} \quad \mathcal{G}(X) = L'V(X)L. \tag{3.17}$$

With these assumptions, we suggest the following modified algorithm for solving (3.1):

$$\begin{aligned} \Psi_{i+1} = \Psi_i &- 2(\Psi_i + K'U_iK) \times [2\Psi_i + K'U_iK + (L'V_iL)^{-1}]^{-1} \\ &\cdot(\Psi_i + K'U_iK) + 2K'U_iK, \end{aligned} \tag{3.18}$$

$$U_{i+1} = \mathcal{S}(U_i, \Psi_i), \tag{3.19}$$

$$V_{i+1} = \mathcal{T}(V_i, \Psi_i). \tag{3.20}$$

Its convergence property is stated below.

THEOREM 3.3. *Consider the system of difference equations* (3.18)–(3.20) *with the initial condition* $(P_0, U_0, V_0) \in \mathcal{P}(n) \times \mathcal{Q}(k) \times \mathcal{Q}(l)$. *Let Assumptions* 2–4 *be enforced. Suppose that there exist* X_1, $X_2 \in \mathcal{P}(n)$ *with* $X_1 \leq X_2$ *such that* (3.4) *is met. If there exists an integer* $N \geq 0$ *such that*

$$X_1 \leq \Psi_N \leq X_2, \quad U(X_1) \leq U_N \leq U(X_2), \quad V(X_2) \leq V_N \leq V(X_1), \tag{3.21}$$

there holds

$$\lim_{i \to \infty} (\Psi_i, U_i, V_i) = (\bar{X}, U(\bar{X}), V(\bar{X})). \tag{3.22}$$

Proof. Without loss of generality we assume $N = 0$. Put

$$(\Psi_0^{(1)}, U_0^{(1)}, V_0^{(1)}) \triangleq (X_1, U(X_1), V(X_1)) \text{ and } (\Psi_0^{(2)}, U_0^{(2)}, V_0^{(2)}) \triangleq (X_2, U(X_2), V(X_2)).$$

Then it is obvious that

$$U_0^{(2)} = \mathcal{S}(U_0^{(2)}, \Psi_0^{(2)}) = U_1^{(2)} \quad \text{and} \quad V_0^{(2)} = \mathcal{T}(V_0^{(2)}, \Psi_0^{(2)}) = V_1^{(2)}.$$

By (2.4) of Lemma 2.1, it follows from the second inequality of (3.4) that

$$\Psi_1^{(2)} = \mathcal{R}(\Psi_0^{(2)}) \leq \Psi_0^{(2)}.$$

Now assume that for some positive integer m,

$$\Psi_m^{(2)} \leq \Psi_{m-1}^{(2)}, \quad U_m^{(2)} \leq U_{m-1}^{(2)}, \quad V_m^{(2)} \geq V_{m-1}^{(2)}. \tag{3.23}$$

Then by Corollary 2.1, we have $\Psi_{m+1}^{(2)} \leq \Psi_m^{(2)}$. Moreover,

$$\begin{aligned} U_{m+1}^{(2)} &= \mathcal{S}(U_m^{(2)}, \Psi_m^{(2)}) \leq \mathcal{S}(U_{m-1}^{(2)}, \Psi_{m-1}^{(2)}) = U_m^{(2)}, \\ V_{m+1}^{(2)} &= \mathcal{T}(V_m^{(2)}, \Psi_m^{(2)}) \geq \mathcal{T}(V_{m-1}^{(2)}, \Psi_m^{(2)}) \geq V_m^{(2)}. \end{aligned}$$

Therefore by induction, (3.23) is valid for any integer $m \geq 0$. In the same manner, it can be shown that

$$\Psi_m^{(1)} \geq \Psi_{m-1}^{(1)}, \quad U_m^{(1)} \geq U_{m-1}^{(1)}, \quad V_m^{(1)} \leq V_{m-1}^{(1)}, \quad m = 1, 2, \ldots. \tag{3.24}$$

Again from Corollary 2.1 together with (3.21), it can be inductively established that

$$\Psi_i^{(1)} \leq \Psi_i \leq \Psi_i^{(2)}, \quad U_1^{(1)} \leq U_i \leq U_i^{(2)}, \quad V_i^{(2)} \leq V_i \leq V_i^{(1)}, \quad i = 0, 1, \ldots. \tag{3.25}$$

As a consequence of (3.23)–(3.25), it follows that

$$\Psi_0^{(1)} \leq \Psi_1^{(1)} \leq \cdots \leq \Psi_i^{(1)} \leq \Psi_i \leq \Psi_i^{(2)} \leq \cdots \leq \Psi_1^{(2)} \leq \Psi_0^{(2)}, \tag{3.26}$$

$$U_0^{(1)} \leq U_1^{(1)} \leq \cdots \leq U_i^{(1)} \leq U_i \leq U_i^{(2)} \leq \cdots \leq U_2^{(1)} \leq U_0^{(2)}, \tag{3.27}$$

$$V_0^{(2)} \leq V_1^{(2)} \leq \cdots \leq V_i^{(2)} \leq V_i \leq V_i^{(1)} \leq \cdots \leq V_1^{(1)} \leq V_0^{(1)}, \tag{3.28}$$

which imply that the limit

$$\lim_{i\to\infty}(\Psi_i^{(j)}, U_i^{(j)}, V_i^{(j)})$$

exists and is in $\mathcal{P}(n)\times\mathcal{Q}(k)\times\mathcal{Q}(l)$ for $j=1,2$. Let $(\Psi^{(j)}, U^{(j)}, V^{(j)})$ denote the limit. By continuity of the operators $\mathcal{S}$ and $\mathcal{T}$ and invertibility of $L'V^{(j)}L$, $(\Psi^{(j)}, U^{(j)}, V^{(j)})$ is a fixed point of the system of difference equations (3.18)–(3.20). In particular, we have

$$U^{(j)} = \mathcal{S}(U^{(j)}, \Psi^{(j)}) \quad \text{and} \quad V^{(j)} = \mathcal{T}(V^{(j)}, \Psi^{(j)}) \tag{3.29}$$

which, by Assumption 4, leads to

$$\mathcal{F}(\Psi^{(j)}) = K'U^{(j)}K \quad \text{and} \quad \mathcal{G}(\Psi^{(j)}) = L'V^{(j)}L.$$

It turns out that $\Psi^{(j)} = \mathcal{R}(\Psi^{(j)})$, that is, $\Psi^{(j)}$ satisfies (3.1). By Assumption 2, there results $\Psi^{(j)} = \bar{X}$. This together with (3.29) gives rise to

$$(\Psi^{(j)}, U^{(j)}, V^{(j)}) = (\bar{X}, U(\bar{X}), V(\bar{X})), \qquad j=1,2.$$

Therefore, again from (3.26)–(3.28), (3.22) follows. □

COROLLARY 3.1. *With the same hypotheses as in Theorem* 3.1, *the solution of the second-order difference equation*

(3.30)

$$\Phi_{i+1} = \Phi_i - 2[\Phi_i + \mathcal{F}(\Phi_{i-1})][2\Phi_i + \mathcal{F}(\Phi_{i-1}) + \mathcal{G}(\Phi_{i-1})^{-1}]^{-1}[\Phi_i + \mathcal{F}(\Phi_{i-1})] + 2\mathcal{F}(\Phi_{i-1})$$

converges to the solution $\bar{X}$ *of* (3.1) *from any initial condition* $(\Phi_{-1}, \Phi_0) \in \mathcal{P}(n)\times\mathcal{P}(n)$ *satisfying*

$$X_1 \le \Phi_{-1}, \qquad \Phi_0 \le X_2, \tag{3.31}$$

with X_1, X_2 *given as in Theorem* 3.1.

Proof. Consider the system of difference equations (3.18)–(3.20) with

$$(\Psi_0, U_0, V_0) = (\Phi_0, \mathcal{F}(\Phi_{-1}), \mathcal{G}(\Phi_{-1})) \tag{3.32}$$

and

$$S(U,X) \triangleq \mathcal{F}(X), \qquad \mathcal{T}(V,X) \triangleq \mathcal{G}(X), \qquad K = L = I. \tag{3.33}$$

Then it is straightforward to check that $\Psi_i = \Phi_i$ for all $i \ge 0$, where Ψ_i denotes the first component of the solution of (3.18)–(3.20) and Φ_i is the solution of (3.30). It is also trivial to see that the operators $\mathcal{S}$ and $\mathcal{T}$ fulfill Assumptions 3 and 4. Finally, note that

$$\mathcal{F}(X_1) \le U_0 \le \mathcal{F}(X_2) \quad \text{and} \quad \mathcal{G}(X_2) \le V_0 \le \mathcal{G}(X_1). \tag{3.34}$$

Thus, by Theorem 3.3, it is concluded that

$$\lim_{i\to\infty}\Phi_i = \lim_{i\to\infty}\Psi_i = \bar{X}. \qquad \square$$

Remark 3.3. By using a similar argument, Theorem 3.1 can also be proved as a consequence of Theorem 3.3.

4. Iterative computation of L^2-sensitivity optimal realizations and Euclidean norm-balancing realizations. In this section, we apply the established general results to two specific problems in system realization theory. One problem is to find L^2-sensitivity optimal realizations of a given system and the other is to find Euclidean norm-balancing realizations. Several iterative algorithms are proposed and proved to possess the convergence property.

Now consider a discrete-time, single-input–single-output, stable system with a transfer function $H(z)$ of order n. Assume that $H(z)$ has an initial minimal realization as follows:

$$\left[\begin{array}{c|c} A & b \\ \hline c & d \end{array}\right] \triangleq c(zI - A)^{-1}b + d. \tag{4.1}$$

The L^2-sensitivity index of the system $H(z)$ with respect to the realization (A, b, c, d) is defined by

$$\Gamma_1(A, b, c) \triangleq \left\| \frac{\partial H}{\partial A} \right\|_2^2 + \left\| \frac{\partial H}{\partial b} \right\|_2^2 + \left\| \frac{\partial H}{\partial c} \right\|_2^2 \tag{4.2}$$

$$= \frac{1}{2\pi i}\text{trace}\left\{ \oint [\mathcal{A}(z)\mathcal{A}(z)^* + \mathcal{B}(z)\mathcal{B}(z)^* + \mathcal{C}(z)^*\mathcal{C}(z)]\frac{dz}{z} \right\}, \tag{4.3}$$

where

$$\mathcal{A}(z) = \left[\begin{array}{cc|c} A & bc & 0 \\ 0 & A & I \\ \hline I & 0 & 0 \end{array}\right], \quad \mathcal{B}(z) = \left[\begin{array}{c|c} A & b \\ \hline I & 0 \end{array}\right], \quad \mathcal{C}(z) = \left[\begin{array}{c|c} A & I \\ \hline c & 0 \end{array}\right]. \tag{4.4}$$

The L^2-sensitivity minimization problem is to find a similarity transformation T so that the L^2-sensitivity index $\Gamma_1(TAT^{-1}, Tb, cT^{-1})$ of $H(z)$ with respect to the transformed realization is minimized. Regarding this problem, we summarize the main known facts from [1] as follows.

Fact 1. $\Gamma_1(TAT^{-1}, Tb, cT^{-1})$ achieves its minimum at $T = T_0$ if and only if $T_0'T_0$ is an equilibrium point of the differential equation

$$\begin{aligned} \dot{P}(t) = \frac{1}{2\pi i} \oint \{ & P(t)^{-1}[\mathcal{A}(z)^*P(t)\mathcal{A}(z) + \mathcal{C}(z)^*\mathcal{C}(z)]P(t)^{-1} \\ & - \mathcal{A}(z)P(t)^{-1}\mathcal{A}(z)^* - \mathcal{B}(z)\mathcal{B}(z)^*\}\frac{dz}{z}. \end{aligned} \tag{4.5}$$

Fact 2. Equation (4.5) has a unique equilibrium $\bar{P}$ in $\mathcal{P}(n)$.

Fact 3. The solution $P(t)$ of (4.5) exponentially converges to $\bar{P}$ from any initial value $P(0) \in \mathcal{P}(n)$.

Although Fact 3 suggests that the equilibrium $\bar{P}$ can be found by solving an initial value problem associated with (4.5), this method lacks computational efficiency and can be numerically ill conditioned, especially when the order n of the system is high.

A similar situation arises when we try to minimize the Euclidean norm defined by

$$\Gamma_2(A, b, c) \triangleq \text{trace}(AA' + bb' + c'c) \tag{4.6}$$

with respect to realizations of $H(z)$. There are three analogous facts [3], but here the relevant equation is

$$\dot{P}(t) = P(t)^{-1}[A'P(t)A + c'c]P(t)^{-1} - AP(t)^{-1}A' - bb'. \tag{4.7}$$

Although (4.7) looks much simpler than (4.5), likewise a computationally attractive method to find its unique equilibrium point has not been proposed.

We are in a position to present several iterative algorithms for finding the unique solution to the matrix equation

(4.8)

$$\frac{1}{2\pi i}\oint\{P^{-1}[\mathcal{A}(z)^*P\mathcal{A}(z)+\mathcal{C}(z)^*\mathcal{C}(z)]P^{-1}-\mathcal{A}(z)P^{-1}\mathcal{A}(z)^*-\mathcal{B}(z)\mathcal{B}(z)^*\}\frac{dz}{z}=0.$$

PROPOSITION 4.1. *Given the initial minimal realization of* $H(z)$ *as in* (4.1) *with* $\bar{P}$ *denoting the solution of* (4.8), *define*

$$W_c(P)\triangleq\frac{1}{2\pi i}\oint[\mathcal{A}(z)P^{-1}\mathcal{A}(z)^*+\mathcal{B}(z)\mathcal{B}(z)^*]\frac{dz}{z}, \tag{4.9}$$

$$W_o(P)\triangleq\frac{1}{2\pi i}\oint[\mathcal{A}(z)^*P\mathcal{A}(z)+\mathcal{C}(z)^*\mathcal{C}(z)]\frac{dz}{z}. \tag{4.10}$$

Let Φ_i *be the solution of the first-order difference equation*

$$\begin{aligned}\Phi_{i+1}&=\Phi_i-2[\Phi_i+W_o(\Phi_i)/\alpha]\\&\quad\times[2\Phi_i+W_o(\Phi_i)/\alpha+\alpha W_c(\Phi_i)^{-1}]^{-1}[\Phi_i+W_o(\Phi_i)/\alpha]+2W_o(\Phi_i)/\alpha\end{aligned} \tag{4.11}$$

from an initial condition $\Phi_0\in\mathcal{P}(n)$ *and* Π_i *the solution of the second-order difference equation*

(4.12)

$$\begin{aligned}\Pi_{i+1}&=\Pi_i-2[\Pi_i+W_o(\Pi_{i-1})/\alpha]\\&\quad\times[2\Pi_i+W_o(\Pi_{i-1})/\alpha+\alpha W_c(\Pi_{i-1})^{-1}]^{-1}[\Pi_i+W_o(\Pi_{i-1})/\alpha]+2W_o(\Pi_{i-1})/\alpha\end{aligned}$$

from $(\Pi_{-1},\Pi_0)\in\mathcal{P}(n)\times\mathcal{P}(n)$, *where* α *is any fixed positive constant. Then there holds*

$$\lim_{i\to\infty}\Phi_i=\lim_{i\to\infty}\Pi_i=\bar{P}. \tag{4.13}$$

Proof. Letting

$$\mathcal{F}(P)=W_o(P)/\alpha\quad\text{and}\quad\mathcal{G}(P)=W_c(P)/\alpha, \tag{4.14}$$

we can easily see that both $\mathcal{F}(P)$ and $\mathcal{G}(P)^{-1}$ are nondecreasing and continuous with respect to $P\in\mathcal{P}(n)$ and that $\bar{P}$ is the unique solution of $\mathcal{F}(P)=P\mathcal{G}(P)P$ in $\mathcal{P}(n)$. Because for any fixed $P\in\mathcal{P}(n)$ there hold

$$\lim_{\mu\to0^+}[\mathcal{F}(\mu P)-(\mu P)\mathcal{G}(\mu P)(\mu P)]=\frac{1}{2\pi\alpha i}\oint\mathcal{C}(z)^*\mathcal{C}(z)\frac{dz}{z}>0, \tag{4.15}$$

$$\lim_{\nu\to+\infty}[\mathcal{F}(\nu P)-(\nu P)\mathcal{G}(\nu P)(\nu P)]/\nu^2=-\frac{1}{2\pi\alpha i}\oint P\mathcal{B}(z)\mathcal{B}(z)^*P\frac{dz}{z}<0. \tag{4.16}$$

Thus, the theorem follows by directly applying Theorem 3.1 and Corollary 3.1. □

Remark 4.1. It is readily seen from Theorem 3.2 that the algorithm

(4.17)

$$\begin{aligned}\Sigma_{i+1} = \Sigma_i &- 2[\Sigma_i + W_c(\Sigma_i^{-1})/\alpha] \\ &\times [2\Sigma_i + W_c(\Sigma_i^{-1})/\alpha + \alpha W_o(\Sigma_i^{-1})^{-1}]^{-1}[\Sigma_i + W_c(\Sigma_i^{-1})/\alpha] + 2W_c(\Sigma_i^{-1})/\alpha\end{aligned}$$

with the initial value $\Sigma_0 \in \mathcal{P}(n)$ also provides an alternative way to compute the equilibrium $\bar{P}$.

Note that the calculation of $W_c(P)$ and $W_o(P)$ inevitably involves intensive iterations given a P. To overcome this drawback, we propose the following modified algorithm, which only needs to evaluate much simpler expressions than $W_c(P)$ and $W_o(P)$ at each iteration.

PROPOSITION 4.2. *Adopt the same hypotheses and notation as in Proposition* 4.1. *Then for any given* $\alpha > 0$ *and initial condition* $(\Psi_0, U_0, V_0) \in \mathcal{P}(n) \times \mathcal{P}(2n) \times \mathcal{P}(2n)$, *the solution* (Ψ_i, U_i, V_i) *of the system of difference equations*

(4.18)

$$\begin{aligned}\Psi_{i+1} = \Psi_i &- 2(\Psi_i + U_i^{11}/\alpha) \\ &\times [2\Psi_i + U_i^{11}/\alpha + \alpha(V_i^{11})^{-1}]^{-1}(\Psi_i + U_i^{11}/\alpha) + 2U_i^{11}/\alpha,\end{aligned}$$

(4.19)

$$U_{i+1} \triangleq \begin{bmatrix} U_{i+1}^{11} & U_{i+1}^{12} \\ U_{i+1}^{21} & U_{i+1}^{22} \end{bmatrix} = \begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix} \begin{bmatrix} U_i^{11} & U_i^{12} \\ U_i^{21} & U_i^{22} \end{bmatrix} \begin{bmatrix} A & 0 \\ bc & A \end{bmatrix} + \begin{bmatrix} c'c & 0 \\ 0 & \Psi_i \end{bmatrix},$$

(4.20)

$$V_{i+1} \triangleq \begin{bmatrix} V_{i+1}^{11} & V_{i+1}^{12} \\ V_{i+1}^{21} & V_{i+1}^{22} \end{bmatrix} = \begin{bmatrix} A & bc \\ 0 & A \end{bmatrix} \begin{bmatrix} V_i^{11} & V_i^{12} \\ V_i^{21} & V_i^{22} \end{bmatrix} \begin{bmatrix} A' & 0 \\ c'b' & A' \end{bmatrix} + \begin{bmatrix} bb' & 0 \\ 0 & \Psi_i^{-1} \end{bmatrix}$$

converges to its unique fixed point $(\bar{P}, \bar{U}, \bar{V}) \in \mathcal{P}(n) \times \mathcal{P}(2n) \times \mathcal{P}(2n)$.

Proof. Define the operators $\mathcal{S}$ and $\mathcal{T}$ on $\mathcal{Q}(2n) \times \mathcal{P}(n)$ by

(4.21)

$$\mathcal{S}(U, X) \triangleq \begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix} U \begin{bmatrix} A & 0 \\ bc & A \end{bmatrix} + \begin{bmatrix} c'c & 0 \\ 0 & X \end{bmatrix},$$

(4.22)

$$\mathcal{T}(V, X) \triangleq \begin{bmatrix} A & bc \\ 0 & A \end{bmatrix} V \begin{bmatrix} A' & 0 \\ c'b' & A' \end{bmatrix} + \begin{bmatrix} bb' & 0 \\ 0 & X^{-1} \end{bmatrix},$$

and let $L = K = \begin{bmatrix} I \\ 0 \end{bmatrix} \in \mathbb{R}^{2n \times n}$ with I an $n \times n$ identity matrix. Because (A, b, c) is minimal, the operators $\mathcal{S}$ and $\mathcal{T}$ continuously map $\mathcal{P}(2n) \times \mathcal{P}(n)$ into $\mathcal{P}(2n)$. Quite obviously, $\mathcal{S}$ and $\mathcal{T}$ meet Assumption 3 in the previous section. Moreover, it is clear that the Lyapunov equations

$$U = \mathcal{S}(U, X) \quad \text{and} \quad V = \mathcal{S}(V, X)$$

have unique solutions $U(X) \in \mathcal{P}(2n)$ and $V(X) \in \mathcal{P}(2n)$ for any $X \in \mathcal{P}(n)$. In addition to this, it is known from [5] that

$$\mathcal{F}(X) = K'U(X)K \quad \text{and} \quad \mathcal{G}(X) = L'V(X)L,$$

where $\mathcal{F}(\cdot)$ and $\mathcal{G}(\cdot)$ are defined as in (4.14). Thus, Assumption 4 is fulfilled.

Next, it is routine to check inductively that $\Psi_i > 0$ for all $i \geq 0$. Hence, from (i) of Lemma 2.3, there exists an integer k_1 such that

$$U_i^{11}, V_i^{11} > \beta I, \quad \forall\, i \geq k_1 \tag{4.23}$$

for some constant $\beta > 0$. By Corollary 2.1, this implies that $\Psi_i > (\beta/\alpha)I$ for all $i > k_1$. Making use of (ii) of Lemma 2.3 yields that there exists $k_2 > k_1$ such that

$$U_i > \tilde{U} \quad \text{and} \quad V_i > \tilde{V}, \quad \forall\, i \geq k_2, \tag{4.24}$$

where $\tilde{U}$ and $\tilde{V}$ are the solutions to the Lyapunov equations

$$U = \begin{bmatrix} A' & c'b' \\ 0 & A' \end{bmatrix} U \begin{bmatrix} A & 0 \\ bc & A \end{bmatrix} + \begin{bmatrix} c'c & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.25}$$

$$V = \begin{bmatrix} A & bc \\ 0 & A \end{bmatrix} V \begin{bmatrix} A' & 0 \\ c'b' & A' \end{bmatrix} + \begin{bmatrix} bb' & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.26}$$

respectively. Now by Lemma 2.2, for sufficiently small $\mu > 0$ and sufficiently large $\nu > 0$ there hold

$$U_{k_2} < U(\nu I) \quad \text{and} \quad V_{k_2} < V(\mu I). \tag{4.27}$$

Because

$$\lim_{\mu \to 0} U(\mu I) = \tilde{U} \quad \text{and} \quad \lim_{\nu \to \infty} V(\nu I) = \tilde{V}, \tag{4.28}$$

it is seen from (4.24) that for sufficiently small $\mu > 0$ and sufficiently large $\nu > 0$ there hold

$$U_{k_2} > U(\mu I) \quad \text{and} \quad V_{k_2} > V(\nu I). \tag{4.29}$$

Also, it is clear that

$$\mu I < \Psi_{k_2} < \nu I \tag{4.30}$$

for sufficiently small $\mu > 0$ and sufficiently large $\nu > 0$. In view of (4.15)–(4.16), a direct application of Theorem 3.3 leads to

$$\lim_{i \to \infty} (\Psi_i, U_i, V_i) = (\bar{P}, U(\bar{P}), V(\bar{P})). \qquad \square$$

Finally, regarding the Euclidean norm-balancing problem with an initial minimal realization (A, B, C), we claim that with the new definitions

$$W_o(P) \triangleq A'PA + C'C \quad \text{and} \quad W_c(P) \triangleq AP^{-1}A' + BB', \tag{4.31}$$

(4.11) and (4.12) are convergent to the unique solution $\mathbb{P}$ of

$$(A'PA + C'C) - P(AP^{-1}A' + BB')P = 0 \tag{4.32}$$

in $\mathcal{P}(n)$. In fact, it suffices to verify by Theorem 3.1 that for any $P \in \mathcal{P}(n)$ there exist P_1, $P_2 \in \mathcal{P}(n)$ with $P_1 \leq P \leq P_2$ such that

$$W_o(P_1) \geq P_1 W_c(P_1) P_1 \quad \text{and} \quad W_o(P_2) \leq P_2 W_c(P_2) P_2. \tag{4.33}$$

However, this is true upon noting that

$$\begin{aligned} W_o(\mu\mathbb{P}) - (\mu\mathbb{P})W_c(\mu\mathbb{P})(\mu\mathbb{P}) &= \mu A'\mathbb{P}A + C'C - \mu^2\mathbb{P}(\mu^{-1}A\mathbb{P}^{-1}A' + BB')\mathbb{P} \\ &= \mu(A'\mathbb{P}A - \mathbb{P}A\mathbb{P}^{-1}A'\mathbb{P}) + (C'C - \mu^2\mathbb{P}BB'\mathbb{P}) \\ &= \mu(\mathbb{P}BB'\mathbb{P} - C'C) + (C'C - \mu^2\mathbb{P}BB'\mathbb{P}) \\ &= (1-\mu)(C'C + \mu\mathbb{P}BB'\mathbb{P}). \end{aligned}$$

Hence, the claim is true.

5. A result on convergence rate. In this section we prove that the convergence of (4.11) is locally exponential by showing that their unique equilibrium is sink. In other words, the linearization of (4.11) at the equilibrium has all its eigenvalues in the open unit disk.

PROPOSITION 5.1. *The linearization of* (4.11) *is asymptotically stable at the fixed point* $\bar{P}$.

Proof. We might as well assume $\alpha = 1$. Now with

(5.1)

$$S_1(X) \triangleq [X + W_o(X)]^{-1}, \quad S_2(X) = [X + W_c(X)^{-1}]^{-1}, \quad T(X) = [S_1(X) + S_2(X)]^{-1},$$

(4.11) can be rewritten as

$$P_{i+1} = \mathcal{R}(P_i) \triangleq 2T(P_i) - P_i. \tag{5.2}$$

Note that $S_1(X), S_2(X), T(X)$ are defined for the realization (A, b, c). Accordingly, we can define $\mathbb{A}(z)$, $\mathbb{W}_o(X)$, and $\mathbb{S}_1$, and so on for the L^2-sensitivity optimal realization $(\bar{P}^{\frac{1}{2}}A\bar{P}^{-\frac{1}{2}}, \bar{P}^{\frac{1}{2}}b, c\bar{P}^{-\frac{1}{2}})$. Then it is routine to check the following relations:

$$\bar{P}^{\frac{1}{2}}S_1(\bar{P})\bar{P}^{\frac{1}{2}} = \mathbb{S}_1(I), \quad \bar{P}^{\frac{1}{2}}S_2(\bar{P})\bar{P}^{\frac{1}{2}} = \mathbb{S}_2(I), \quad \bar{P}^{-\frac{1}{2}}T(\bar{P})\bar{P}^{-\frac{1}{2}} = \mathbb{T}(I). \tag{5.3}$$

Because $\mathcal{R}(X)$ is an operator from $R^{n\times n}$ to $R^{n\times n}$, its Fréchet derivative at $X = \bar{P}$ is a linear operator from $R^{n\times n}$ to $R^{n\times n}$ given by

$$\begin{aligned} \mathcal{R}'(\bar{P})X = {} & 2T(\bar{P})\{S_1(\bar{P})[X + W_o'(\bar{P})X]S_1(\bar{P}) \\ & + S_2(\bar{P})[X - W_c(\bar{P})^{-1}W_c'(\bar{P})XW_c(\bar{P})^{-1}]S_2(\bar{P})\}T(\bar{P}) - X, \end{aligned} \tag{5.4}$$

with

(5.5)

$$W_o'(\bar{P})X = \frac{1}{2\pi i}\oint \mathcal{A}(z)^*X\mathcal{A}(z)\frac{dz}{z} \quad \text{and} \quad W_c'(\bar{P})X = -\frac{1}{2\pi i}\oint \mathcal{A}(z)\bar{P}^{-1}X\bar{P}^{-1}\mathcal{A}(z)^*\frac{dz}{z}.$$

Using (5.3), we have

(5.6)

$$\begin{aligned} & \bar{P}^{-\frac{1}{2}}[\mathcal{R}'(\bar{P})X]\bar{P}^{-\frac{1}{2}} \\ & \quad = 2\mathbb{T}(I)\{\mathbb{S}_1(I)\left[\bar{P}^{-\frac{1}{2}}X\bar{P}^{-\frac{1}{2}} + \frac{1}{2\pi i}\oint \mathcal{A}(z)^*(\bar{P}^{-\frac{1}{2}}X\bar{P}^{-\frac{1}{2}})\mathcal{A}(z)\frac{dz}{z}\right]\mathbb{S}_1(I) \\ & \qquad + \mathbb{S}_2(I)\left[\bar{P}^{-\frac{1}{2}}X\bar{P}^{-\frac{1}{2}} + \mathbb{W}_c(I)^{-1}\frac{1}{2\pi i}\oint \mathcal{A}(z)\bar{P}^{-\frac{1}{2}}X\bar{P}^{-\frac{1}{2}}\mathcal{A}(z)^*\frac{dz}{z}\mathbb{W}_c(I)^{-1}\right] \\ & \qquad \cdot \mathbb{S}_2(I)\}\mathbb{T}(I) - \bar{P}^{-\frac{1}{2}}X\bar{P}^{-\frac{1}{2}} \\ & \quad = \mathbb{R}'(I)(\bar{P}^{-\frac{1}{2}}X\bar{P}^{-\frac{1}{2}}), \end{aligned} \tag{5.6}$$

from which it is not hard to see that the linear operator $\mathcal{R}'(\bar{P})$ is asymptotically stable in the discrete time sense if and only if $\mathbb{R}'(I)$ is also. Because $\mathbb{W}_c(I) = \mathbb{W}_o(I)$, it follows that

$$\mathbb{R}'(I)X$$

$$= 2\mathbb{T}\mathbb{S}_1 \left[X + \mathbb{W}_o X \mathbb{W}_o + \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* X \mathbb{A}(z) + \mathbb{A}(z) X \mathbb{A}(z)^*] \frac{dz}{z} \right] \mathbb{S}_1 \mathbb{T} - X \tag{5.7}$$

$$= 2\mathbb{S}_1 \left[X + \mathbb{W}_o X \mathbb{W}_o + \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* X \mathbb{A}(z) + \mathbb{A}(z) X \mathbb{A}(z)^*] \frac{dz}{z} \right] \mathbb{S}_1 - X, \tag{5.8}$$

where $\mathbb{T}$ is understood to be $\mathbb{T}(I)$ and so on. Thus, the matrix representation of $\mathbb{R}'(I)$ is expressed by

$$2(\mathbb{S}_1 \otimes \mathbb{S}_1) \left\{ I + \mathbb{W}_o \otimes \mathbb{W}_o + \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* \otimes \mathbb{A}(z)^\tau + \mathbb{A}(z) \otimes \mathbb{A}(\bar{z})] \frac{dz}{z} \right\} - I. \tag{5.9}$$

It remains to show that this matrix has all its eigenvalues in the open unit disk, which is equivalent to saying that so is the matrix

$$\Gamma \triangleq 2(\mathbb{S}_1^{1/2} \otimes \mathbb{S}_1^{1/2}) \left\{ I + \mathbb{W}_o \otimes \mathbb{W}_o + \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* \otimes \mathbb{A}(z)^\tau + \mathbb{A}(z) \otimes \mathbb{A}(\bar{z})] \frac{dz}{z} \right\} \cdot (\mathbb{S}_1^{1/2} \otimes \mathbb{S}_1^{1/2}) - I.$$

Since Γ is symmetric and $\Gamma > -I$, it suffices to prove that $\Gamma < I$. To do this, note that

$$\begin{aligned}
&\Gamma < I \\
&\iff I + \mathbb{W}_o \otimes \mathbb{W}_o + \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* \otimes \mathbb{A}(z)^\tau + \mathbb{A}(z) \otimes \mathbb{A}(\bar{z})] \frac{dz}{z} < \mathbb{S}_1^{-1} \otimes \mathbb{S}_1^{-1} \\
&\iff \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* \otimes \mathbb{A}(z)^\tau + \mathbb{A}(z) \otimes \mathbb{A}(\bar{z})] \frac{dz}{z} < \mathbb{W}_o \otimes I + I \otimes \mathbb{W}_o \\
&\iff \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* \otimes \mathbb{A}(z)^\tau + \mathbb{A}(z) \otimes \mathbb{A}(\bar{z})] \frac{dz}{z} \\
&\quad < \frac{1}{2\pi i} \oint [\mathbb{A}(z)^* \mathbb{A}(z) \otimes I^\tau + I \otimes \mathbb{A}(\bar{z}) \mathbb{A}(z)^\tau + \mathbb{C}(z)^* \mathbb{C}(z) \otimes I + I \otimes \mathbb{B}(z) \mathbb{B}(z)^*] \frac{dz}{z} \\
&\iff \frac{1}{2\pi i} \oint (\mathbb{A}(z) \otimes I - I \otimes \mathbb{A}(z)^\tau)^* (\mathbb{A}(z) \otimes I - I \otimes \mathbb{A}(z)^\tau) \frac{dz}{z} + \Sigma_o \otimes I + I \otimes \Sigma_c \\
&> 0,
\end{aligned}$$

where Σ_o and Σ_c denote the observability and controllability Gramians. Hence, it follows that $\Gamma < I$. □

6. Simulation results. The purpose of this section is to demonstrate the effectiveness of the algorithms proposed in the previous section by simulation on a SPARCstation. To do this, consider a specific minimal statespace realization (A, b, c) with

$$A = \begin{bmatrix} 0.5 & 0 & 1.0 \\ 0 & -0.25 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, \quad b - \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad c = [1 \quad 5 \quad 10]. \tag{6.1}$$

Recall that there exists a unique positive definite matrix $\bar{P} \in \mathbb{R}^{3\times 3}$ such that the realization (TAT^{-1}, Tb, cT^{-1}) is L^2-sensitivity optimal for any similarity transformation T with $T'T =$

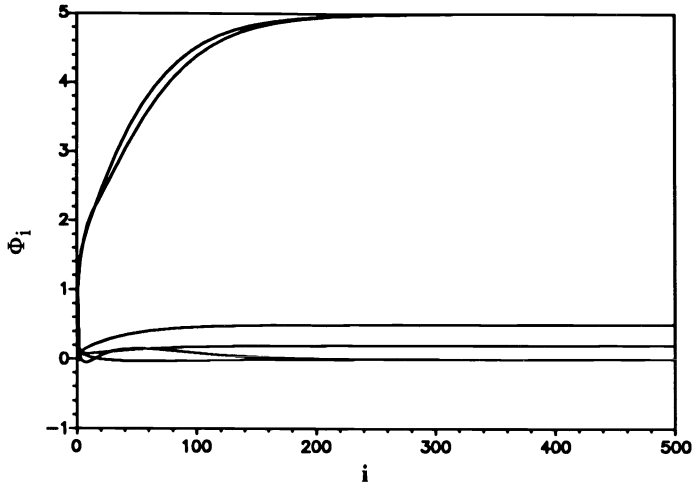


FIG. 6.1. *The trajectory of* Φ_i *of* (4.11) *with* $\alpha = 300$.

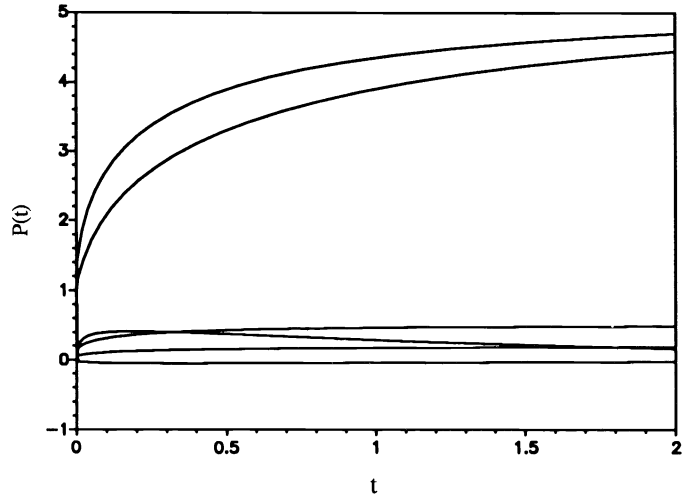


FIG. 6.2. *The trajectory of* $P(t)$ *of* (4.5).

$\bar{P}$. It turns out that $\bar{P}$ is exactly given by

$$\bar{P} = \begin{bmatrix} 0.2 & 0 & 0.5 \\ 0 & 5.0 & 0 \\ 0.5 & 0 & 5.0 \end{bmatrix}, \tag{6.2}$$

which indeed satisfies (4.8).

We first take (4.11) with $\alpha = 300$ and implement it starting from the identity matrix using MATLAB. The resulting trajectory Φ_i during the first 500 iterations is shown in Fig. 6.1 and is clearly seen to converge very fast to $\bar{P}$. The time taken for this implementation is less than three minutes. In contrast, if an ordinary differential equation (ODE) algorithm in MABLAB is used to solve (4.5) with the same initial condition, it is found that it takes about 45 minutes to compute the solution $P(t)$ on the time interval [0, 2], which is depicted in Fig. 6.2. In fact, more than 2,400 iterations are performed during that time interval. Even so, the solution does not appear to be close enough to $\bar{P}$ though it very slowly tends to $\bar{P}$.

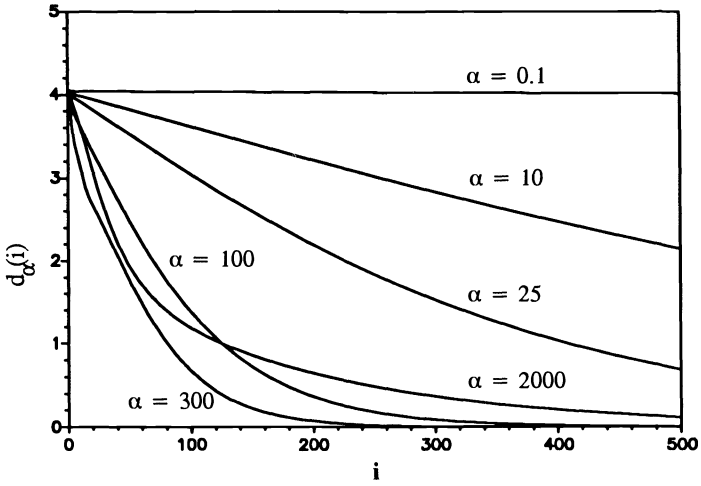


FIG. 6.3. *Effect of different* α *on the convergence rate of* (4.11).

Next, we examine the effect of α on the convergence rate (4.11). For this purpose, define the deviation between Φ_i and the true solution $\bar{P}$ in (6.2) as

$$d_\alpha(i) = \|\Phi_i - \bar{P}\|_2, \tag{6.3}$$

where $\|\cdot\|_2$ denotes the spectral norm of a matrix. Implement (4.11) with

$$\alpha = 0.1,\ 10,\ 25,\ 100,\ 300,\ 2000,$$

respectively, and depict the evolution of the associated deviation $d_\alpha(i)$ for each α in Fig. 6.3. Then we can see that $\alpha = 300$ is the best choice. In addition, as long as $\alpha \leq 300$, the larger α, the faster the convergence of the algorithm. On the other hand, it should be observed that a larger α is not always better than a smaller α and that too small an α can make the convergence extremely slow.

Finally, let us turn to two algorithms (4.12) and (4.18)–(4.20) with $\alpha = 300$. All the initial matrices required for the implementation are set to identity matrices of appropriate dimension. Define

$$f(i) = \|\Pi_i - \bar{P}\|_2 \quad \text{and} \quad g(i) = \|\Psi_i - \bar{P}\|_2$$

as the deviations from the true solution $\bar{P}$ for the two algorithms, respectively. Their evolutions are depicted in Fig. 6.4 and manifestly exhibit the convergence of the algorithms. Indeed, the algorithm (4.18)–(4.20) is fastest in terms of the execution time, but with the same number of iterations it does not produce a solution as satisfactory as (4.11) or (4.12).

Some concluding remarks are in order.

Remark 6.1. Adding a scalar factor to (4.5) does not help much in reducing the CPU time required for solving it on a digital computer.

Remark 6.2. Because α does play an important role in speeding up the algorithms, it is worthwhile to do further study to find some helpful guidelines for choosing a suitable α.

Remark 6.3. It appears that the proposed algorithms are quite robust against nonsymmetric or indefinite disturbances. This is demonstrated by implementing (4.11) with $\alpha = 300$ and the initial matrix

$$\Phi_0 = \begin{bmatrix} 1 & 10 & -10 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

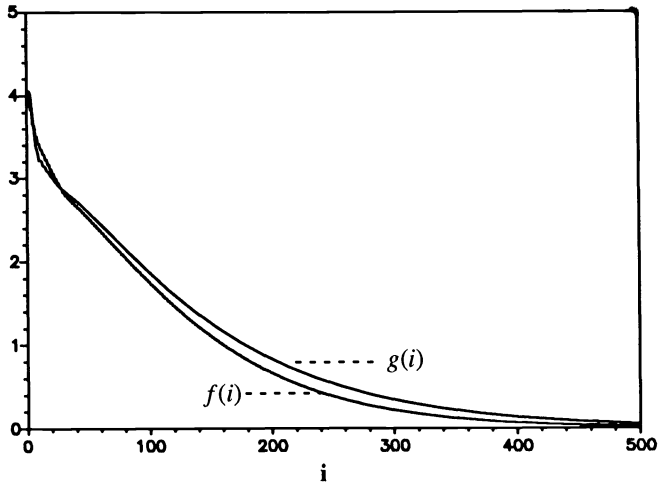


FIG. 6.4. *Convergence of algorithms* (4.12) *and* (4.18)–(4.20) *with* $\alpha = 300$.

which is obviously nonsymmetric and indefinite. It turns out that the effect of the nonsymmetry and indefiniteness almost completely vanishes after 30 iterations. Afterwards, the algorithm converges to $\bar{P}$ with no oscillatory behavior.

7. Conclusions. Two types of difference equation have been proposed and studied with the aim of solving a class of nonlinear matrix equations. The main contribution of this paper is twofold. First, we characterize a set of initial conditions from which the solution of the proposed difference equations is guaranteed to converge to the solution of the matrix equation of concern. Second, the general results have been successfully employed to derive a number of efficient iterative algorithms for finding L^2-sensitivity optimal realizations and Euclidean norm-balancing realizations. These algorithms are simple to implement without any requirement of step-size adjustment. Another feature of them is that their convergence rate can be significantly improved by proper choice of a constant scalar in advance. In addition, the convergence of one algorithm is locally exponential. The effectiveness of the algorithms are demonstrated by simulation.

REFERENCES

[1] U. HELMKE AND J. B. MOORE, *L^2 sensitivity minimization of linear system representations via gradient flows*, J. Math. Systems Control Theory, to appear.

[2] K. L. HITZ AND B. D. O. ANDERSON, *Iterative method of computing the limiting solution of the matrix Riccati differential equation*, Proceedings of IEE, 119 (1972), pp. 1402–1406.

[3] J. E. PERKINS, U. HELMKE, AND J. B. MOORE, *Balanced realizations via gradient flows*, Systems Control Lett., 14 (1990), pp. 369–380.

[4] L. T. WATSON, *Engineering Applications of the Chow–Yorke Algorithm, Homotopy Methods and Global Convergence*, Plenum Press, New York, 1983, pp. 287–307.

[5] W.-Y. YAN AND J. B. MOORE, *On L^2-sensitivity minimization of linear state-space systems*, IEEE Trans. Circuits and Systems, 39 (1992), pp. 641–648.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1577–1596, November 1994

005

OPTIMAL PROBLEMS FOR NONLINEAR PARABOLIC BOUNDARY CONTROL SYSTEMS*

H. O. FATTORINI† AND T. MURPHY†

Abstract. This paper considers optimal problems for boundary control systems with the control acting through a general nonlinear boundary condition. The problems include constraints on the control and target conditions. The final result is a version of Pontryagin's maximum principle. This setup covers such physical situations as Stefan–Boltzmann boundary conditions, and is based on an abstract theory of nonlinear programming problems.

Key words. boundary control systems, maximum principle, optimal control

AMS subject classifications. 93E20, 93E25

1. Introduction. Let Ω be a bounded domain with boundary $\partial\Omega$ in m-dimensional Euclidean space $\mathbb{R}^m$. We consider a boundary control system described by the heat equation in Ω,

$$y_t(t,x) = \Delta y(t,x) \qquad (x \in \Omega,\ 0 < t \leq T), \tag{1.1}$$

$$y(0,x) = \zeta(x), \qquad (x \in \Omega), \tag{1.2}$$

with a nonlinear boundary condition

$$\partial_\nu y(t,x) = g(t, y(t,x)) + u(t,x) \qquad (x \in \partial\Omega,\ 0 < t \leq T) \tag{1.3}$$

(∂_ν is the outer normal derivative at the boundary). Controls are taken in $L^\infty((0,T) \times \partial\Omega)$. The optimal control problem is that of minimizing a *cost functional* $y_0(t,u)$ among all controls satisfying a constraint

$$u(t,\cdot) \in U = \textit{ control set } \subseteq L^\infty(\partial\Omega) \qquad (0 \leq t \leq \bar{t}) \tag{1.4}$$

whose corresponding solutions $y(t,x,u)$ satisfy a *target condition*

$$y(\bar{t},\cdot,u) \in Y = \textit{target set } \subseteq C(\bar{\Omega}) \tag{1.5}$$

($C(\bar{\Omega})$ is the space of all continuous functions in $\bar{\Omega}$ endowed with the supremum norm). The time $\bar{t}$ at which the optimal process terminates may be fixed or free. A physically realistic instance is that where $m = 1$, 2, or 3 and the body Ω undergoes Stefan–Boltzmann cooling at the boundary,

$$\partial_\nu y(t,x) = -y(t,x)^4 + u(t,x) \qquad (x \in \partial\Omega,\ 0 < t \leq \bar{t}). \tag{1.6}$$

The term $u(t,x)$ represents heat applied at the boundary $\partial\Omega$ and satisfies constraints on $\partial\Omega$, for instance

$$0 \leq u(t,x) \leq 1 \qquad (x \in \partial\Omega,\ 0 \leq t \leq \bar{t}). \tag{1.7}$$

As a very particular case of the results in this paper, we obtain existence and necessary conditions for the time optimal problem under (1.6) and (1.7).

* Received by the editors October 1, 1992; accepted for publication (in revised form) June 25, 1993. This research was supported by National Science Foundation grant DMS-9001793.

† Department of Mathematics, University of California, Los Angeles, California 90024.

EXISTENCE THEOREM 1.1. *Assume that a control* $u(\cdot,\cdot) \in L^\infty((0,\bar{t})\times\Omega)$ *satisfying* (1.7) *almost everywhere and driving an initial state* $\zeta(\cdot) \in C(\bar{\Omega})$ *to a closed target set* $Y \subseteq C(\bar{\Omega})$ *exists. Then there exists a time optimal control* $\bar{u}(t,x)$.

THEOREM 1.2 (Pontryagin's maximum principle). *Let* $\bar{u}(\cdot,\cdot) \in L^\infty((0,\bar{t})\times\Omega)$ *be a time optimal control and* $y(t,x,\bar{u})$ *be the solution corresponding to* $\bar{u}$. *Let the target set* $Y \subseteq C(\bar{\Omega})$ *be defined by*

$$|y(x) - \bar{y}(x)| \le \varepsilon \quad (x \in \bar{\Omega}), \qquad y(x_j) = c_j \quad (j = 1,2,\dots,N),$$

where $\bar{y}(\cdot) \in C(\bar{\Omega})$, $\varepsilon > 0$, $x_j \in \bar{\Omega}$, *and the* c_j *are arbitrary constants. Then there exists a finite Borel measure* ν *in* $\bar{\Omega}$, $\nu \neq 0$, *such that*

$$\int_{\partial\Omega} z(t,x)\bar{u}(t,x)d\sigma = \max_{v\in L^\infty(\partial\Omega),0\le v(x)\le 1} \int_{\partial\Omega} z(t,x)v(x)d\sigma$$

almost everywhere in $0 \le t \le \bar{t}$ = *optimal time, where* $d\sigma$ *is the area differential in* $\partial\Omega$ *and* $z(t,x)$ *is the solution of the backwards problem*

$$z_t(t,x) = -\Delta z(t,x) \qquad (x \in \Omega,\ 0 \le t < \bar{t}), \tag{1.8}$$

$$z(\bar{t},\cdot) = \nu \qquad (x \in \bar{\Omega}), \tag{1.9}$$

$$z_\nu(t,x) = -4y(t,x,\bar{u})^3 z(t,x), \qquad (x \in \partial\Omega,\ 0 \le t < \bar{t}). \tag{1.10}$$

For a proof of (a more general version of) Theorem 1.2, see §6, where versions of Pontryagin's maximum principle are also given for other optimal problems; the proofs are based on an abstract theory of nonlinear programming problems in Banach spaces [10], [11], which generalizes some of the results in [16]. The earlier Hilbert space theory is in [7], [8], [14], [15]. Existence of optimal controls is treated in §4. We clarify in §§3 and 4 the definition of solutions of (1.1)–(1.3) and of other auxiliary equations (such as (1.8)–(1.10)), and compute in §5 the directional derivatives of the solution map $u \to y(t,y,u)$ needed to apply the abstract nonlinear programming theory. Finally, some open problems are pointed out in §6.

The method chosen to construct solutions of (1.1)–(1.3) is the classical one of integral equations with the Neumann function as kernel. It has been extensively used in boundary control problems for nonlinear parabolic equations, for instance in [30], [26], [28], [29], solutions taking instantaneous values in L^p or Sobolev spaces. The only novelty in our treatment may be that solutions take values in the space $C(\bar{\Omega})$. There are some advantages in this. The first is conceptual. If the equation describes a heat process, the supremum norm in $C(\bar{\Omega})$ is more natural than the $L^p(\Omega)$ norm—the latter allows for large temperature spikes possibly exceeding the melting point of the material and invalidating (1.1)–(1.3) as a model (of course, this could also be avoided via Sobolev norms). Also, the supremum norm allows us to use smaller target sets than L^p norms. The second advantage is one purely of convenience; the supremum norm allows easy treatment of continuous nonlinearities in the equation (no growth conditions needed), and trace theorems become trivial.

The reason the method of integral equations is preferred over others in this paper is that expressions for the solutions are formally similar to those for distributed parameter systems (see for instance [11]), so that proving continuity of the solution map and computing directional derivatives in §5 can be done in well-understood ways. On the other hand, the existence results that we obtain in this way for (1.1)–(1.3) are by no means the best available. For instance, much more general boundary conditions are dealt with in [1] and other papers of the same

author. Also, different approaches using monotonicity rather than Lipschitz continuity [2], [3] yield results for some nonsmooth nonlinearities.

A large number of recent papers examine boundary control problems for parabolic equations, and we have included some of them in the references. In the linear case, [5] deals with the time optimal problem with a point target (see comments at the end of §6). Optimal approximation in the supremum norm is studied in [18]; the results in [22] and [23] include target conditions as well as state constraints. In the nonlinear case, [26] deals with boundary conditions including (among many others) Stefan–Boltzmann; the problem is optimum approximation of a target. Similar problems in dimension ≥ 1 are considered in [27]. The integral equation approach is used in [30], [19], and (combined with semigroup methods) in [28]. Finally, many control problems, distributed parameter and boundary are outlined in [29] with a substantial bibliography.

In the one-dimensional case, the abstract nonlinear programming approach to optimal problems applied in this paper was outlined (but not carried out in detail) in [8].

2. The Neumann function. Because of hypoellipticity of the heat operator, any solution of

$$y_t(t,x) = \Delta y(t,x) \qquad ((t,x) \in \Omega \times (0,T]). \tag{2.1}$$

$$y(0,x) = \zeta(x) \qquad (x \in \Omega), \tag{2.2}$$

$$\partial_\nu y(t,x) = g(t,x) \qquad ((t,x) \in (0,T) \times \partial\Omega) \tag{2.3}$$

(∂_ν the outer normal derivative) is analytic in $(0,T] \times \Omega$. If $\zeta(\cdot)$ is continuous in $\bar{\Omega}$ and $g(\cdot,\cdot)$ is continuous in $[0,T] \times \partial\Omega$, a solution $y(t,x)$ is called *strong* or *classical* if it is continuous in $[0,T] \times \bar{\Omega}$, continuously differentiable in $((0,T] \times \bar{\Omega})$, and satisfies (2.1) in $(0,T] \times \Omega$, (2.2) in $\bar{\Omega}$, and (2.3) literally in $(0,T] \times \partial\Omega$; the normal derivative $\partial_\nu y(t,x)$ is assumed to be continuous in $[0,T] \times \partial\Omega$. A solution is *semistrong* if it is continuous in $[0,T] \times \bar{\Omega}$ and satisfies (2.1) in $(0,T] \times \Omega$ and (2.2) in $\bar{\Omega}$ where $\partial_\nu y(t,x)$ at a boundary point x is understood as the limit of $\partial_\nu y(t,\eta)$, where $\eta \to x$ in K, K any finite closed cone with vertex in x such that $K \subseteq \Omega$. Similar considerations apply to the initial-boundary value problem in an interval $\tau \leq t \leq T$.

We call $\Gamma(t,x;\,\tau,\xi)$ the *fundamental solution* of the heat equation in $\mathbb{R}^m$,

$$\Gamma(t,x;\,\tau,\xi) = \frac{\exp(-|x-\xi|^2/4(t-\tau))}{(4\pi)^{m/2}(t-\tau)^{m/2}},$$

$x, \xi \in \mathbb{R}^m$, $x \neq \xi$, $\tau < t$. The *Neumann function* $N(t,x;\,\tau,\xi)$ of the heat equation in $[0,T] \times \Omega$ is defined for $t \geq \tau$, $\xi \in \Omega$ by

$$N(t,x;\,\tau,\xi) = \Gamma(t,x;\,\tau,\xi) - V(t,x;\,\tau,\xi) \qquad ((t,x) \in [\tau,T] \times \Omega),$$

where $V(t,x;\,\tau,\xi)$ satisfies [17, p. 155]

$$V_t(t,x;\,\tau,\xi) = \Delta_x V(t,x;\,\tau,\xi), \qquad ((t,x) \in (\tau,T] \times \Omega)),$$

$$V(\tau,x;\,\tau,\xi) = 0, \qquad (x \in \Omega),$$

$$\partial_{\nu,x} V(t,x;\,\tau,\xi) = \partial_{\nu,x}\Gamma(t,x;\,\tau,\xi), \qquad ((t,x) \in (\tau,T] \times \partial\Omega).$$

A domain $\Omega \subseteq \mathbb{R}^m$ is said to be *of class* $C^{(2)}$ if, locally, its boundary $\partial\Omega$ can be represented in the form $x_j = h(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_m)$ with h twice continuously differentiable; if h is continuously differentiable with λ-Hölder continuous derivatives, Ω is *of class* $C^{(1+\lambda)}$. If $\xi \in \Omega$, $\partial_{\nu,x}\Gamma(t,x;\,\tau,\xi)$ is infinitely differentiable in the cylinder $[\tau,T]\times\bar{\Omega}$, thus [17, Thm. 2, p. 144] $V(t,x;\,\tau,\xi)$ exists; if Ω is of class $C^{(2)}$, $V(t,x;\,\tau,\xi)$, as a function of t, x is analytic in $(\tau,T]\times\Omega$, continuous in $[\tau,T]\times\bar{\Omega}$, and continuously differentiable in $(\tau,T]\times\bar{\Omega}$. The boundary condition is understood literally. Application of the maximum principle shows that

$$N(t,x;\,\tau,\xi) > 0, \qquad (t > \tau, x, \xi \in \Omega). \tag{2.4}$$

If $N^-(t,x;\,\tau,\xi)$ is the Neumann function of the backward heat equation $y_t = -\Delta y$ in $[0,T]\times\Omega$, then

$$N^-(t,x;\,\tau,\xi) = N(\tau,\xi,t,x) \qquad (t > \tau, x, \xi \in \Omega). \tag{2.5}$$

(The proof is similar to that for the Green function, see [17, p. 84].) Equality (2.5) translates to (τ,ξ) the (t,x)-dependence properties of the Neumann function. An application of the divergence theorem shows that if $y(t,x)$ is a strong solution of (2.1)–(2.3), then

$$\begin{aligned} y(t,x) = &\int_\Omega N(t,x;\,\tau,\xi)\zeta(\xi)d\xi \\ &+ \int_{(\tau,t)\times\partial\Omega} N(t,x;\,r,\xi)g(r,\xi)d\sigma_\xi\,dr \qquad (t > \tau,\ x \in \Omega), \end{aligned} \tag{2.6}$$

where $d\sigma_\xi$ is the area differential on $\partial\Omega$. On the other hand, if $g(t,\xi)$ is continuously differentiable in $[0,T]\times\partial\Omega$ and $\zeta(x)$ is continuously differentiable in $\bar{\Omega}$ with $\partial_\nu y(0,x) = g(0,x)$, a strong solution exists [21] and thus (2.6) provides a strong solution for such ζ and g. An approximation argument using the continuous dependence properties of the solution of (2.1)–(2.3) on ζ, and g [17, p. 146] shows that (2.6) extends to semistrong solutions. Formula (2.6) will be used in §3 to define weak solutions of (1.1)–(1.3) via suitable estimates for the Neumann function. Because $N(t,x,\tau,\xi)$ inherits the t-singularity of $\Gamma(t,x;\,\tau,\xi)$ at $(t,x) = (\tau,\xi)$, which is not integrable (except in dimension 1), we use the properties of the negative exponential to quench this singularity at the cost of introducing an "artificial singularity" in x in the estimations. Although the results below (in particular, Theorem 2.1) are apparently known to all specialists, we were unable to locate precise references and include a sketch of the proofs.

We construct $V(t,x;\,\tau,\xi)$ using a method in [17, p. 144] to solve the initial-boundary value problem (2.1)–(2.3). We try

$$V(t,x;\,\tau,\xi) = \int_{(\tau,t)\times\partial\Omega} \Gamma(t,x;\,r,\eta)\rho(r,\eta;\,\tau,\xi)d\sigma_\eta\,dr \tag{2.7}$$

with density $\rho(r,\eta;\,\tau,\xi)$ to be determined. Using formula (2.10) in [17, p. 137] for the derivative of single layer potentials and the boundary condition for V, we obtain the integral equation

$$\rho(t,x;\,\tau,\xi) = 2\partial_{\nu,x}\Gamma(t,x;\,\tau,\xi) - 2\int_{(\tau,t)\times\partial\Omega} \partial_{\nu,x}\Gamma(t,x;\,r,\eta)\rho(r,\eta;\,\tau,\xi)d\sigma_\eta\,dr \tag{2.8}$$

for $\rho(r,\eta;\,\tau,\xi)$ in x, $\xi \in \partial\Omega$. We solve this equation by the following iteration process:

$$\rho(t,x;\,\tau,\xi) = 2\partial_{\nu,x}\Gamma(t,x;\,\tau,\xi) - 2\int_{(\tau,t)\times\partial\Omega} \partial_{\nu,x}\Gamma(t,x;\,r,\eta)R(r,\eta;\,\tau,\xi)d\sigma_\eta\,dr \tag{2.9}$$

$$R(r,\eta;\,\tau,\xi)=\sum_{j=0}^{\infty}\rho_j(r,\eta;\,\tau,\xi), \tag{2.10}$$

the ρ_j defined inductively by $\rho_0(t,x;\,\tau,\xi)=2\partial_{\nu,\eta}\Gamma(t,x;\,\tau,\xi)$,

$$\rho_{j+1}(t,x;\,\tau,\xi)=-2\int_{(\tau,t)\times\partial\Omega}\partial_{\nu,x}\Gamma(t,x;\,r,\eta)\rho_j(r,\eta;\,\tau,\xi)d\sigma_\eta\,dr. \tag{2.11}$$

Estimation of the successive terms depends on the inequality

$$\int_\Sigma\frac{d\sigma_\eta}{|x-\eta|^\alpha|\eta-\xi|^\beta}\le\begin{cases}C|x-\xi|^{m-1-\alpha-\beta} & \text{if } \alpha+\beta>m-1,\\ C & \text{if } \alpha+\beta<m-1,\end{cases} \tag{2.12}$$

valid in a bounded $C^{(1)}$ hypersurface in $\mathbb{R}^m$ for arbitrary α, β and x, ξ in a bounded subset $A\subseteq\mathbb{R}^m$. If $x,\,\xi\notin\Sigma$, then α and β are arbitrary. If $x\in\Sigma$ (respectively, $\xi\in\Sigma$), we must have $\alpha<m-1$ (respectively, $\beta<m-1$); in all cases, the constant C only depends on A, α, β. For a proof, see [24]; a somewhat less general formula is in [17, p. 137].

The estimates for $\Gamma(t,x;\,\tau,\xi)$ are based on the inequality $e^{-x}\le Cx^{-\beta}$ ($\beta>0$, $0<x\le\infty$) and are detailed in [17, pp. 137–38]. The result for a domain of class $C^{(1+\lambda)}$, $0<\lambda<1$ is as follows. If $0<\mu<1$, then

$$|\partial_{\nu,x}\Gamma(t,x,\tau,\xi)|\le\frac{C}{(t-\tau)^\mu|x-\xi|^{m+1-2\mu-\lambda}}, \tag{2.13}$$

where $m+1-2\mu-\lambda<m-1$ if $\mu>1-\lambda/2$. We use this and (2.12) to estimate the η-integral in (2.11); to estimate the r-part, we use the gamma formula

$$\int_\tau^t(t-r)^{-\alpha}(r-\tau)^{-\mu}\,dr=\frac{\Gamma(1-\alpha)\Gamma(1-\mu)}{\Gamma(2-\alpha-\mu)}(t-\tau)^{1-\alpha-\mu}, \tag{2.14}$$

valid for $\alpha,\,\mu<1$. The η-part of the integrals is estimated using (2.12). In the first step the exponent is $m-1-(m+1-2\mu-\lambda)-(m+1-2\mu-\lambda)=-m-3+4\mu+2\lambda$; the second exponent is $-m-5+6\mu+3\lambda$ and, in general, the nth exponent is $-m-(2n+1)+(2n+2)\mu+(n+1)\lambda=-m-1+2\mu+\lambda+n(2\mu+\lambda-2)=a+bn$, where $b>0$ if $\mu>1-\lambda/2$. For the r-integral we use formula (2.14) inductively for $\alpha=-(n-1+n\mu)$, $n=1,2,\dots$. The integral equals $\{\Gamma(n(1-\mu))\Gamma(1-\mu)/\Gamma((n+1)(1-\mu))\}(t-\tau)^{n-(n+1)\mu}$. The final result is

$$|\rho_n(t,x;\,\tau,\xi)|\le\frac{\Gamma(1-\mu)^{n+1}}{\Gamma((n+1)(1-\mu))}M_n(t-\tau)^{n-(n+1)\mu}|x-\xi|^{a+bn},$$

where the constant M_n is the product of all constants C_n, $C_{n-1},\dots,C_1$ arising from the estimate (2.12) at each step. Obviously, it is enough to show that the M_n are bounded by the powers of an absolute constant C. This in turn follows from the fact that the power $|x-\xi|$ in (2.12) becomes positive after a finite number of iterations; precisely, (2.12) is used for $\alpha=m+1-2\mu-\lambda$ fixed and $\beta=-a-bn$. Accordingly, $|x-\xi|^{a+bn}$ can be uniformly bounded by a C^{a+bn} if $n\ge-a/b$, where C is the diameter of Σ. We have thus established estimates sufficient to ensure the uniform convergence of a tail of (2.10). Because the powers of $t-\tau$ and $|x-\xi|$ become higher with the index of ρ_j, the final estimate for R is the same as that for ρ_0,

$$|R(t,x,\tau,\xi)|\le\frac{C}{(t-\tau)^\mu|x-\xi|^{m+1-2\mu-\lambda}}, \tag{2.15}$$

which implies

$$|\rho(t,x,\tau,\xi)| \le \frac{C}{(t-\tau)^{\mu}|x-\xi|^{m+1-2\mu-\lambda}}. \tag{2.16}$$

THEOREM 2.1. *Let C be a bounded domain of class $C^{(2)}$, $\mu > 1/2$. The Neumann function $N(t,x;\tau,\xi)$ $(\xi \in \partial\Omega)$ can be extended to $t > \tau$, $x \in \bar{\Omega}$, $\xi \in \partial\Omega$, $x \neq \xi$, is continuous there and satisfies*

$$|N(t,x,\tau,\xi)| \le \frac{C}{(t-\tau)^{\mu}|x-\xi|^{m-2\mu}}. \tag{2.17}$$

Proof. The fundamental solution $\Gamma(t,x;\tau,\xi)$ satisfies (2.17) [17, p. 134]. Thus, we only have to estimate $V(t,x;\tau,\xi)$, which we do using (2.12) and (2.16), noting that $m-1-(m-2\mu)-(m+1-2\mu-\lambda) = -(m+2-4\mu-\lambda) > -(m-2\mu)$ if $\mu > 1-\lambda/2$. The fundamental solution $\Gamma(t,x;\tau,\xi)$ is infinitely differentiable for $t \ge \tau$, $x \neq \xi$, thus only continuity of $V(t,x;\tau,\xi)$ has to be proved, and it is enough to show continuity in sets Υ_δ defined by

$$(t,x,\tau,\xi) \in ([\tau,T]\times\bar{\Omega})\times([\tau,T]\times\partial\Omega), \qquad (t-\tau\ge\delta,\ |x-\xi|\ge\delta),$$

where $\delta > 0$. Continuity is obvious for $\rho_0(t,x;\tau,\xi)$. Assume it has been proved for $\rho_j(t,x;\tau,\xi)$, and consider the integral (2.11) defining $\rho_{j+1}(t,x;\tau,\xi)$. Extend all functions as zero in $t \le \tau$, so that we may integrate over $[\tau,T]\times\partial\Omega$ in (2.11). If $\rho_{j+1}(t,x;\tau,\xi)$ is not continuous in Υ_δ, there exist sequences $\{(t_n,x_n,\tau_n,\xi_n)\} \subset \Upsilon_\delta$ with $\{(t_n,x_n,\tau_n,\xi_n)\} \to (\bar{t},\bar{x},\bar{\tau},\bar{\xi}) \in \Upsilon_\delta$ and

$$|\rho_{j+1}(t_n,x_n,\tau_n,\xi_n) - \rho_{j+1}(\bar{t},\bar{x},\bar{\tau},\bar{\xi})| \ge \varepsilon > 0. \tag{2.18}$$

Continuity of $\rho_j(r,\eta;\tau,\xi)$ for $\eta \neq \xi$ implies

$$\partial_{\nu,\eta}\Gamma(t_n,x_n;r,\eta)\rho_j(r,\eta;\tau_n,\xi_n) \to \partial_{\nu,\eta}\Gamma(\bar{t},\bar{x};r,\eta)\rho_j(r,\eta;\bar{\tau},\bar{\xi}) \tag{2.19}$$

almost everywhere in $[0,T]\times\partial\Omega$. On the other hand, this function is bounded by

$$C(t_n-r)^{-\mu}(r-\tau_n)^{n-(n+1)\mu}|x_n-\eta|^{m+1-2\mu-\lambda}|\eta-\xi_n|^{a+\iota n}$$

(extended to zero if $r > t_n$ or $r < \tau_n$). Now, the integrals of these functions are equicontinuous in $[\tau,T]\times\partial\Omega$, so Vitali's theorem applies to show that $\rho_j(t_n,x_n,\tau_n,\xi_n) - \rho_j(\bar{t},\bar{x},\bar{\tau},\bar{\xi})$ as $n \to \infty$. This contradicts (2.18) and thus shows continuity of $\rho_j(t,x;\tau,\xi)$ in Υ_δ. Using the estimates for the terms of (2.10), we obtain continuity of $R(t,x;\tau,\xi)$ in $t > \tau$, $x \neq \xi$. Arguing in a similar way first with (2.8), then with (2.7), we show continuity of $\rho(t,x;\tau,\xi)$ and then of $V(t,x;\tau,\xi)$ in $t > \tau$, $x \neq \xi$. This ends the proof of Theorem 2.1. □

3. Solutions of the nonlinear initial-boundary value problem. We consider (1.1)–(1.3) in a domain of class $C^{(2)}$ with $g(t,y)$ defined and continuous in $[0,T]\times\mathbb{R}$ and the control $u(\cdot,\cdot)$ in $L^\infty((0,T)\times\partial\Omega)$. By definition, $y(t,x)$ is a *weak solution* of (1.1)–(1.3) for $\zeta(\cdot) \in C(\bar{\Omega})$ if and only if it is continuous in $[0,T]\times\bar{\Omega}$ and satisfies

$$\begin{aligned} y(t,x) = &\int_\Omega N(t,x;0,\xi)\zeta(\xi)\,d\xi \\ &+ \int_{(0,t)\times\partial\Omega} N(t,x;\tau,\xi)\{g(\tau,y(\tau,\xi)) + u(\tau,\xi)\}\,d\sigma_\xi\,d\tau, \end{aligned} \tag{3.1}$$

where $N(t,x;\tau,\xi)$ is the Neumann function of $[0,T]\times\Omega$. For $x\in\partial\Omega$, (3.1) is an integral equation for the restriction $\phi(t,x)$ of $y(t,x)$ to $\partial\Omega$. Once solved, we construct $y(t,x)$ from (3.1).

Consider the two parameter family of operators

$$(\mathbf{N}(t,\tau)\psi)(x)=\int_{\partial\Omega}N(t,x;\tau,\xi)\psi(\xi)d\sigma_\xi \qquad (\tau<t) \tag{3.2}$$

for $\psi\in L^\infty(\partial\Omega)$. Theorem 2.1 shows that $\mathbf{N}(t,\tau)$ is a bounded operator from $L^\infty(\partial\Omega)$ into $C(\bar\Omega)$; moreover, $\mathbf{N}(t,\tau)$ is continuous in $\tau<t$ in the norm of $L(L^\infty(\partial\Omega),\,C(\bar\Omega))$ in $0\le\tau<t\le T$ and

$$\|\mathbf{N}(t,\tau)\|_{L(L^\infty(\partial\Omega),C(\bar\Omega))}\le B(t-\tau)^{-\mu} \qquad (0\le\tau<t\le T), \tag{3.3}$$

with $1/2\le\mu<1$, where $L(X,Y)$ is the Banach space of all bounded operators from the Banach space X into the Banach space Y. Let $\mathbf{M}(t,\tau)=\Pi\circ\mathbf{N}(t,\tau)$, $\Pi: C(\bar\Omega)\to C(\partial\Omega)$ the restriction operator; $\mathbf{M}(t,\tau)$ is continuous in the norm of the space $L(L^\infty(\partial\Omega),\,C(\partial\Omega))$ and satisfies the companion of (3.3) in that norm. The *solution operator* $\mathbf{S}(t,\tau)$ of the heat equation in Ω (with Neumann boundary conditions) is given by the first term of (2.6), that is

$$\mathbf{S}(t,\tau)\zeta(x)=\int_\Omega N(t,x;\tau,\xi)\zeta(\xi)d\xi \tag{3.4}$$

and is a bounded operator from $C(\bar\Omega)$ into $C(\bar\Omega)$; moreover, $\mathbf{S}(t,\tau)$ is differentiable in the norm of $L(C(\bar\Omega),\,C(\bar\Omega))$ for $t>\tau$ and strongly continuous and uniformly bounded in $0\le\tau\le t\le T$, with $S(\tau,\tau)=I$. Finally, $\mathbf{T}(t,\tau)=\Pi\circ\mathbf{S}(t,\tau)$; $\mathbf{T}(t,\tau)$ is differentiable in the norm of $L(C(\bar\Omega),\,C(\partial\Omega))$ in $0\le\tau<t\le T$ and strongly continuous for $t\ge\tau$. All these operators depend on $t-\tau$, although we keep the present notation in view of the treatment of time-dependent equations (see §6). The function $g(t,y)$ produces an operator $\mathbf{g}:[0,T]\times C(\partial\Omega)\to C(\partial\Omega)$ defined by $\mathbf{g}(t,\boldsymbol{\phi})(x)=g(t,\phi(t,x))$ that is continuous in t in the norm of $(C(\partial\Omega),\,C(\partial\Omega))$. Using this operator jargon, the integral equation to be solved is

$$\begin{aligned}\boldsymbol{\phi}(t)&=\mathbf{T}(t,0)\zeta+\int_0^t\mathbf{M}(t,\tau)\mathbf{u}(\tau)d\tau+\int_0^t\mathbf{M}(t,\tau)\mathbf{g}(\tau,\boldsymbol{\phi}(\tau))d\tau\\&=\mathbf{f}(t)+\int_0^t\mathbf{M}(t,\tau)\mathbf{g}(\tau,\boldsymbol{\phi}(\tau))d\tau,\end{aligned} \tag{3.5}$$

where $\boldsymbol{\phi}(t)(x)=\phi(t,x)$, $\mathbf{u}(t)(x)=u(t,x)$; setting $\mathbf{y}(t)=y(t,x)$, formula (3.1) becomes

$$\mathbf{y}(t)=\mathbf{S}(t,0)\zeta+\int_0^t\mathbf{N}(t,\tau)\mathbf{u}(\tau)d\tau+\int_0^t\mathbf{N}(t,\tau)\mathbf{g}(\tau,\boldsymbol{\phi}(\tau))d\tau. \tag{3.6}$$

Note that $\boldsymbol{\phi}(\cdot)\in C([0,T];\,C(\partial\Omega))$ (respectively, $\mathbf{y}(\cdot)\in C([0,T];\,C(\bar\Omega))$ if and only if $\phi(\cdot,\cdot)\in C([0,T]\times\partial\Omega)$ (respectively, $y(\cdot,\cdot)\in C([0,T]\times\bar\Omega)$); here, if X is a Banach space, $C([0,T];\,X)$ denotes the space of continuous X-valued functions defined in $[0,T]$ endowed with the supremum norm. On the other hand, the first integrals on the right sides of (3.5) and (3.6) need explanation because the function $\mathbf{u}(\cdot)$ may not be strongly measurable as an $L^\infty(\partial\Omega)$-valued function. We use the fact that $L^\infty((0,T)\times\partial\Omega)$ can be identified (algebraically and metrically) with the space $L^\infty_W(0,T;\,L^\infty(\partial\Omega))$ of all $L^1(\Omega)$-weakly measurable $L^\infty(\partial\Omega)$-essentially bounded functions defined in $0\le t\le T$ endowed with the essential supremum norm (if $\mathbf{u}(\cdot)\in L^\infty_W(0,T;\,L^\infty(\partial\Omega))$, then $\|\mathbf{u}(t)\|=\sup\langle y_n,\mathbf{u}(t)\rangle$, where $\{y_n\}$ is any sequence dense in the unit ball of $L^1(\partial\Omega)$, so that $t\to\|\mathbf{u}(t)\|$ is measurable).

LEMMA 3.1. *Let* $\mathbf{u}(\cdot) \in L^\infty_W(0,T;\, L^\infty(\partial\Omega))$. *Then*

$$\tau \to \mathbf{M}(t,\tau)\mathbf{u}(\tau) \qquad (0 \le \tau < t) \tag{3.7}$$

is a strongly measurable $C(\bar\Omega)$*-valued function.*

Proof. Because the space $C(\bar\Omega)$ is separable, it is enough to show that (3.7) is $C(\bar\Omega)^*$-weakly measurable, where the dual $C(\bar\Omega)^*$ is the space $\Sigma(\bar\Omega)$ of all finite Borel measures on $\bar\Omega$ endowed with the total variation norm. Let $\nu \in \Sigma(\bar\Omega)$. Then

$$\langle \nu, \mathbf{N}(t,\tau)\mathbf{u}(\tau)\rangle = \int_{\bar\Omega}\int_{\partial\Omega} N(t,x;\,\tau,\xi)\nu(dx)u(\tau,\xi)d\sigma_\xi.$$

Using the continuity properties of $N(t,x;\,\tau,\xi)$, we check that $\tau \to \langle \nu, \mathbf{N}(t,\tau)\mathbf{u}(\tau)\rangle$ is continuous in $0 < \tau < t$, which is more than enough. □

THEOREM 3.2. *Assume* $g(t,y)$ *satisfies a local Lipschitz condition in* y*; for every* $C > 0$ *there exists* $K = K(C)$ *such that*

$$|g(t,y') - g(t,y)| \le K|y' - y| \qquad (0 \le t \le T,\ |y|,\ |y'| \le C). \tag{3.8}$$

Then a unique solution of (3.5) *exists in an interval* $0 \le t \le T_0$, $T_0 \le T$.

Proof. Inequality (3.8) implies a local Lipschitz condition for $\mathbf{g}$. Lemma 3.1 and the properties of the operator $\mathbf{M}(t,\tau)$ imply that $\mathbf{f}(t)$ in (3.5) belongs to $C([0,T];\, C(\partial\Omega))$. We consider the map

$$\mathbf{L}\boldsymbol{\phi}(t) = \mathbf{f}(t) + \int_0^t \mathbf{M}(t,\tau)\mathbf{g}(\tau,\boldsymbol{\phi}(\tau))d\tau$$

in a closed ball $B(\mathbf{f}(\cdot),\rho)$ of $C([0,T_0];\, C(\partial\Omega))$ (endowed with the supremum norm) where $\rho > 0$. Using the companion of (3.3) for $\mathbf{M}(t,\tau)$, local boundedness of $\mathbf{g}$, and the local Lipschitz condition, we show that $\mathbf{L}$ is a contraction for T_0 small enough. Thus a unique fixed point $\boldsymbol{\phi} = \mathbf{L}\boldsymbol{\phi}$ exists. The estimations are based on the *generalized Gronwall inequality* [20, p. 188]: if $b \ge 0$, $\beta > -1$, $a(\cdot)$, $b(\cdot) \in L^1_{\mathrm{loc}}(0,T)$, and

$$u(t) \le a(t) + b\int_\tau^t (t-r)^\beta u(r)dr \qquad (\tau \le t \le T), \tag{3.9}$$

then there exists a constant c depending only on β and T such that

$$u(t) \le a(t) + c\int_\tau^t (t-r)^\beta a(r)dr \qquad (\tau \le t \le T). \tag{3.10}$$

The solution $\boldsymbol{\phi}(t)$ of (3.5) is not necessarily defined in $0 \le t \le T$. If it is not, we can define a maximal interval of existence $0 \le t < T_m$ at the end of which $\boldsymbol{\phi}(t)$ must blow up.

LEMMA 3.3. *Let* $\boldsymbol{\phi}(t)$ *be a solution of* (3.5) *in* $0 \le t < \bar t$. *Assume that* $\|\boldsymbol{\phi}(t)\| \le C$. *Then* $\boldsymbol{\phi}(t)$ *can be continued as a solution in* $0 \le t \le \tilde t > \bar t$.

The proof is classical. We use the integral equation to show that $\{\boldsymbol{\phi}(t);\ \bar t - \delta \le t < \bar t\}$ is Cauchy as $t \to \bar t$, so that $\lim_{t\to\bar t}\boldsymbol{\phi}(t)$ exists and $\boldsymbol{\phi}(t)$ can be continued as a solution to $0 \le t \le \bar t$. Then we solve

$$\begin{aligned}\boldsymbol{\psi}(t) = \mathbf{T}(t,0)\zeta &+ \int_0^t \mathbf{M}(t,\tau)\mathbf{u}(\tau)d\tau \\ &+ \int_0^{\bar t} \mathbf{M}(t,\tau)\mathbf{g}(\tau,\boldsymbol{\phi}(\tau))d\tau + \int_{\bar t}^t \mathbf{M}(t,\tau)\mathbf{g}(\tau,\boldsymbol{\psi}(\tau))d\tau\end{aligned}$$

and extend $\boldsymbol{\phi}(t)$ with $\boldsymbol{\psi}(t)$. □

Consider the linear nonhomogeneous version of (1.1)–(1.3),

$$\xi_t(t,x) = \Delta\xi(t,x) \qquad (x \in \Omega,\ 0 < t < \bar{t}), \tag{3.11}$$

$$\xi(0,x) = \zeta(x) \qquad (x \in \Omega), \tag{3.12}$$

$$\partial_\nu \xi(t,x) = a(t,x)\xi(t,x) + \delta(t-s)v(x) \qquad (x \in \partial\Omega,\ 0 < t \le T), \tag{3.13}$$

where $0 < s < T$, $\zeta(\cdot) \in C(\bar{\Omega})$, $a(\cdot,\cdot) \in C([0,T] \times \partial\Omega)$, $v(\cdot) \in L^\infty(\partial\Omega)$, and $\delta(t)$ is the Dirac delta. This type of boundary condition is not covered by the preceding nonlinear theory, but we can stretch the treatment to accommodate (3.13). The equation for $\boldsymbol{\psi}(t) = \Pi \circ \boldsymbol{\xi}(t)$ is

$$\begin{aligned} \boldsymbol{\psi}(t) &= \mathbf{T}(t,0)\zeta + h(t-s)\mathbf{M}(t,s)v \\ &\quad + \int_0^t \mathbf{M}(t,\tau)\mathbf{a}(\tau)\boldsymbol{\psi}(\tau)d\tau \\ &= \mathbf{f}(t) + h(t-s)\int_s^t \mathbf{M}(t,\tau)\mathbf{a}(\tau)\boldsymbol{\psi}(\tau)d\tau, \end{aligned} \tag{3.14}$$

where $h(t) = 1$ for $t \ge 0$, $h(t) = 0$ for $t < 0$, and $\mathbf{a}(t)$ is the linear bounded operator defined by $(\mathbf{a}(t)\boldsymbol{\phi})(x) = a(t,x)\phi(t,x)$. By definition, the solution of (3.14) is given by

$$\begin{aligned} \boldsymbol{\xi}(t) &= \mathbf{S}(t,\tau)\zeta + h(t-s)\mathbf{N}(t,s)v \\ &\quad + \int_s^t \mathbf{N}(t,\tau)\mathbf{a}(\tau)\boldsymbol{\psi}(\tau)d\tau. \end{aligned} \tag{3.15}$$

In view of (3.3), $\mathbf{f}(t)$ is continuous in $t < s$ and $t > s$ with $\|\mathbf{f}(t)\| \le C(t-s)^{-\mu}$ $(t > s)$. Extending $\mathbf{f}(t)$ to $t = s$ by $\mathbf{f}(s) = \lim_{t \to s-} \mathbf{f}(t)$, we may actually consider $\mathbf{f}(t)$ continuous in $t \le s$. This time, we construct $\boldsymbol{\psi}(\cdot)$ in the space $C_{s,\mu}([0,T];\, C(\partial\Omega))$ of all $C(\partial\Omega)$-valued $\boldsymbol{\psi}(t)$ continuous in $0 \le t \le s$ and $s < t \le T$ endowed with the weighted supremum norm with weight $\omega(t,\mu) = 1$ $(t \le s)$, $\omega(t,\mu) = (t-s)^\mu$, $(t > s)$. Equation (3.14) may be directly solved by successive approximations in the interval $0 \le t \le T$ starting with $\boldsymbol{\psi}_0(t) = \mathbf{f}(t)$, estimations based on the gamma formula. The solution belongs to $C_{s,\mu}([0,T];\, C(\partial\Omega))$, thus

$$\|\boldsymbol{\psi}(t)\| \le C\omega(t,-\mu) \tag{3.16}$$

so that, in view of (3.6) and the gamma formula, the solution $\mathbf{z}(t)$ will satisfy a bound of the same form.

We will also have to stretch the (linear) theory to accommodate the backwards initial value (or final value) problem

$$z_t(t,x) = -\Delta z(t,x) - f(t,x) \qquad (x \in \Omega,\ 0 \le t \le \bar{t}), \tag{3.17}$$

$$z(\bar{t},\cdot) = \nu \qquad (x \in \Omega), \tag{3.18}$$

$$\partial_\nu \xi(t,x) = b(t,x)\xi(t,x) + v(t,x) \qquad (x \in \partial\Omega,\ 0 \le t \le \bar{t}), \tag{3.19}$$

with $\nu \in \Sigma(\bar{\Omega})$, $f(\cdot,\cdot) \in C([0,T] \times \bar{\Omega})$, $b(\cdot,\cdot) \in C([0,T] \times \partial\Omega)$, and $v(\cdot,\cdot) \in L^\infty((0,T) \times \partial\Omega)$. Because of the symmetry (2.5) of the Neumann function, if $\mathbf{S}^-(t,\tau)$, $\mathbf{N}^-(t,\tau)$, $\mathbf{M}^-(t,\tau)$ $(t < \tau)$ are the operators associated with (3.17)–(3.19), we have $\mathbf{S}^-(t,\tau) = \mathbf{S}(\tau,t)$, with

similar equalities for $\mathbf{N}$ and $\mathbf{M}$. Accordingly, the associated integral equation for $\boldsymbol{\psi}(t) = \Pi \circ \mathbf{z}(t)$ is

$$\begin{aligned}\boldsymbol{\psi}(t) &= \mathbf{T}(\bar{t}, t)\nu + \int_t^{\bar{t}} \mathbf{T}(\tau, t)\mathbf{f}(\tau)d\tau \\ &+ \int_t^{\bar{t}} \mathbf{M}(\tau, t)\mathbf{v}(\tau)d\tau + \int_t^{\bar{t}} \mathbf{M}(\tau, t)\mathbf{a}(\tau)\boldsymbol{\psi}(\tau)d\tau.\end{aligned} \tag{3.20}$$

As to the first term, we extend $\mathbf{S}(t, \tau)$ to $\Sigma(\bar{\Omega})$ by

$$(\mathbf{S}(t, \tau)\nu)(x) = \int_{\bar{\Omega}} N(t, \xi; \tau, x)\nu(d\xi) \tag{3.21}$$

because $N(t, x; \tau, \cdot)$ is continuous in $(x, \xi) \in \bar{\Omega} \times \bar{\Omega}$ for $t > \tau$. Moreover, $\mathbf{S}(t, \tau)\Sigma(\bar{\Omega}) \subseteq C(\bar{\Omega})$ and $\mathbf{S}(t, \tau)$ are continuous in the norm of $L(\Sigma(\bar{\Omega}), C(\bar{\Omega}))$ in $t > \tau$; again, $\mathbf{T}(t, \tau) = \Pi \circ \mathbf{S}(t, \tau)$ and $\mathbf{T}(t, \tau) : \Sigma(\bar{\Omega}) \to C(\partial\Omega)$. It can be shown by semigroup theoretic methods [4] that $\|\mathbf{S}(t, \tau)\|$ is uniformly bounded in $t > \tau$ and that $\mathbf{S}(t, \tau)\nu \to \nu$ $C(\bar{\Omega})$-weakly in $\Sigma(\bar{\Omega})$. We solve (3.20) by successive approximations in the space $C([0, \bar{t}); C(\partial\Omega))$ of all $C(\partial\Omega)$-valued functions $\boldsymbol{\psi}(t)$ bounded and continuous in $0 \leq t < \bar{t}$ endowed with the supremum norm and obtain the solution from the backwards analogue of (3.6).

4. Existence theory for the optimal control problem. We denote by $\mathbf{U}_{\text{ad}}(0, T; U)$ the *admissible control space*, that is, the set of all $\mathbf{u}(\cdot) \subseteq L^\infty_W(0, T; L^\infty(\partial\Omega))$ satisfying the control constraint (1.4) almost everywhere in $0 \leq t \leq T$; if $\mathbf{u}(\cdot) \in \mathbf{U}_{\text{ad}}(0, T; U)$ (or, more generally, if $\mathbf{u}(\cdot) \in L^\infty_W(0, T; L^\infty(\partial\Omega))$), $\boldsymbol{\phi}(t, \mathbf{u})$ is the $C(\partial\Omega)$-valued solution of the integral equation (3.5) corresponding to a the control $\mathbf{u}(\cdot) \in L^\infty_W(0, T; L^\infty(\partial\Omega))$. Likewise, $\mathbf{y}(t, \mathbf{u})$ is the $C(\bar{\Omega})$-valued solution defined by (3.6). Both functions need not be defined in the whole interval $0 \leq t \leq T$.

LEMMA 4.1. *The operators* $\mathbf{N}(t, \tau) : L^\infty(\partial\Omega) \to C(\bar{\Omega})$ *and* $\mathbf{M}(t, \tau) : L^\infty(\partial\Omega) \to C(\partial\Omega)$ *are compact.*

Proof. We only have to consider $\mathbf{N}(t, \tau)$. Let $\{u_n(\cdot)\} \subset L^\infty(\partial\Omega)$ be $L^1(\partial\Omega)$-weakly convergent to zero. If $\mathbf{N}(t, \tau)$ is not convergent to zero in $C(\bar{\Omega})$, there exists a sequence $\{x_n\} \subset \bar{\Omega}$ such that

$$|\mathbf{N}(t, \tau)u_n(x_n)| \geq \varepsilon > 0 \tag{4.1}$$

(we may assume $x_n \to \bar{x} \in \bar{\Omega}$). However, by (2.17) and Vitali's theorem, $N(t, x_n, \tau, \cdot) \to N(t, \bar{x}, \tau, \cdot)$ in $L^1(\partial\Omega)$, thus $\mathbf{N}(t, \tau)u_n(x_n) \to 0$, a contradiction. □

THEOREM 4.2. *The operator*

$$(\boldsymbol{\Delta}u)(t) = \int_0^t \mathbf{N}(t, \tau)\mathbf{u}(\tau)d\tau$$

from $L^\infty_W(0, T; L^\infty(\partial\Omega))$ *into* $C([0, T]; C(\bar{\Omega}))$ *is compact.*

The proof is essentially the same as that of Lemma 6.1 in [9] and is omitted; the result implies compactness of the operator $\boldsymbol{\Lambda} = \Pi \circ \boldsymbol{\Delta}$.

A cost functional $y_0(t, \mathbf{u})$ is *weakly lower semicontinuous* if

$$y_0(t, \bar{\mathbf{u}}) \leq \limsup_{n \to \infty} y_0(t_n, \mathbf{u}_n) \tag{4.2}$$

for every sequence $\{t_n\} \subseteq \mathbb{R}$ with $t_n \to \bar{t}$ and every sequence $\{\mathbf{u}_n\} \subseteq L^\infty_W(0, T; L^\infty(\partial\Omega))$ with $\mathbf{u}_n \to \bar{\mathbf{u}}$ $L^1(0, T; L^1(\partial\Omega))$-weakly.

Let m be the minimum of the cost functional $y_0(t, \mathbf{u})$ subject to the control constraint (1.4) and the target condition (1.5). Assuming that $-\infty < m < \infty$, a sequence $\{\mathbf{u}^n\}$, $\mathbf{u}^n \in \mathbf{U}_{\text{ad}}(0, t_n; U)$ is called a *minimizing sequence* if and only if

$$\limsup_{n\to\infty} y_0(t_n, \mathbf{u}^n) \le m, \qquad \lim_{n\to\infty} \operatorname{dist}(\mathbf{y}(t_n, \mathbf{u}^n), Y) \to 0. \tag{4.3}$$

THEOREM 4.3. *Let $\mathbf{U}_{\text{ad}}(0,T;U)$ be $L^1(0,T;L^1(\partial\Omega))$-weakly compact in $L^\infty_W(0,T;L^\infty(\partial\Omega))$, and assume $y_0(t,\mathbf{u})$ is weakly lower semicontinuous. Let $\{\mathbf{u}^n(\cdot)\}$, $\mathbf{u}^n(\cdot) \in \mathbf{U}_{\text{ad}}(0,t_n;U)$ be a minimizing sequence with $t_n \to \bar t$ and $\|\boldsymbol{\phi}(t,\mathbf{u}^n)\|_{C(\partial\Omega)} \le C$ $(0 \le t \le \bar t)$. Then (a subsequence of) $\{\mathbf{u}^n\}$ converges $L^1(0,\bar t;L^1(\partial\Omega))$-weakly to an optimal control $\bar{\mathbf{u}}(\cdot) \in \mathbf{U}_{\text{ad}}(0,\bar t;U)$.*

Proof. We write (3.5) for each t_n, u^n, $\boldsymbol{\phi}(t,\mathbf{u}^n)$ in the interval $0 \le t \le \bar t$; if $t_n < \bar t$, we define, say, $\mathbf{u}^n(t) = \mathbf{u}^n(t_n)$, $\boldsymbol{\phi}(t,\mathbf{u}^n) = \boldsymbol{\phi}(t_n,\mathbf{u}^n)$ in $t \ge t_n$. We then obtain

$$\tilde{\boldsymbol{\phi}}(t,\mathbf{u}^n) = \mathbf{T}(t,s)\zeta + \int_0^{\bar t} \mathbf{M}(t,\tau)\mathbf{u}^n(\tau)d\tau + \int_0^{\bar t} \mathbf{M}(t,\tau)g(\tau,\boldsymbol{\phi}(\tau,\mathbf{u}^n))d\tau \tag{4.4}$$

in $0 \le t \le \bar t$, where if $t_n \ge \bar t$, $\tilde{\boldsymbol{\phi}}(t,\mathbf{u}^n) = \boldsymbol{\phi}(t,\mathbf{u}^n)$. If $t_n \le \bar t$, $\tilde{\boldsymbol{\phi}}(t,\mathbf{u}^n)$ is defined by (4.4) in terms of the extended $\boldsymbol{\phi}$. Passing, if necessary, to a subsequence, we may assume that $\mathbf{u}^n(\cdot) \to \bar{\mathbf{u}}(\cdot) \in \mathbf{U}_{\text{ad}}(0,\bar t;U)$ $L^1(0,T;L^1(\partial\Omega))$-weakly. Using Theorem 4.2 and passing again to a subsequence, we deduce that $\tilde{\boldsymbol{\phi}}(\cdot,\mathbf{u}^n) \to \boldsymbol{\phi}(\cdot) \in C([0,\bar t];\partial\Omega)$ in the norm of $C([0,\bar t];\partial\Omega)$. Accordingly, the limit of (4.4) is the integral equation (3.5) for $\boldsymbol{\phi}(t)$ and $\bar{\mathbf{u}}(\cdot)$, which shows that $\boldsymbol{\phi}(t) = \boldsymbol{\phi}(t,\bar{\mathbf{u}})$ and ends the proof. □

The proof of Theorem 4.3 depends essentially on the fact that controls enter linearly into the boundary condition. If this is not the case (for instance, if the boundary condition is of the form $\partial_\nu y(t,x) = g(t,y(t,x),u(t,x))$, the infinite-dimensional theory of relaxed controls [12], [13] can be applied.

We examine global existence for (3.5). Lemma 4.4 below shows that it extends from a control $\bar{\mathbf{u}}$ to neighboring controls.

LEMMA 4.4. *Let $\bar{\mathbf{u}}(\cdot) \in L^\infty_W(0,\bar t;L^\infty(\partial\Omega))$ be such that $\boldsymbol{\phi}(t,\bar{\mathbf{u}})$ exists in $0 \le t \le \bar t$, and let $q > 1/(1-\mu)$, μ the constant in (3.3). Then, if $\|\mathbf{u}-\bar{\mathbf{u}}\|_{L^q(0,\bar t;L^q(\partial\Omega))} \le \rho$, the solution $\boldsymbol{\phi}(t,\mathbf{u}^n)$ of (3.5) exists in $0 \le t \le \bar t$. Moreover, if $\mathbf{v}(\cdot) \in L^\infty_W(0,\bar t;L^\infty(\partial\Omega))$ is such that $\|\mathbf{v}-\bar{\mathbf{u}}\|_{L^q(0,\bar t;L^q(\partial\Omega))} \le \rho$ as well, then*

$$\|\boldsymbol{\phi}(t,\mathbf{v}) \to \boldsymbol{\phi}(t,\mathbf{u})\| \le C\|\mathbf{v}-\mathbf{u}\|_{L^q(0,\bar t;L^q(\partial\Omega))}. \tag{4.5}$$

Proof. Let $[0,t_0]$ be the maximal interval where $\boldsymbol{\phi}(t,\mathbf{u})$ exists and satisfies $\|\boldsymbol{\phi}(t,\mathbf{u}) - \boldsymbol{\phi}(t,\bar{\mathbf{u}})\| \le 1$. Using (3.5) and estimating via Hölder's inequality in the first integral and local Lipschitz continuity of $q(t,y)$ in the second we obtain

$$\|\boldsymbol{\phi}(t,\mathbf{u}) - \boldsymbol{\phi}(t,\bar{\mathbf{u}})\| \le C\|\mathbf{u}-\bar{\mathbf{u}}\|_{L^q((0,T)\times\partial\Omega)} + C'\int_0^t (t-\tau)^{-\mu}\|\boldsymbol{\phi}(\tau,\mathbf{u}) - \boldsymbol{\phi}(\tau,\bar{\mathbf{u}})\|d\tau.$$

Taking ρ small enough and using the generalized Gronwall's inequality, we produce a contradiction unless $[0,t_0] = [0,\bar t]$ and prove (4.5) for $\mathbf{v} = \bar{\mathbf{u}}$. A similar argument deals with the general pair $\mathbf{v}, \mathbf{u}$. □

Formula (3.6) produces an estimate of the form of (4.5) for solutions $\mathbf{y}(t)$ and shows in particular (taking $\mathbf{v} \in C([0,T];\partial\Omega)$ and approximating with bounded $L^\infty_W(0,\bar t;L^\infty(\partial\Omega))$-norm) that the following corollary holds.

COROLLARY 4.5. *A weak solution $\mathbf{y}(t,\mathbf{u})$ of (1.1)–(1.3) can be uniformly approximated in $[0,\bar t]\times\bar\Omega$ by semistrong solutions $\mathbf{y}_n(t,\mathbf{u}_n)$ with $\mathbf{u}_n(\cdot) \in C([0,T]\times\partial\Omega)$ uniformly bounded. If $\mathbf{u}(t) \ge 0$ almost everywhere, the $\mathbf{u}_n(t)$ can be chosen such that $\mathbf{u}_n(t) > 0$.*

In the result below, $\mathbf{U}_{\text{ad}}(0,T;\,U)$ is equipped with the distance

$$d(\mathbf{u},\mathbf{v}) = \text{meas}\{t;\ \mathbf{u}(t) \neq \mathbf{v}(t)\}. \tag{4.6}$$

(Note that $\{t;\ \mathbf{u}(t) \neq \mathbf{v}(t)\}$ is the union of the countable family of measurable sets $\{t;\ \langle y_n, \mathbf{u}(t)\rangle \neq \langle y_n, \mathbf{v}(t)\rangle\}$, $\{y_n\}$ a countable dense set in $L^1(\partial\Omega)$, thus is measurable.)

COROLLARY 4.6. *Let $\bar{\mathbf{u}}(\cdot) \in \mathbf{U}_{\text{ad}}(0,\bar{t};\,U)$ be such that $\boldsymbol{\phi}(t,\bar{\mathbf{u}})$ exists in $0 \leq t \leq \bar{t}$. Then there exists $\rho > 0$ such that if $d(\mathbf{u},\bar{\mathbf{u}}) \leq \rho$ then $\boldsymbol{\phi}(t,\mathbf{u})$ also exists in $0 \leq t \leq \bar{t}$. Moreover,*

$$\|\boldsymbol{\phi}(t,\mathbf{v}) - \boldsymbol{\phi}(t,\mathbf{u})\| \leq C d(\mathbf{v},\mathbf{u})^{1-\mu} \qquad (0 \leq t \leq \bar{t}, \mathbf{u}, \mathbf{v} \in B(\bar{\mathbf{u}},\rho)). \tag{4.7}$$

Proof. Take $\mu > \mu' > 1/2$, $q = 1/(1-\mu) > 1/(1-\mu')$, and apply Lemma 4.4, noting that $\|\mathbf{v} - \mathbf{u}\|_{L^q(0,\bar{t};L^q(\partial\Omega))} \leq C d(\mathbf{v},\mathbf{u})^{1-\mu}$. □

LEMMA 4.7. *Assume $\zeta(x) \geq 0$, $g(t,0) = 0$, $g(t,y) < 0$ for $y \neq 0$, and let $\mathbf{u}(\cdot) \in L^\infty_W(0,\bar{t};\,L^\infty(\partial\Omega))$, $\mathbf{u}(t) \geq 0$ almost everywhere. Then the solution $\boldsymbol{\phi}(t,\mathbf{u})$ of* (3.5) *exists in $0 \leq t \leq \bar{t}$ and satisfies*

$$0 \leq \boldsymbol{\phi}(t,\mathbf{u}) \leq \|\zeta\| + Ct^{1-\mu}\|u\| \qquad (0 \leq t \leq \bar{t}). \tag{4.8}$$

Proof. We solve first with boundary condition $\partial_\nu y(t,\bar{x}) = g(t,y(t,x)\text{sign } y(t,x) + u(t,x)$ and show that the solution $y(t,x)$ is nonnegative. By Lemma 4.3 we may assume that $\mathbf{u}(\cdot) \in C([0,T]\times\partial\Omega)$ and thus that the solution is semistrong. Let $[0,t_0]$ be an interval where $\boldsymbol{\phi}(t,\mathbf{u})$ exists. By the maximum principle, $y(t,x,u)$ must attain its minimum at $\bar{\Omega} \cup ([0,T]\times\partial\Omega)$. If the minimum lies in $\bar{\Omega}$, it is nonnegative. If $y(t,x)$ attains a negative minimum $\bar{x}$ in $[0,T]\times\partial\Omega$, then $\partial_\nu y(t,\bar{x}) \leq 0$, which contradicts the boundary condition. It follows then that $y(t,x)$ must coincide with the (unique) solution with the original boundary condition $\partial_\nu y(t,x) = g(t,y(t,x)) + u(t,x)$. To show that $\boldsymbol{\phi}(t,\mathbf{u})$ exists in $0 \leq t \leq \bar{t}$ we use Lemma 3.3 in combination with a priori bound. This bound is obtained from (3.5); in fact, by positivity of the kernels, the third term is nonpositive so that we have $\|\boldsymbol{\phi}(t,\mathbf{u})\| \leq \|\zeta\| + Ct^{1-\mu}\|\mu\|$. This ends the proof. □

THEOREM 4.8. *Let $\xi(t,x)$ be a solution of* (3.11)–(3.13) *with $\zeta = 0$ and let $z(t,x)$ be a solution of* (3.17)–(3.19). *Then*

$$\begin{aligned}\int_{\bar{\Omega}} \xi(\bar{t},x)\nu(dx) &= \int_{\partial\Omega} v(x)z(s,x)d\sigma \\ &+ \int_{(s,\bar{t})\times\partial\Omega} \{\xi(\tau,x)a(\tau,x)z(\tau,x) - \xi(\tau,x)(b(\tau,x)z(\tau,x) + v(\tau,x))\}d\sigma\,d\tau \\ &- \int_{(s,\bar{t})\times\Omega} \xi(\tau,x)f(\tau,x)dx\,d\tau.\end{aligned}$$

The proof is a consequence of the divergence theorem for smooth solutions. For general solutions, an approximation argument is used [24].

5. Directional derivatives. Let V be a metric space, E be a Banach space, and $g : V \to E$. An element $\xi \in E$ is a (one-sided) *directional derivative* of g at $u \in V$ if and only if there exists $u : [0,\delta] \to V$ with $d(u(h),u) \leq h$ and

$$g(u(h)) = g(u) + h\xi + o(h) \quad \text{as } h \to 0+.$$

The set of all directional derivatives of g at u is denoted $\partial g(u)$; it is star-shaped and closed. We compute below certain directional derivatives of the function $\mathbf{f} : \mathbf{U}_{\text{ad}}(0,\bar{t};U) \to E$ defined by

$$\mathbf{f}(\mathbf{u}) = \mathbf{y}(\bar{t},\mathbf{u}) \in C(\bar{\Omega}), \tag{5.1}$$

where $\mathbf{u} \in \mathbf{U}_{\text{ad}}(0, \bar{t}; U)$ is such that $\boldsymbol{\phi}(t, \mathbf{u})$ (thus $\mathbf{y}(t, \mathbf{u})$) exists in $0 \le t \le \bar{t}$. The space $\mathbf{U}_{\text{ad}}(0, \bar{t}; U)$ is equipped with the distance (4.6). Derivatives are constructed by means of "multispike perturbations" $\mathbf{u}(h) = \mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}}(t)$ defined by

$$\mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}}(t) = \begin{cases} v_j & (s_j - p_j h \le t \le s_j,\ j = 1, 2, \dots, m), \\ \mathbf{u}(t) & \text{elsewhere}, \end{cases} \tag{5.2}$$

where $\mathbf{s} = (s_1, s_2, \dots, s_m)$, $0 < s_1 < s_2 < \cdots < s_m < \bar{t}$, $\mathbf{p}$ is a probability vector $(p_1, p_2, \dots, p_m)$, $p_j \ge 0$, $\sum p_j = 1$, and $\mathbf{v} = (v_1, v_2, \dots, v_m)$, $v_j \in U$. The number m of spikes is arbitrary.

THEOREM 5.1. *Assume that $g(t, y)$ is differentiable with respect to y and the partial $\partial_y g(t, y)$ is continuous in $[0, \bar{t}] \times \mathbb{R}$. Then there exists a set e of full measure in $0 \le s \le \bar{t}$ such that if $s_j \in e$ $(j = 1, 2, \dots, m)$, then*

$$\boldsymbol{\xi}(t, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v}) = \lim_{h \to 0+} \frac{\mathbf{y}(t, \mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}}) - \mathbf{y}(t, \mathbf{u})}{h} \tag{5.3}$$

exists in the norm of $C(\bar{\Omega})$, convergence being uniform outside of the intervals $|s - s_j| < \varepsilon$, $j = 1, 2, \dots, m$ for any $\varepsilon > 0$. We have

$$\boldsymbol{\xi}(t, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v}) = \sum_{j=1}^{m} p_j \boldsymbol{\xi}(t, s_j, \mathbf{u}, v_j), \tag{5.4}$$

where $\xi(t, x, s, u, v) = \boldsymbol{\xi}(t, s, \mathbf{u}, v)(x)$ $(0 \le s \le \bar{t},\ v \in U)$ is the solution of the linear initial value problem

$$\xi_t(t, x, s, u, v) = \Delta_x \xi(t, x, s, u, v) \qquad ((t, x) \in (0, \bar{t}] \times \Omega), \tag{5.5}$$

$$\xi(0, x, s, u, v) = 0 \qquad (x \in \Omega), \tag{5.6}$$

$$\begin{aligned} \partial_\nu \xi(t, x, s, u, v) &= \partial_y g(t, y(t, x, u)) \xi(t, x, s, u, v) \\ &\quad + \delta(t - s)(v(x) - u(s, x)) \qquad ((t, x) \in (0, \bar{t}] \times \partial\Omega). \end{aligned} \tag{5.7}$$

Moreover,

$$\|h^{-1}(\mathbf{y}(t, \mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}}) - \mathbf{y}(t, \mathbf{u})) - \boldsymbol{\xi}(t, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v})\|_{C(\bar{\Omega})} \le C \sum_{j=1}^{m} \kappa(t, s_j, h, 1 - 2\mu) \tag{5.8}$$

for sufficiently small h, with $\kappa(t, h, s, \beta) = h(t - (s - h))$ for $\beta \ge 0$; for $\beta < 0$,

$$\kappa(t, h, s, \beta) = \begin{cases} 0 & (0 \le t \le s - h), \\ (t - (s - h))^\beta & (s - h \le t \le s), \\ (t - s)^\beta & (s \le t \le \bar{t}). \end{cases} \tag{5.9}$$

We note that, in view of the representation (3.6) for solutions of the nonlinear problem and the corresponding formula (3.15) for the linear problem, it is enough to prove (5.3), (5.4), and (5.8) for $\boldsymbol{\phi}(t, \mathbf{u})$, $\boldsymbol{\phi}(t, \mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}})$, $\boldsymbol{\psi}(t, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v})$, where $\boldsymbol{\phi}(t, \mathbf{u})$ (respectively, $\boldsymbol{\phi}(t, \mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}})$) is the boundary restriction of $\mathbf{y}(t, u)$ (respectively, $\mathbf{y}(t, \mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}})$) and

$$\boldsymbol{\psi}(t, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v}) = \sum_{j=1}^{m} p_j \boldsymbol{\psi}(t, s_j, u, v_j),$$

$\boldsymbol{\psi}(t,s,\mathbf{u},v)$ is the solution of the integral equation

$$\begin{aligned}\boldsymbol{\psi}(t,s,\mathbf{u},v) &= h(t-s)\mathbf{M}(t,s)(v-\mathbf{u}(s))\\ &\quad+\int_0^t \mathbf{M}(t,\tau)\partial_y\mathbf{g}(\tau,\boldsymbol{\phi}(\tau,\mathbf{u}))\boldsymbol{\psi}(\tau,s,\mathbf{u},v)d\tau,\end{aligned}\tag{5.10}$$

where $\partial_y\mathbf{g}(t,\boldsymbol{\phi})(x) = \partial_y g(t,\phi(x))$ ($\partial_y\mathbf{g}$ is the Fréchet derivative of $\mathbf{g}$). Also, existence of $\boldsymbol{\phi}(t,\mathbf{u}_{\mathbf{s,p},h,\mathbf{v}})$ in $0\le t\le \bar t$ is covered by Lemma 4.6. The proof of Theorem 5.1 is very similar to that of [11, Thm. 5.2], thus we only sketch it. In particular, we only treat the case of a single spike $\mathbf{u}_{s,h,v}(t) = v\in U$ $(s-h\le t\le s)$, $\mathbf{u}_{s,h,v}(t) = \mathbf{u}(t)$ elsewhere (for justification of a similar simplification see [11, §5]).

Define a $C(\partial\Omega)$-valued function by

$$\boldsymbol{\eta}(t,s,h) = h^{-1}(\boldsymbol{\phi}(t,\mathbf{u}_{s,h,v}) - \boldsymbol{\phi}(t,\mathbf{u})) - \boldsymbol{\psi}(t,s,\mathbf{u},v)).$$

Obviously, $\boldsymbol{\eta}(t,s,h) = 0$ for $t < s-h$. For $t\ge s-h$,

$$\begin{aligned}\boldsymbol{\eta}(t,s,h) =& \int_0^t \mathbf{M}(t,\tau)\partial_y\mathbf{g}(\tau,\boldsymbol{\phi}(\tau,\mathbf{u}))\boldsymbol{\eta}(\tau,s,h)d\tau\\ &+\int_0^t \mathbf{M}(t,\tau)h^{-1}\boldsymbol{\rho}(\tau,\boldsymbol{\phi}(\tau,\mathbf{u}),\boldsymbol{\phi}(\tau,\mathbf{u}_{s,h,v}))d\tau\\ &+\frac{1}{h}\int_{s-h}^s \mathbf{M}(t,\tau)(v-\mathbf{u}(\tau))d\tau - \mathbf{M}(t,s)(v-\mathbf{u}(s)),\end{aligned}\tag{5.11}$$

where $\boldsymbol{\rho}(t,\boldsymbol{\phi},\boldsymbol{\phi}') = \mathbf{g}(t,\boldsymbol{\phi}') - \mathbf{g}(t,\boldsymbol{\phi}) - \partial_y\mathbf{g}(t,\boldsymbol{\phi})(\boldsymbol{\phi}'-\boldsymbol{\phi})$. Taking norms,

$$\begin{aligned}\|\boldsymbol{\eta}(t,s,h)\| \le C\int_{s-h}^t (t-\tau)^{-\alpha}\|\boldsymbol{\eta}(\tau,s,h)\|d\tau\\ +\|\boldsymbol{\delta}_1(t,s,h)\| + \|\boldsymbol{\delta}_2(t,s,h)\|,\end{aligned}\tag{5.12}$$

where $\boldsymbol{\delta}_1$ (respectively, $\boldsymbol{\delta}_2$) is the second integral on the right-hand side of (5.11) (respectively, the combination of the third integral and the nonintegral term). Writing $r(t,s,h) = \|\boldsymbol{\delta}_1(t,s,h)\| + \|\boldsymbol{\delta}_2(t,s,h)\|$, we obtain from the generalized Gronwall inequality that

$$\|\boldsymbol{\eta}(t,s,h)\| \le r(t,s,h) + C\int_0^t (t-\tau)^{-\alpha}r(\tau,s,h)d\tau.\tag{5.13}$$

Assume we can show that

$$r(t,s,h)\le C\kappa(t,s,-\mu)\qquad (0\le t\le\bar t,\ 0<h\le\delta)\tag{5.14}$$

and that for every $\varepsilon>0$ we have

$$r(t,s,h)\to 0\quad\text{uniformly in } s+\varepsilon\le t\le\bar t.\tag{5.15}$$

Then all the claimed properties of $\boldsymbol{\eta}(t,s,h)$ will be a consequence of the result below, where $\mu<1$, $\{r(t,s,h);\ 0\le h\le\delta\}$ is a family of nonnegative functions in $L^1(0,\bar t)$ with $\mu(t,s,h) = 0$ for $t\le s-h$, and $\nu(t,s,h)$ is defined by

$$\nu(t,s,h) = \int_0^t (t-\tau)^{-\mu}r(\tau,s,h)d\tau.\tag{5.16}$$

LEMMA 5.2. *Assume that* $\{r(t,s,h);\ 0<h\le\delta\}$ *satisfies* (5.14). *Then*

$$\nu(t,s,h)\le C'\kappa(t,s,h,\,1-2\mu)\qquad(0\le t\le\bar t,\ 0\le h\le\delta). \tag{5.17}$$

Moreover, if $\{r(t,s,h)\}$ *is uniformly bounded in* $s+\varepsilon\le t\le\bar t$ *for any* $\varepsilon>0$ *and* $r(t,s,h)\to 0$ *as* $h\to 0$ *almost everywhere in* $s\le t\le\bar t$, *then* $\nu(t,s,h)\to 0$ *as* $h\to 0$ *uniformly for* $s+\varepsilon\le t\le\bar t$ *for any* $\varepsilon>0$.

For a proof see [11, Lem. 5.3]. To show (5.15) we note that, because of the uniform boundedness of the $\boldsymbol{\phi}(t,\mathbf{u}_{s,h,v})$ and the local Lipschitz continuity of $g(t,y)$, we have

$$\begin{aligned}\|\boldsymbol{\phi}(t,\mathbf{u}_{s,h,v})-\boldsymbol{\phi}(t,\mathbf{u})\|&\le C\int_0^t(t-\tau)^{-\mu}\|\boldsymbol{\phi}(\tau,\mathbf{u}_{s,h,v})-\boldsymbol{\phi}(\tau,\mathbf{u})\|d\tau\\&\le C'\int_{(0,t)\cap(s-h,s)}(t-\tau)^{-\mu}d\tau\le Ch\kappa(t,s,h,-\mu)\end{aligned} \tag{5.18}$$

[11, §5] and use again the local Lipschitz continuity of $\mathbf{g}$. $\boldsymbol{\delta}_2(t,s,h)$ is estimated in the same way using (3.3) and the last inequality (5.18) in the integral term. To show (5.15) for $\boldsymbol{\delta}_1$, we use the fact that $h^{-1}\boldsymbol{\rho}(t,\boldsymbol{\phi}(t,\mathbf{u}),\boldsymbol{\phi}(t,\mathbf{u}_{s,h,v}))\to 0$ for $t>s$. To deal with $\boldsymbol{\delta}_2$, let $\{t_n\}$ be an enumeration of the rationals in $0\le t\le\bar t$, and let $e_n\subseteq[0,t_n]$ be the set of the (left) Lebesgue points of the function $\tau\to\mathbf{N}(t_n,\tau)u(\tau)$ (which is strongly measurable by Lemma 3.1). Define $e=\cap_{n\ge1}(e_n\cup[t_n,\bar t])$. Then e has full measure in $0\le t\le\bar t$ and if $s\in e$, then $\boldsymbol{\delta}_2(t_n,s,h)\to 0$ for each n. On the other hand, if $t>s$ and $t_n\ge t$,

$$\begin{aligned}\|\boldsymbol{\delta}_2(t_n,s,u)-\boldsymbol{\delta}_2(t,s,u)\|\le{}&\frac{C}{h}\int_{s-h}^{s}\{(t_n-\tau)^{-\mu}-(t-\tau)^{-\mu}\}d\tau\\&+C\{(t_n-s)^{-\mu}-(t-s)^{-\mu}\}\end{aligned}$$

that can be made arbitrarily small taking t_n-t small enough. Thus (5.15) holds for $\boldsymbol{\delta}_2$ as well. This ends the proof of Theorem 5.1.

We compute directional derivatives (for $t=\bar t$) of

$$\begin{aligned}\mathbf{f}_0(\mathbf{u})=y_0(t,\mathbf{u})={}&\int_{(0,t)\times\Omega}f_0(\tau,y(\tau,x,u))dx\,d\tau\\&+\int_{(0,t)\times\partial\Omega}g_0(\tau,y(\tau,x,u),\,u(\tau,x))d\sigma\,d\tau.\end{aligned} \tag{5.19}$$

THEOREM 5.3. *Assume that* $f_0(t,y)$ *and* $g_0(t,y,u)$ *are continuous in all variables and continuously differentiable with respect to* y. *Let* $\mathbf{u}\in\mathbf{U}_{\text{ad}}(0,\bar t;\,U)$, $\mathbf{s}$, $\mathbf{p}$, h, $\mathbf{v}$, *be as in Theorem* 5.1. *Then there exists a set* $e_0\subseteq e$ *of full measure in* $0\le t\le\bar t$ *such that if* $s_j\in e_0$, *then*

$$\xi_0(\bar t,\mathbf{s},\mathbf{p},\mathbf{u},\mathbf{v})=\lim_{h\to0+}\frac{y_0(\bar t,\mathbf{u}_{\mathbf{s},\mathbf{p},h,\mathbf{v}})-y_0(\bar t,\mathbf{u})}{h} \tag{5.20}$$

exists and equals

$$\xi_0(t,x,\mathbf{s},\mathbf{p},\mathbf{u},\mathbf{v})=\sum_{j=1}^{m}p_j\xi_0(t,x,s_j,\mathbf{u},v_j), \tag{5.21}$$

where

$$\xi_0(t,x,s,\mathbf{u},v) = \int_{(s,t)\times\Omega} \partial_y f_0(\tau, y(\tau,x,u))\xi(\tau,x,s,u,v)dx\,d\tau$$

$$+ \int_{(s,t)\times\partial\Omega} \partial_y g_0(\tau, y(\tau,x,u))\xi(\tau,x,s,u,v)d\sigma\,d\tau \tag{5.22}$$

$$+ \int_{\partial\Omega} g_0(s, y(s,x,u),\, v(x))d\sigma - \int_{\partial\Omega} g_0(s, y(s,x,u), u(s,x))d\sigma.$$

The proof is based on (5.3), (5.4), and (5.8). We omit the details.

6. The maximum principle. We apply the abstract nonlinear programming theory in [11, §2] to the optimal control problem written in the form

$$\text{minimize } \mathbf{f}_0(\mathbf{u}) \quad \text{subject to } \mathbf{f}(\mathbf{u}) \in Y, \tag{6.1}$$

where $\mathbf{f}$ (respectively, $\mathbf{f}_0$) is defined by (5.1) (respectively, (5.19)) in the space $\mathbf{U}_{\text{ad}}(0,\bar{t};\, U)$, equipped with the distance (4.6) that makes the space complete. The function $\mathbf{f}$ is (Hölder) continuous (Corollary 4.6).

LEMMA 6.1. *Let* $\bar{\mathbf{u}}$, ρ, μ *be as in Corollary* 4.6. *Then*

$$|y_0(t,\mathbf{v}) - y_0(t,\mathbf{u})| \le Cd(\mathbf{v},\mathbf{u})^{1-\mu} \qquad (\mathbf{u},\mathbf{v} \in B(\bar{\mathbf{u}},\rho)). \tag{6.2}$$

The estimation of the first integral in (5.19) is immediate from (4.7) and y-differentiability of f_0. In the second, we rewrite the integrand in the form $g_0(t, y(t,x,v), v(t,x)) - g_0(t, y(t,x,v), u(t,x)) + g_0(t, y(t,x,v), u(t,x)) - g_0(t, y(t,x,u), u(t,x))$. The second integral is estimated again using (4.7) and differentiability of g_0; the integrand in the first vanishes outside of $\{t;\, \mathbf{u}(t) \ne \mathbf{v}(t)\}$, so that the integral is $\le Cd(\mathbf{v},\mathbf{u})$.

Having checked (with excess) all hypotheses to apply Theorem 2.8 in [11], let $\bar{\mathbf{u}}$ be a solution of (6.1) (that is, an optimal control). Theorem 2.8 provides a sequence $\{\mathbf{u}^n\} \in \mathbf{U}_{\text{ad}}(0,\bar{t};\, U)$ with $\mathbf{u}^n \to \bar{\mathbf{u}}$ ($d(\mathbf{u}^n,\bar{\mathbf{u}}) \to 0$ as fast as we wish) such that, if $\{\mathbf{D}_n\}$ is a sequence of convex sets with $\mathbf{D}_n \subseteq \partial(\mathbf{f}_0,\mathbf{f})(\mathbf{u}^n)$, there exists a sequence $\{(\mu_n,\nu_n)\} \subseteq \mathbb{R} \times C(\bar{\Omega})^* = \mathbb{R} \times \Sigma(\bar{\Omega})$ such that

$$\mu_n^2 + \|\nu_n\|^2 = 1, \quad \mu_n \ge 0, \quad \mu_n\eta^n + \langle \nu_n, \xi^n\rangle \ge -\delta_n \tag{6.3}$$

for $(\eta^n,\xi^n) \in \mathbf{D}_n$, where $\delta_n \to 0$. If (μ,ν) is the $\mathbb{R} \times C(\bar{\Omega})$-weak limit of (a subsequence of) $\{(\mu_n,\nu_n)\}$, we obtain the Kuhn–Tucker inequality

$$\mu\eta + \langle \nu, \xi\rangle \ge 0 \tag{6.4}$$

for (η,ξ) in the set $\liminf_{n\to\infty} \mathbf{D}_n$ of all limits of sequences $\{(\eta^n,\xi^n)\}$, $(\eta^n,\xi^n) \in \mathbf{D}_n$. Here, $\mathbf{D}_n$ consists of all elements

$$(\xi_0(\bar{t},\mathbf{s},\mathbf{p},\mathbf{u},\mathbf{v}), \boldsymbol{\xi}(\bar{t},\mathbf{s},\mathbf{p},\mathbf{u},\mathbf{v})), \tag{6.5}$$

where $\mathbf{u} = \mathbf{u}^n$, $\boldsymbol{\xi}(\bar{t},\mathbf{s},\mathbf{p},\mathbf{u}^n,\mathbf{v})$ (respectively, $\xi_0(\bar{t},\mathbf{s},\mathbf{p},\mathbf{u}^n,\mathbf{v})$) is given by (5.4) (respectively, (5.21)). So that $\mathbf{D}_n$ is convex, we must allow for the possibility that two or more of the s_j in the vector $\mathbf{s}$ coincide.

LEMMA 6.2. *Let* $\mathbf{u} \in \mathbf{U}_{\text{ad}}(0,\bar{t};\, U)$. *Then there exists a set* e *of full measure in* $0 \le t \le \bar{t}$ *such that* (6.5) *belongs to* $\partial(\mathbf{f}_0,\mathbf{f})(\mathbf{u})$ *for* $\mathbf{s} = (s_1, s_2, \dots, s_m)$, $s \in e$, $0 < s_1 \le s_2 \le \cdots \le s_m < \bar{t}$.

Proof. We limit ourselves to the case $\mathbf{s} = (s, s)$ with $0 < s < \bar{t}$, which puts in evidence the general argument. We only have to construct two sequences $\{s_n^-\}$, $\{s_n^+\} \subset e$, e the set in Theorem 5.1 such that if $\mathbf{s}_n = (s_n^-, s_n^+)$, then $\boldsymbol{\xi}(\bar{t}, \mathbf{s}_n, \mathbf{p}, \mathbf{u}, \mathbf{v}) \to \boldsymbol{\xi}(\bar{t}, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v})$, $\boldsymbol{\xi}_0(\bar{t}, \mathbf{s}_n, \mathbf{p}, \mathbf{u}, \mathbf{v}) \to \boldsymbol{\xi}_0(\bar{t}, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v})$.

In view of the integral equation (3.15), the boundary projection $\boldsymbol{\psi}(t, s, \mathbf{u}, v) = \Pi \circ \boldsymbol{\xi}(t, s, \mathbf{u}, v)$ is

$$\boldsymbol{\psi}(t, s, \mathbf{u}, v) = \mathbf{U}(t, s, \mathbf{u})(v - \mathbf{u}(s)), \tag{6.6}$$

where the operator $\mathbf{U}(t, s, \mathbf{u})$ is given by the integral equation

$$\mathbf{U}(t, s, \mathbf{u}) = \mathbf{M}(t, s) + \int_0^t \mathbf{M}(t, \tau)\partial_y \mathbf{g}(\tau, \mathbf{y}(\tau, \mathbf{u}))\mathbf{U}(\tau, s, \mathbf{u})d\tau. \tag{6.7}$$

Solving by successive approximations and using the properties of $\mathbf{M}(t, s)$, we show that $\mathbf{U}(t, \tau, \mathbf{u}) \in L(L^\infty(\partial\Omega), C(\bar{\Omega}))$ and that $\mathbf{U}(t, \tau, \mathbf{u})$ is continuous in $t, \tau, \mathbf{u}$ for $0 \le \tau < t \le \bar{t}$, $\mathbf{u} \in \mathbf{U}_{\text{ad}}(0, \bar{t}; U)$ ($\mathbf{U}_{\text{ad}}(0, \bar{t}; U)$ endowed with the distance (4.6)) in the norm of $L(L^\infty(\partial\Omega), C(\bar{\Omega}))$. Moreover,

$$\|\mathbf{U}(t, \tau, \mathbf{u})\|_{L(L^\infty(\partial\Omega), C(\bar{\Omega}))} \le B(t - \tau)^{-\mu} \qquad (0 \le \tau < t \le \bar{t}). \tag{6.8}$$

Using the argument in Lemma 3.1, we show that $s \to \boldsymbol{\psi}(\bar{t}, s, \mathbf{u}, v) = \mathbf{U}(\bar{t}, s, \mathbf{u})(v - \mathbf{u}(s))$ is strongly measurable in $0 \le t \le \bar{t}$. Hence if s is a Lebesgue point of $\boldsymbol{\psi}(\bar{t}, \cdot, \mathbf{u}, v)$ and

$$e(s, h) = \{\sigma;\ s - h \le \sigma \le h;\ \|\boldsymbol{\psi}(\bar{t}, \sigma, \mathbf{u}, v) - \boldsymbol{\psi}(\bar{t}, s, \mathbf{u}, v)\| \ge h\},$$

then $\text{meas}\, e(s, h) = o(h)$. It follows that there exists a sequence $\{s_n^-\}$ in the set e of Theorem 5.1, $s_n < s$ such that $s_n^- \to s$, $\boldsymbol{\psi}(\bar{t}, s_n^-, \bar{\mathbf{u}}, v) \to \boldsymbol{\psi}(\bar{t}, s, \bar{\mathbf{u}}, v)$. A similar argument produces a corresponding sequence $\{s_n^+\}$ of similar properties with $s_n^+ > s$ and thus completes the proof. The treatment of $\xi_0(\bar{t}, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v})$ is similar. □

LEMMA 6.3. *Let $\{t_n\}$ be a sequence in $0 \le t \le \bar{t}$ with $t_n \to \bar{t}$ and $\{\mathbf{u}^n\}$ be a sequence in $\mathbf{U}_{\text{ad}}(0, t_n, U)$ such that $\sum d_n(\mathbf{u}^n, \bar{\mathbf{u}}) < \infty$. Then there exists a set e of full measure in $0 \le s \le \bar{t}$ such that if $s_j \in e$ and n is large enough, then the directional derivatives $\boldsymbol{\xi}(\bar{t}, \mathbf{s}, \mathbf{p}, \mathbf{u}^n, \mathbf{v})$, $\xi_0(\bar{t}, \mathbf{s}, \mathbf{p}, \mathbf{u}^n, \mathbf{v})$ exist and converge to* (6.5).

Proof. Let $e_n = \{t;\ 0 \le t \le t_n, \bar{\mathbf{u}}(t) \ne \mathbf{u}^n(t)\}$. Then $\sum_n \text{meas}(c_n) < \infty$, and if $e = \cap_{m \ge 1} \cup_{n \ge m} c_n$, $\text{meas}(e) = 0$; if $t \notin e$, $\mathbf{u}^n(t) = \bar{\mathbf{u}}(t)$ for sufficiently large n. We use formula (6.6) and $\mathbf{u}$-continuity of the operator $U(t, \tau, \mathbf{u})$. The argument for $\xi_0(\bar{t}, \mathbf{s}, \mathbf{p}, \mathbf{u}, \mathbf{v})$ is similar. □

Lemma 6.3 shows that there exists a set e of full measure in $0 \le t \le \bar{t}$ such that (6.5) belongs to $\liminf_{n \to \infty} \mathbf{D}_n$ if $s_j \in e$.

THEOREM 6.4. *Let $\bar{u}(t, x) \in L^\infty((0, T) \times \partial\Omega) = L_W^\infty(0, T; L^\infty(\partial\Omega))$ be an optimal control. Then there exists a set e of full measure in $0 \le t \le \bar{t}$ and $(\mu, \nu) \in \mathbb{R} \times \Sigma(\bar{\Omega})$, $\mu \ge 0$ such that, if $z(t, x)$ is the solution of the final value problem*

$$z_t(t, x) = -\Delta z(t, x) - \mu\partial_y f_0(t, y(t, x, \bar{u})) \qquad ((t, x) \in [0, \bar{t}] \times \Omega), \tag{6.9}$$

$$z(\bar{t}, \cdot) = \nu \qquad (x \in \bar{\Omega}), \tag{6.10}$$

$$\partial_\nu z(t, x) = \partial_y g(t, y(t, x, \bar{u}))z(t, x) - \mu\partial_y g_0(t, y(t, x, \bar{u}) \qquad (t, x) \in [0, \bar{t}] \times \partial\Omega, \tag{6.11}$$

then we have

$$\begin{aligned}&\int_{\partial\Omega}\{z(s,x)\bar{u}(s,x)+\mu g_0(s,y(s,y(s,x,\bar{u}),\bar{u}(s,x)))\}d\sigma\\&\quad=\min_{v\in U}\int_{\partial\Omega}\{z(s,x)v(x)+\mu g_0(s,y(s,x,\bar{u}),v(x))\}d\sigma\end{aligned}\tag{6.12}$$

almost everywhere in $0\le t\le\bar{t}$.

Proof. We use the Kuhn–Tucker inequality (6.4) for the elements (6.5) of $\liminf_{n\to\infty}\mathbf{D}_n$. For single spikes we obtain

$$\mu\xi_0(\bar{t},x,s,\bar{u},v)+\langle\nu,\xi(\bar{t},x,s,\bar{u},v)\rangle\ge 0$$

so that

$$\begin{aligned}&\mu\int_{(s,\bar{t})\times\Omega}\partial_y f_0(\tau,y(\tau,x,\bar{u}))\xi(\tau,x,s,\bar{u},v)dx\,d\tau\\&\quad+\mu\int_{(s,\bar{t})\times\partial\Omega}\partial_y g_0(\tau,y(\tau,x,\bar{u}))\xi(\tau,x,s,\bar{u},v)d\sigma\,d\tau\\&\quad+\mu\int_{\partial\Omega}\{g_0(s,y(t,x,\bar{u}),v(x))-g_0(s,y(t,x,\bar{u}),\bar{u}(s,x))\}d\sigma\\&\quad+\int_{\bar{\Omega}}\xi(\bar{t},x,s,\bar{u},v)\nu(dx)\ge 0.\end{aligned}\tag{6.13}$$

Using Theorem 4.8, (6.12) results. □

For the time optimal problem apply Theorem 2.4 in [11] to the sequence of functions $\mathbf{f}_n(\mathbf{u})=\mathbf{y}(t_n,\mathbf{u}),\mathbf{f}_n:\mathbf{U}_{\text{ad}}(0,t_n;U)\to C(\bar{\Omega})$, where $\{t_n\}$ is a sequence with $t_n<\bar{t}=$ optimal time, $t_n\to\bar{t}$ (see [11] for details). If $\bar{\mathbf{u}}$ is a time optimal control, Theorem 2.4 provides a sequence $\{\mathbf{u}^n\}$, $\mathbf{u}^n\in\mathbf{U}_{\text{ad}}(0,t_n;U)$ (with $d_n(\mathbf{u}^n,\bar{\mathbf{u}})\to 0$ as fast as we wish, d_n the distance (4.6)) such that, if $\{\mathbf{D}_n\}$ is a sequence of convex sets with $\mathbf{D}_n\subseteq\partial\mathbf{f}_n(\mathbf{u}^n)$, there exists a sequence $\{\nu_n\}\subseteq C(\bar{\Omega})^*=\Sigma(\bar{\Omega})$ such that

$$\|\nu_n\|^2=1,\qquad\langle\nu_n,\xi^n\rangle\ge-\delta_n\tag{6.14}$$

for every $\xi^n\in\mathbf{D}_n$, where $\delta_n\to 0$. Taking limits,

$$\langle\nu,\xi\rangle\ge 0.\tag{6.15}$$

For single-spike perturbations,

$$\int_{\bar{\Omega}}\xi(\bar{t},x,s,\bar{u},v)\nu(dx)\ge 0\tag{6.16}$$

for all s in a total set e and all $v\in e$. Operating as above, we obtain the following theorem.

THEOREM 6.5. *Let $\bar{u}(t,x)$ be an optimal control for the time optimal problem. Then there exists $\nu\in\Sigma(\bar{\Omega})$ such that, if $z(t,x)$ is the solution of*

$$z_t(t,x)=-\Delta z(t,x)\qquad(t,x)\in[0,\bar{t})\times\Omega,\tag{6.17}$$

$$z(\bar{t},\cdot)=\nu\qquad(x\in\bar{\Omega}),\tag{6.18}$$

$$\partial_\nu z(t,x)=\partial_y g(t,y(t,x,u))z(t,x)\qquad(t,x)\in[0,\bar{t})\times\partial\Omega,\tag{6.19}$$

then we have

$$\int_{\partial\Omega} z(s,x)\bar{u}(s,x)d\sigma = \min_{v\in U}\int_{\partial\Omega} z(s,x)v(x)d\sigma \quad \textit{almost everywhere in } 0\le t\le \bar{t}. \tag{6.20}$$

Without special assumptions, the multipliers (μ,ν) in Theorem 6.4 and ν in Theorem 6.5 may be trivial. Conditions for the abstract nonlinear programming problem that prevent this are given in [11, Lem. 2.5]. For the general problem, this result requires a compact set Q such that

$$\Delta = \bigcap_{n\ge 1}(\overline{\text{conv}}(\Delta_n)+Q) \tag{6.21}$$

contains an interior point, where $\Delta_n = \Pi(\mathbf{D}_n)-N_Y(\mathbf{y}^n)+Q$; $\{\mathbf{D}_n\}$, the sequence in provided by [15, Thm. 2.8]; Π, the canonical projection from $\mathbb{R}\times C(\bar{\Omega})$ into $C(\bar{\Omega})$; $\{\mathbf{y}_n\}$, a sequence in the target set Y with $\mathbf{y}^n\to\bar{\mathbf{y}}=\mathbf{y}(\bar{t},\bar{\mathbf{u}})$; and $N_Y(\bar{\mathbf{y}})$, the tangent cone to Y at $\bar{\mathbf{y}}$. It can be shown that, due to the smoothing properties of the heat equation, $\Pi(\mathbf{D}_n)+Q$ will never satisfy this condition by itself, thus we must rely on a "large" target set Y. A suitable target set is, for instance $Y=C\cap B(x,\varepsilon)$, where C is the variety in $C(\bar{\Omega})$ defined by $\phi_1(y)=\phi_2(y)=\cdots=\phi_p(y)=0$, the ϕ_j continuously differentiable functionals with $\{\partial_j\phi(y)\}$ linearly independent in $C(\bar{\Omega})^*=\Sigma(\bar{\Omega})$. In the time optimal case, $\Delta_n=\mathbf{D}_n-N_Y(\mathbf{y}^n)+Q$ and the same comments apply. (See [11] for a similar situation for distributed parameter systems.)

There are some indications that the point target case for the general control problem as well as for the time optimal problem could be treated in controllability subspaces such as those provided in [25]. This has been done in the linear case (using separation theorems) in [5], [6]. However, the nonlinear case seems to be open.

Acknowledgment. The authors are grateful to the referee for numerous bibliographical indications.

REFERENCES

[1] H. Amann, *Parabolic evolution equations with nonlinear boundary conditions*, in Nonlinear Functional Analysis and Applications, F. Browder, ed., Proc. Symp. Pure Appl. Math., Vol. 45 (part I), 1986, pp. 17–27.

[2] V. Barbu, *Boundary control problems with nonlinear state equations*, SIAM J. Control Optim., 20 (1982), pp. 46–65.

[3] V. Barbu and F. Pavel, *Optimal control problems for boundary control systems*, preprint.

[4] H. O. Fattorini, *The Cauchy Problem*, Cambridge University Press, Cambridge, UK, 1983.

[5] H. O. Fattorini, *The time optimal control problem in Banach spaces*, Appl. Math. Optim., 1 (1974/75), pp. 163–188.

[6] H. O. Fattorini, *The time-optimal problem for boundary control of the heat equation*, in Calculus of Variations and Control Theory, Academic Press, New York, 1976, pp. 305–320.

[7] H. O. Fattorini, *The maximum principle for nonlinear nonconvex systems in infinite dimensional spaces*, in Distributed Parameter Systems, Springer Lecture Notes in Control and Information Sciences, Vol. 75, 1985, pp. 162–178.

[8] H. O. Fattorini, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Applied Math. Optim., 15 (1987), pp. 141–185.

[9] H. O. Fattorini, *Optimal control of nonlinear systems: Convergence of suboptimal controls*, I, in Lecture Notes in Pure and Applied Mathematics, Vol. 108, Marcel Dekker, New York, 1987, pp. 159–199.

[10] H. O. Fattorini, *Optimal control problems for distributed parameter systems governed by semilinear parabolic equations in L^1 and L^∞ spaces*, in Optimal Control of Partial Differential Equations, Springer Lecture Notes in Control and Information Sciences, Vol. 149, 1991, pp. 68–80.

[11] H. O. Fattorini, *Optimal control problems in Banach spaces*, Appl. Math. Optim., 28 (1993), pp. 225–257.

[12] H. O. Fattorini, *Relaxed controls in infinite-dimensional control systems*, Internat. Ser. Numer. Math., 100 (1991), pp. 115–128.

[13] H. O. FATTORINI, *Relaxed controls, differential inclusions, existence theorems and the maximum principle in nonlinear infinite-dimensional control theory*, in Differential Equations, Optimal Control and Biomathematics P. Clément and G. Lumer, eds. Marcel Dekker, New York, NY 1993, pp. 177–195.

[14] H. O. FATTORINI AND H. FRANKOWSKA, *Necessary conditions for infinite dimensional control problems*, in Springer Lecture Notes in Control and Information Sciences, Vol. 111, 1990, pp. 381–392.

[15] H. O. FATTORINI AND H. FRANKOWSKA, *Necessary conditions for infinite dimensional control problems*, Math. Control Signals Systems, 4 (1990), pp. 41–67.

[16] H. FRANKOWSKA, *Some inverse mapping theorems*, Ann. Inst. Henri Poincaré, 7 (1990), pp. 183–234.

[17] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1964.

[18] K. GLASHOFF AND N. WECK, *Boundary control of parabolic differential equations in arbitrary dimension: Supremum-norm problems*, SIAM J. Control Optim., 14 (1976), pp. 662–681.

[19] H. GOLDBERG AND F. TRÖLTZSCH, *Second order optimality conditions for a class of control problems governed by nonlinear integral equations with an application to parablic boundary control*, Optimization, 20 (1989), pp. 687–698.

[20] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Springer, Berlin, 1981.

[21] O. LADIZHENSKAYA, V. SOLONNIKOV, AND N. URALCEVA, *Linear and quasilinear equations of parabolic type*, Izd. Nauka, Moscow, 1967.

[22] U. MACKENROTH, *Time-optimal parabolic boundary control problems with pointwise state constraints*, Num. Funct. Anal. Optim., 3 (1981), pp. 285–300.

[23] U. MACKENROTH, *On parabolic distributed optimal control problems with restrictions on the gradient*, Appl. Math. Optim., 10 (1983), pp. 69–95.

[24] T. MURPHY, *Optimal control problems for nonlinear boundary control systems*, Ph.D. thesis, University of California, Department of Mathematics, Los Angeles, CA, 1992.

[25] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.

[26] E. SACHS, *A parabolic control problem of the Stefan–Boltzmann type*, Z. Angew Math. Mech, 58 (1978), pp. 443–449.

[27] E. J. P. G. SCHMIDT, *Boundary control for the heat equation with nonlinear boundary condition*, J. Differential Equations, 78 (1989), pp. 98–121.

[28] F. TRÖLTZSCH, *On the semigroup approach for optimal control of semilinear parabolic equations including distributed and boundary control*, Zeitschrift für Anal. und Ihre Awend., 8 (1989), pp. 431–443.

[29] F. TRÖLTZSCH, *Optimality Conditions for Parablic Control Problems and Applications*, Teubner-Texte zur Mathematik, Band 62, B. G. Teubner Verlagsgesselschaft, Leipzig, 1984.

[30] L. v. WOLFERSDORF, *Optimal control for processes governed by mildly nonlinear differential equations of parabolic type* I, II, Z. Angew Math. Mech, 56 (1976), pp. 531–538; 57 (1977), pp. 11–17.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1597–1604, November 1994

006

ON BROCKETT'S CONDITION FOR SMOOTH STABILIZABILITY AND ITS NECESSITY IN A CONTEXT OF NONSMOOTH FEEDBACK*

E. P. RYAN†

Abstract. The necessity of Brockett's condition for stabilizability of nonlinear systems by smooth feedback is shown, by an argument based on properties of a degree for set-valued maps, to persist when the class of controls is enlarged to include discontinuous feedback.

Key words. degree, discontinuous control, nonlinear systems, nonsmooth feedback, set-valued maps, stabilizability

AMS subject classifications. 34D99, 55M25, 93B55, 93D15

1. Introduction.

Consider the control system

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = x^0 \in \mathbb{R}^N, \quad f(0,0) = 0, \tag{1}$$

with $f : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$ continuous. In the case of linear f, it is well known that the system is (globally) asymptotically null controllable if and only if it is stabilizable by (linear) feedback. Brockett [2] has shown that an analogous equivalence of (local) asymptotic null controllability and (nonlinear) feedback stabilizability does not hold for smooth (by which we mean C^1) nonlinear systems. In particular, he proved a result that implies the following necessary condition for (local) smooth stabilizability—henceforth referred to as Brockett's condition.

BROCKETT'S CONDITION. *Let $f \in C^1$. If (1) is C^1 stabilizable (in the sense that there exists a time-invariant C^1 feedback that renders $\{0\}$ both Lyapunov stable and an attractor), then the image of f contains an open neighbourhood of 0.*

If f is linear, that is, if $f(x,u) = Ax + Bu$, then the necessary condition for stabilizability is simply the requirement that $[A : B]$ be of full rank, and this is implied by asymptotic null controllability of the linear system. However, for general $f \in C^1$, (local asymptotic) null controllability of (1) does not imply that f has the above property, a (now classic) illustration is the case

$$f : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^3, \qquad (x,u) = (x_1, x_2, x_3, u_1, u_2) \mapsto (u_1, u_2, x_2u_1 - x_1u_2)$$

that defines a completely controllable bilinear system (1) for which $(0,0,\epsilon) \notin \mathrm{im}(f)$ for all $\epsilon \neq 0$, and so this system is not C^1 stabilizable.

Such examples are counterintuitive. It is tempting to conjecture that the "gap" between controllability and feedback stabilizability is due to the restriction to the class of smooth (C^1) time-invariant feedbacks. As in Sontag [11], the investigation readily extends to time-invariant feedbacks that are only locally Lipschitz (in fact, even this requirement is too strong, its consequence, *uniqueness of the solution* of the feedback-controlled initial-value problem, suffices as in [13]) and the gap is found to persist. Furthermore, Zabczyk [13] has shown that the necessity of Brockett's condition on f also persists when "stabilizability by time-invariant continuous feedback is interpreted in either of the following senses: (i) that of rendering $\{0\}$ a global attractor (which, of course, does not imply Lyapunov stability of $\{0\}$), or (ii) in the case of $n \leq 2$, that of rendering $\{0\}$ Lyapunov stable. Two possible avenues for further investigation suggest themselves naturally: (a) time-varying feedback and (b) discontinuous feedback. The former avenue has been followed by Coron [5]. In the case of $f(x,u) = \sum_{i=1}^{M} u_i f_i(x)$ with

* Received by the editors August 10, 1992; accepted for publication (in revised form) July 5, 1993.
† School of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom.

$f_i \in C^\infty(\mathbb{R}^N)$, he has established that the accessibility rank condition, $\dim \mathrm{Lie}(\Phi)(x) = N$ for all $x \in \mathbb{R}^N \backslash \{0\}$ (where $\mathrm{Lie}(\Phi)$ denotes the Lie algebra of vector fields generated by $\Phi = (f_1, \ldots, f_M)$), is sufficient for the existence of T-periodic C^∞ feedbacks that globally asymptotically stabilize (1). In particular, this result applies to the example cited above. In the present paper, we take the second avenue and restrict to time-invariant feedbacks.

Discontinuous feedbacks arise naturally in many areas of control theory (see [7]) and practice (indeed, bang-bang or relay-type control actions permeate much of the early development of the field). It is not difficult to construct examples that fail to be locally asymptotically stabilizable by continuous feedback, but that are so stabilizable by discontinuous feedback. One such example is system (1) with $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $(x, u) \mapsto x + |x|u$. Therefore, the additional dynamic behaviours engendered by discontinuous feedbacks (that subsume the continuous case) raise the question of whether or not their adoption might close the controllability-stabilizability gap. Here this question is answered negatively. We show that with f only required to be continuous and to have property (2), below, and with the class of time-invariant feedbacks taken to be that of upper semicontinuous set-valued maps with nonempty convex and compact values (a class into which a wide variety of discontinuous strategies may be embedded and within which continuous feedbacks may be identified with the subclass of singleton-valued maps), the necessity of Brockett's condition on f again persists.

2. Class of systems and statement of main result. We study systems of form (1) and assume only that continuous f has the property (see also Remark 1, below)

$$K \subset \mathbb{R}^M \text{ convex} \Longrightarrow f(x, K) \subset \mathbb{R}^N \text{ convex}. \tag{2}$$

Evidently, (2) holds for systems that are linear in the control.

As *admissible feedback controls* for (1), we take the class $\mathcal{K}$ of upper semicontinuous maps $x \mapsto k(x) \subset \mathbb{R}^M$ on $\mathbb{R}^N$, with nonempty convex and compact values and with $0 \in k(0)$. For example, in the case $M = 1$, discontinuous feedbacks of the form $x \mapsto \gamma(x)\mathrm{sgn}(\xi(x))$, with γ and ξ continuous and such that the product $\gamma(0)\xi(0)$ is zero, fall within our framework if the signum function is interpreted as the upper semicontinuous set-valued map

$$v \mapsto \mathrm{sgn}(v) := \begin{cases} \{+1\}, & v > 0, \\ [-1, 1], & v = 0, \\ \{-1\}, & v < 0. \end{cases}$$

For every feedback $k \in \mathcal{K}$, the map $x \mapsto f(x, k(x))$ is also upper semicontinuous with nonempty convex and compact values. Therefore, for each $x^0 \in \mathbb{R}^N$, the initial-value problem

$$\dot{x}(t) \in f(x(t), k(x(t))), \qquad x(0) = x^0 \tag{3}$$

has at least one solution (see [1, Thm. 2.1.3]), that is, a function $x : [0, \omega) \to \mathbb{R}^N$, with $x(0) = x^0$, that is absolutely continuous on compact subintervals and that satisfies the differential inclusion in (3) almost everywhere. Moreover, every solution x can be maximally extended. Furthermore, if x is bounded on its maximal interval of existence $[0, \omega)$, then $\omega = \infty$ (see, for example, [10]). We say that $\{z\}$ is an equilibrium of (3) if $0 \in f(z, k(z))$. Note that, for each $k \in \mathcal{K}$, $\{0\}$ is an equilibrium of (3).

In contrast with the smooth case, the property of uniqueness of the solution for the initial-value problem (3) clearly does not hold in our general nonsmooth framework. Implicit in the following definition is a notion of local asymptotic stability wherein we impose "equi-attractivity" of the equilibrium $\{0\}$. In essence, attraction to this equilibrium is required to be uniform with respect to nonunique solutions.

DEFINITION 1. *A feedback control $k \in \mathcal{K}$ is said to be* equi-asymptotically stabilizing *for* (1) *if it renders the equilibrium* $\{0\}$ *of* (3) *equi-asymptotically stable in the sense that the following two properties hold.*

(i) Lyapunov stability of the equilibrium $\{0\}$: *for each* $\rho > 0$, *there exists* $\delta > 0$ *such that*

$$\|x^0\| \leq \delta \Longrightarrow \|x(t)\| < \rho \quad \text{for all } t \geq 0$$

for every maximal solution x *of the initial-value problem* (3).

(ii) Equi-attractivity of the equilibrium $\{0\}$: *there exists* $\delta > 0$ *and, to each* $\tau > 0$, *there corresponds* $T > 0$ *such that*

$$\|x^0\| \leq \delta \Longrightarrow \|x(t)\| < \tau \quad \text{for all } t \geq T$$

for every maximal solution x *of the initial-value problem* (3).

Although the above definition is intrinsic to the problem, the following weaker (but somewhat artificial) property of the feedback is all that is required in the analysis.

DEFINITION 2. *A feedback control $k \in \mathcal{K}$ is said to be* equi-constricting *for* (1) *if* (3) *has the following property. There exist scalars* $\rho > \delta > \tau > 0$ *and* $T > 0$ *such that*

$$\|x^0\| \leq \delta \Longrightarrow \|x(t)\| < \rho \quad \text{for all } t \geq 0 \quad \text{and} \quad \|x(t)\| < \tau \quad \text{for all } t \in [T, 2T]$$

for every maximal solution x *of* (3).

It is clear that, if $k \in \mathcal{K}$ is an equi-asymptotically stabilizing feedback for (1), then k is an equi-constricting feedback for (1). While the former concept is manifestly more natural from an applications viewpoint, the latter is considerably weaker. In particular, Definition 2 simply invokes the existence of *some* quadruple (ρ, δ, τ, T), assuring the requisite properties. In essence, solutions of (3) are required only to be bounded uniformly with respect to initial data in some closed ball (of radius δ) and, on an interval $[T, 2T]$, to take their values in some smaller ball (of radius $\tau < \delta$).

The main result we will prove is the following.

THEOREM 1. *Let f be continuous with property* (2). *If there exists an equi-constricting feedback control $k \in \mathcal{K}$ for* (1), *then the image of f contains an open neighbourhood of* 0.

A simple modification to the proof of Theorem 1 will yield the following generalization of Brockett's condition.

COROLLARY 1. *Let f be continuous with property* (2). *If there exists an equi-asymptotically stabilizing feedback control $k \in \mathcal{K}$ for* (1), *then, for each open neighbourhood $\mathcal{N}$ of $0 \in \mathbb{R}^N$, $f(\mathcal{N} \times \mathbb{R}^M)$ contains an open neighbourhood of* 0.

Remark 1. If $\mathcal{K}$ is replaced by the class of C^1 feedbacks and attention is restricted to functions $f \in C^1$, then condition (2), which plays its role only in assuring that the right hand side of (3) takes convex values, may be removed; furthermore, the qualifier "equi" in Definition 2 is redundant. In this manner, Brockett's original result for smooth systems may be recovered as a special case of the above. It is in this sense that we regard Corollary 1 as a generalization of Brockett's condition.

The proof of Theorem 1, which is given in §4, is degree-theoretic in nature and similar in concept to the approaches of [8, §52], [11, §4.8], and [13, §2]. However, in the present nonsmooth setting, we first require some appropriate notion of degree for set-valued maps. This has been investigated by Cellina and Lasota [3] (see also [9], [12], [6]), and a distillation of results pertinent to our application is given in the next section.

3. Degree for set-valued maps. Here the objective is to reiterate, within the framework of [3], [12] but tailored to our immediate purpose, some results pertaining to degree for set-valued maps. The approach to defining degree for a (suitably regular) set-valued map F is via the Brouwer degree for single-valued approximate selections for F. With this in mind, some basic definitions and properties of upper semicontinuous maps and approximate selections (for details, see [1], [6]) are initially assembled.

3.1. Upper semicontinuous maps and approximate selections. For notational convenience, write $X := \mathbb{R}^N$. The ball of radius $r > 0$, centred at $c \in X$, will be denoted $B_r(c)$; when $c = 0$, we simply write B_r. For nonempty subsets U, V of a Banach space Y, define

$$d(y, V) := \inf_{v \in V} \|y - v\| \quad \text{for all } y \in Y, \quad \text{and} \quad d^*(U, V) := \sup_{u \in U} d(u, V).$$

Let $x \mapsto F(x) \subset X$, with domain $\mathrm{dom}(F) = D \subset X$, have nonempty values. F is upper semicontinuous if it is upper semicontinuous at each $x \in D$: for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$F(w) \subset F(x) + B_\epsilon \quad \text{for all } w \in D \cap B_\delta(x).$$

If $C \subset D$ is compact and F is upper semicontinuous with compact values, then $F(C)$ is compact.

THEOREM 2 (Approximate selection theorem). *Let F be an upper semicontinuous map with domain $D \subset X$ and taking nonempty convex compact values in X. For each $\epsilon > 0$, there exists a locally Lipschitz single-valued function $f_\epsilon : D \to \mathrm{co}(F(D))$ such that*

$$d^*(\mathrm{graph}(f_\epsilon), \mathrm{graph}(F)) < \epsilon.$$

(Any such f_ϵ will be referred to as an approximate selection for F.)

3.2. Construction and properties of degree. Initially, we recall *Brouwer degree* in the context of single-valued maps. As before, let $X := \mathbb{R}^N$. Henceforth, $\Omega \subset X$ is a bounded open set, with closure $\bar{\Omega}$ and boundary $\partial\Omega$. Let

$$\mathcal{M} = \{(f, \Omega, p) | X \supset \Omega \text{ open bounded}, f : \bar{\Omega} \to X \text{ continuous}, p \in X \backslash f(\partial\Omega)\},$$

then the Brouwer degree $\deg_B$ is the unique map $\mathcal{M} \to \mathbb{Z}$ with the following properties:

B-1. $\deg_B(I, \Omega, p) = 1$ for all $p \in \Omega$;

B-2. If $\deg_B(f, \Omega, p) \neq 0$, then $p = f(x)$ for some $x \in \Omega$;

B-3. (Homotopic invariance). If $h : [0, 1] \times \bar{\Omega} \to X$ and $q : [0, 1] \to X$ are continuous with $q(t) \notin h(t, \cdot)(\partial\Omega)$ for all $t \in [0, 1]$, then $\deg_B(h(t, \cdot), \Omega, q(t))$ is independent of $t \in [0, 1]$;

B-4. (Odd mappings). If Ω contains, and is symmetric about, the origin in X and $f(-x) = -f(x)$ for all $x \in \partial\Omega$, then $\deg_B(f, \Omega, 0)$ is odd (and so is nonzero).

The class of set-valued maps F, to which the ensuing construction [3], [12] of degree applies, are precisely those satisfying the hypotheses of Theorem 2: upper semicontinuous maps $x \mapsto F(x) \subset X$ from $\mathrm{dom}(F) \subset X$ to the nonempty convex compact subsets of X.

For every open bounded Ω, with closure $\bar{\Omega} \subset \mathrm{dom}(F) \subset X$, and every $p \in X \backslash F(\partial\Omega)$, we define an integer $\deg(F, \Omega, p)$, the degree of F (with respect to the set Ω and point p).

3.2.1. Construction. Let $p \in X \backslash F(\partial\Omega)$ and let $F^{(p)}$ denote the map defined on compact $\bar{\Omega}$ by $x \mapsto F(x) - \{p\} = \{v - p | v \in F(x)\}$. By Theorem 2, for each $\epsilon > 0$ there exists an approximate selection f_ϵ for $F^{(p)}$. We first show that, for all $\epsilon > 0$ sufficiently small, every such approximate selection f_ϵ has no zeros in $\partial\Omega$. Suppose otherwise. Then there

exist sequences (ϵ_n), (f_{ϵ_n}), and $(x_n) \subset \partial\Omega$, with $\epsilon_n \downarrow 0$, $0 = f_{\epsilon_n}(x_n) \in \text{co}(F^{(p)}(\bar{\Omega}))$, and $0 \in F^{(p)}(y_n) + B_{\epsilon_n}$ for some $y_n \in \bar{\Omega}$ with $\|x_n - y_n\| < \epsilon_n$. By compactness of $\bar{\Omega}$, (y_n) has a convergent subsequence (that we do not relabel), with limit z say, and so $x_n \to z \in \partial\Omega$ as $n \to \infty$. By upper semicontinuity of $F^{(p)}$, for each $\epsilon > 0$, $0 \in F^{(p)}(z) + B_{\epsilon_n+\epsilon}$ for all n sufficiently large. Therefore, $0 \in \overline{F^{(p)}(z)} = F^{(p)}(z)$ and so $p \in F(z)$ with $z \in \partial\Omega$, a contradiction. It follows that for all $\epsilon > 0$ sufficiently small, $\deg_B(f_\epsilon, \Omega, 0)$ is well defined for every approximate selection f_ϵ for $F^{(p)}$.

Let f_ϵ and g_ϵ be any two such approximate selections. Define the continuous function

$$h_\epsilon : [0,1] \times \bar{\Omega} \to X, \qquad (t,x) \mapsto tf_\epsilon(x) + (1-t)g_\epsilon(x).$$

For all $\epsilon > 0$ sufficiently small, $h_\epsilon(t,\cdot)$ has no zeros in $\partial\Omega$ for every $t \in [0,1]$. This can be argued (in a similar manner to above) by contradiction. Suppose otherwise; then there exist $t \in [0,1]$, a sequence (ϵ_n) with $\epsilon_n \downarrow 0$, and a sequence $(x_n) \subset \partial\Omega$ such that

$$0 = h_{\epsilon_n}(t,x_n) = tf_{\epsilon_n}(x_n) + (1-t)g_{\epsilon_n}(x_n) \in tF^{(p)}(y_n) + (1-t)F^{(p)}(z_n) + B_{\epsilon_n}$$

for some $y_n, z_n \in \bar{\Omega}$ with $\|x_n - y_n\|$, $\|x_n - z_n\| < \epsilon_n$. By compactness of $\bar{\Omega}$, without loss of generality we may assume that $z_n \to z$ and so $y_n \to z$ and $x_n \to z \in \partial\Omega$ as $n \to \infty$. By upper semicontinuity of $F^{(p)}$ and convexity of its values, for each $\epsilon > 0$, $0 \in F^{(p)}(z) + B_{\epsilon_n+\epsilon}$ for all n sufficiently large and so $0 \in F^{(p)}(z)$, contradicting the fact that $p \notin F(\partial\Omega)$. Therefore, for all $\epsilon > 0$, $0 \notin h_\epsilon(t,\cdot)(\partial\Omega)$ for all $t \in [0,1]$. Thus, by property B-3, $\deg_B(h_\epsilon(t,\cdot),\Omega,0)$ is independent of $t \in [0,1]$, and so we may conclude that, for all $\epsilon > 0$ sufficiently small,

$$\deg_B(f_\epsilon,\Omega,0) = \deg_B(h(1,\cdot),\Omega,0) = \deg_B(h(0,\cdot),\Omega,0) = \deg_B(g_\epsilon,\Omega,0).$$

Simply stated, for all $\epsilon > 0$ sufficiently small, $\deg_B(f_\epsilon,\Omega,0)$ is well defined for every approximate selection f_ϵ and is independent of the particular selection chosen.

In summary, the above construction ensures that the following concept of degree for the set-valued map F is well defined:

$$\deg(F,\Omega,p) := \lim_{\epsilon \downarrow 0} \deg_B(f_\epsilon,\Omega,0).$$

3.2.2. Properties.

THEOREM 3. *Let $x \mapsto F(x) \subset X$ be upper semicontinuous on compact $\bar{\Omega} \subset X$ with nonempty, convex, and compact values.*

(i) *If $q : [0,1] \to X \backslash F(\partial\Omega)$ is continuous, then $\deg(F,\Omega,q(t))$ is independent of $t \in [0,1]$.*

(ii) *If $p \in X \backslash F(\partial\Omega)$ is such that $\deg(F,\Omega,p) \neq 0$, then $p \in F(x)$ for some $x \in \Omega$.*

Proof. By the above construction, all degrees in the assertions of the theorem are well defined.

Assertion (i) is an immediate consequence of the construction together with B-3.

(ii) Because $\deg_B(F,\Omega,p) \neq 0$, there exists a sequence (ϵ_n), with $\epsilon_n \downarrow 0$, and an associated sequence (f_{ϵ_n}) of approximate selections for $F^{(p)}$ such that $\deg_B(f_{\epsilon_n},\Omega,0) \neq 0$ for all n sufficiently large. By B-2, for each n sufficiently large, there exists $x_n \in \Omega$ such that $0 = f_{\epsilon_n}(x_n)$. By compactness of $\bar{\Omega}$, without loss of generality we may assume that $x_n \to x \in \bar{\Omega}$. Moreover, because the functions f_{ϵ_n} are approximate selections, for each n there exists $y_n \in \bar{\Omega}$, with $\|x_n - y_n\| < \epsilon_n$, such that

$$0 = f_{\epsilon_n}(x_n) \in F^{(p)}(y_n) + B_{\epsilon_n}.$$

Arguing as before (using semicontinuity of F and compactness of its values), it follows that $0 \in F^{(p)}(x)$ and so $p \in F(x)$. This proves assertion (ii) of the theorem. □

4. Proof of the main result. We now turn attention to the proof of Theorem 1. Again write $X := \mathbb{R}^N$. Assume $k \in \mathcal{K}$ is an equi-constricting feedback for (1). Then there exist $\rho > \delta > \tau > 0$ and $T > 0$ such that

$$\|x^0\| \leq \delta \Longrightarrow \|x(t)\| < \rho \quad \text{for all } t \geq 0 \quad \text{and} \quad \|x(t)\| < \tau \quad \text{for all } t \in [T, 2T]$$

for every maximal solution $x(\cdot)$ of (3).

Define the set-valued map F on X as follows:

$$F : x \mapsto \begin{cases} f(x, k(x)), & \|x\| \leq \rho, \\ f(\rho\|x\|^{-1}x, k(\rho\|x\|^{-1}x)), & \|x\| > \rho. \end{cases} \tag{4}$$

It is evident that F is upper semicontinuous with nonempty convex and compact values, and so $F(\bar{B}_\rho) \equiv F(X)$ is compact. By the construction in §3.2.1, for every open bounded set $\Omega \subset X$ and every $p \in X \backslash F(\partial\Omega)$, $\deg(F, \Omega, p)$ is well defined.

Consider the initial-value problem

$$\dot{x}(t) \in F(x(t)), \qquad x(0) = x^0. \tag{5}$$

By compactness of $F(X)$ we may deduce that, for each $x^0 \in X$, every solution of (5) has maximal interval of existence $\mathbb{R}^+ := [0, \infty)$. Observe that, for each x^0 with $\|x^0\| \leq \delta$, the set of maximal solutions of (5) is precisely the set of maximal solutions of (3).

Write $\Omega^0 := B_\delta$, with closure $\overline{\Omega^0}$. By the equi-constricting property, the annulus $\overline{\Omega^0} \backslash B_\tau$ cannot contain an equilibrium of (5) (or, equivalently, a point x such that $0 \in F(x)$). Therefore, $0 \notin F(\partial\Omega^0)$ and $\deg(F, \Omega^0, 0)$ is well defined. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of locally Lipschitz approximate selections for F with

$$d^*(\text{graph}(f_n), \text{graph}(F)) \to 0 \quad \text{as } n \to \infty$$

and such that $\deg(F, \Omega^0, 0) = \deg_{\text{B}}(f_n, \Omega^0, 0)$ for all n.

Write $I := [0, 2T]$ and $Y := C(I; X)$ with the uniform norm. On $\overline{\Omega^0}$ we define the map

$$\mathcal{F} : x^0 \mapsto \{x \in Y | \dot{x}(t) \in F(x(t)) \text{ a.e., } x(0) = x^0\}.$$

For each n, define the map $\phi_n : \overline{\Omega^0} \to Y$ as follows: $\phi_n(x^0)$ is the unique element x of Y such that

$$\dot{x}(t) = f_n(x(t)) \quad \text{for all } t \in I, \quad \text{and} \quad x(0) = x^0.$$

By the classical theory of ordinary differential equations, the map $(t, x^0) \mapsto (\phi_n(x^0))(t)$ is continuous.

We claim that, for every $\epsilon > 0$,

$$d^*(\text{graph}(\phi_n), \text{graph}(\mathcal{F})) < \epsilon \quad \text{for some } n.$$

Suppose otherwise. Then there exist $\epsilon > 0$ and a sequence $(x_n^0) \subset \overline{\Omega^0}$ such that

$$d((x_n^0, \phi_n(x_n^0)), \text{graph}(\mathcal{F})) \geq \epsilon \quad \text{for all } n.$$

For notational convenience, we write $x_n = \phi_n(x_n^0)$. Arguing as in the first proof of Theorem 2.1.3 of [1] (see also [4, Thm 3.1.7]) and extracting a subsequence if necessary, we may assume that $(x_n) \subset Y$ converges uniformly to an absolutely continuous function $x : I \to X$, $x(0) =$

$x^0 \in \overline{\Omega^0}$, satisfying $\dot{x}(t) \in F(x(t))$ almost everywhere, whence the following contradiction: $(x_n^0, x_n) \to (x^0, x) \in \text{graph}(\mathcal{F})$ as $n \to \infty$.

Let $0 < \epsilon < \delta - \tau$ and let m be such that $d^*(\text{graph}(\phi_m), \text{graph}(\mathcal{F})) < \epsilon$. We assert that

$$\text{for all } x^0 \in \overline{\Omega^0}, \quad (\phi_m(x^0))(t) \in \Omega^0 \quad \text{for all } t \in [T, 2T]. \tag{6}$$

This may be shown as follows. Let $x^0 \in \overline{\Omega^0}$ be arbitrary. There exists $y^0 \in \overline{\Omega^0}$, with $\|x^0 - y^0\| < \epsilon$, and $y \in \mathcal{F}(y^0)$ such that $\|(\phi_m(x^0))(t) - y(t)\| < \epsilon$ for all $t \in I$. Because the set $\{y(t) | y \in \mathcal{F}(\overline{\Omega^0})\}$ lies in the ball B_τ for all $t \in [T, 2T]$, the assertion must hold.

Define a function $h : [0, 1] \times \overline{\Omega^0} \to X$ by

$$h(s, x^0) := \begin{cases} f_m(x^0), & s = 0, \\ \frac{1}{s}[(\phi_m(x^0))(sT) - x^0], & 0 < s \leq 1. \end{cases}$$

That h is continuous is readily verified. Furthermore, $h(s, x^0) \neq 0$ for all $(s, x^0) \in [0, 1] \times \partial\Omega^0$ by the following argument. Suppose $h(0, x^0) = f_m(x^0) = 0$ for some $x^0 \in \partial\Omega^0$. Then $(\phi_m(x^0))(t) = x^0 \in \partial\Omega^0$ for all $t \in I$, which contradicts (6). Now suppose $h(s, x^0) = 0$ for some $(s, x^0) \in (0, 1] \times \partial\Omega^0$. Then $(\phi_m(x^0))(nsT) = x^0 \in \partial\Omega^0$ for all $n \in \mathbb{N}$ with $ns \leq 2$.

In particular, there exists $n \in \mathbb{N}$ such that

$$1 \leq ns \leq 2 \quad \text{and} \quad (\phi_m(x^0))(nsT) = x^0 \in \partial\Omega^0.$$

This contradicts (6).

We have now established h as a homotopic connection of the functions f_m and

$$g_m : x^0 \mapsto (\phi_m(x^0))(T) - x^0.$$

It is evident that $h_0 : [0, 1] \times \overline{\Omega^0}$, $(s, x^0) \mapsto (1 - s)g_m(x^0) - sx^0$ defines a homotopic connection of g_m and the odd map $x^0 \mapsto -x^0$. By properties B-3 and B-4, we may conclude that

$$\deg(F, \Omega^0, 0) = \deg_B(f_m, \Omega^0, 0) = \deg_B(g_m, \Omega^0, 0) \neq 0.$$

Since $0 \notin F(\partial\Omega^0)$, $d(0, F(x)) > 0$ for all $x \in \partial\Omega^0$. Next, we show that $x \mapsto d(0, F(x))$ is lower semicontinuous on $\partial\Omega^0$. Let $x \in \partial\Omega^0$ be arbitrary and let $(x_n) \subset \partial\Omega^0$ be a convergent sequence with limit x. Let subsequence (x_{n_k}) be such that

$$\lim_{k \to \infty} d(0, F(x_{n_k})) = \liminf_{n \to \infty} d(0, F(x_n)).$$

For each k, let y_k be a minimizer of $\|\cdot\|$ over compact $F(x_{n_k})$, that is, $\|y_k\| = d(0, F(x_{n_k}))$. By upper semicontinuity of F, for each $\epsilon > 0$ we have $y_k \in F(x_{n_k}) \subset F(x) + B_\epsilon$ for all k sufficiently large. By compactness of $F(x)$, it follows that (y_k) has a convergent subsequence (that we do not relabel) with limit $y \in F(x)$, whence

$$d(0, F(x)) = \min_{v \in F(x)} \|v\| \leq \|y\| = \lim_{k \to \infty} \|y_k\| = \liminf_{n \to \infty} d(0, F(x_n)).$$

Thus, $x \mapsto d(0, F(x))$ is positive-valued and lower semicontinuous on compact $\partial\Omega^0$ and so attains a positive minimum value thereon. We may now conclude the existence of a scalar $\mu > 0$ such that $p \notin F(\partial\Omega^0)$ for all $p \in B_\mu$. By Theorem 3(i) we deduce that, for every such p,

$$\deg(F, \Omega^0, p) = \deg(F, \Omega^0, 0) \neq 0.$$

Therefore, by Theorem 3(ii), for each $p \in B_\mu$ there exists $x \in \Omega^0 = B_\delta$ such that $p \in F(x) = f(x, k(x))$. It immediately follows that each $p \in B_\mu$ is the image, under f, of some point $(x, u) \in B_\delta \times \mathbb{R}^M$. This completes the proof of Theorem 1.

It remains only to prove Corollary 1. Let $\mathcal{N}$ be any open neighbourhood of $0 \in X$ and let $\rho > 0$ be such that $B_\rho \subset \mathcal{N}$. Let $k \in \mathcal{K}$ be equi-asymptotically stabilizing. Then there exist scalars $T > 0$ and δ, τ, with $0 < \tau < \delta < \rho$, such that the equi-constricting property of Definition 2 holds. Now, arguing exactly as in the proof of Theorem 1, it follows that $f(B_\delta \times \mathbb{R}^M)$ (and so, a fortiori, $f(\mathcal{N} \times \mathbb{R}^M)$) contains an open neighbourhood of $0 \in X$.

Acknowledgment. The author is indebted to his colleague, J. F. Toland of the University of Bath, for many helpful discussions.

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, New York, 1984.
[2] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, & H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.
[3] A. Cellina and A. Lasota, *A new approach to the definition of topological degree for multivalued mappings*, Rend. Acc. Naz. Lincei, 47 (1969), pp. 434–440.
[4] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[5] J. M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.
[6] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, New York, 1985.
[7] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer, Dordrecht, 1988.
[8] M. A. Krasnosel'skiĭ and P. P. Zabreĭko, *Geometrical Methods of Nonlinear Analysis*, Springer-Verlag, Berlin, 1984.
[9] W. V. Petryshyn and P. M. Fitzpatrick, *A degree theory, fixed point theorems, and mapping theorems for multivalued noncompact mappings*, Trans. Amer. Math. Soc., 194 (1974), pp. 1–25.
[10] E. P. Ryan, *Discontinuous feedback and universal adaptive stabilization*, in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser, Boston, 1990, pp. 245–258.
[11] E. D. Sontag, *Mathematical Control Theory*, Springer-Verlag, New York, 1990.
[12] J. R. L. Webb, *On degree theory for multivalued mappings and applications*, Bol. Un. Math. Ital., 9 (1974), pp. 137–158.
[13] J. Zabczyk, *Some comments on stabilizability*, Appl. Math. Optim., 19 (1989), pp. 1–9.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1605–1620, November 1994

007

ABNORMAL MINIMIZERS*

RICHARD MONTGOMERY†

Abstract. This paper constructs the first example of a singular, abnormal minimizer for the Lagrange problem with linear velocity constraints and quadratic definite Lagrangian, or, equivalently, for an optimal control system of linear controls, with k controls, n states, and a running cost function that is quadratic positive-definite in the controls. In the example, $k = 2$, $n = 3$, and the system is completely controllable. The example is stable: if both the control law and cost are perturbed, the singular minimizer persists. Its importance is due, in part, to the fact that it is a counterexample to a theorem that has appeared several times in the differential geometry literature. There, the problem is called the problem of finding minimizing sub-Riemannian geodesics, and it has been claimed that all minimizers are normal Pontryagin extremals [*The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962]. (If the number of states equals the number of controls, then the problem is that of finding Riemannian geodesics.) The main difficulty is proving minimality. To do this, the length (cost) of the abnormal is compared with all competing normal extremals. A detailed asymptotic analysis of the differential equations governing the normals shows that they are all longer.

Key words. nonholonomic distributions, sub-Riemannian or singular geometry, geodesics, abnormal extremals

AMS subject classifications. primary 49J15, 49K40, 49Q; secondary 58E10, 58B20, 58E25, 58A17, 58O30, 53B99, 53C15

1. Introduction.

Consider a system

$$\dot{q} = \sum_{i=1}^{k} u_i X_i(q),$$

linear in the k controls u_i. Here the point q evolves on an n-manifold Q, and we assume $k < n$. If Q is connected and if the vector fields X_i are bracket generating, then according to Chow's theorem we can find piecewise continuous controls $u_i(t)$ that steer between any two points of Q. Consider the optimal control problem of finding controls $u_i(t), 0 \leq t \leq 1$ that steer between two given points in such a way as to minimize the L^2-norm $\frac{1}{2}\int_0^1 \sum_i u_i^2\, dt$ over all controls that steer between these points. If desired, this can be reformulated as a minimum time problem. Impose the bounds $\sum_i u_i^2 \leq 1$ and find the controls joining the two points in the minimum possible time.

This problem can be viewed as a generalization of the problem of finding geodesics on a Riemannian manifold. Let $D_q = \text{Span}\{X_i(q) : i = 1, 2, \ldots, k\}$. We assume that the X_i are everywhere linearly independent. Then $D = \cup_{q\in Q} D_q$ forms a *distribution*, that is, a linear subbundle of the tangent bundle TQ of Q. By declaring the X_i to be orthonormal, we define an inner product $\langle \cdot, \cdot \rangle$ on the k-planes D_q. The pair $(D, \langle \cdot, \cdot \rangle)$ is called a *sub-Riemannian metric* on Q. And the X_i form a *framing* of D.

Call an absolutely continuous path *horizontal*, *integrable*, or a *D-curve* if its derivative lies in D wherever it is defined. The *length* of such a path is $\int ds$, where $ds = \sqrt{\langle \dot{\gamma}, \dot{\gamma} \rangle}\, dt$. The *distance* between two points is the infimum of the lengths of all horizontal curves joining them. Our problem is to find a horizontal curve joining the two given points whose length realizes the distance between them. We call such a path a minimizing geodesic.

The Pontryagin maximum principle [19] provides necessary conditions for a curve to be a minimizing geodesic. The Hamiltonian equations that govern the normal Pontryagin extremals will be called the *geodesic equations*. They are similar to the geodesic equations

* Received by the editors March 3, 1993; accepted for publication (in revised form) July 11, 1993. Part of this work was completed while the author was visiting Mathematical Sciences Research Institute, University of California, Berkeley, California with the support of National Science Foundation grant DMS-8807219.

† Department of Mathematics, University of California, Santa Cruz, California 95064.

of Riemannian geometry, the principle difference being that the fiber-quadratic form defining the Hamiltonian has rank $k < n$. We will also say that a curve in Q satisfies the geodesic equations if it is the projection of a solution to the geodesic equations.

Our main result is the construction of a sub-Riemannian metric on $\mathbf{R}^3$ and a horizontal curve in this space that is a minimizing geodesic but does not satisfy the geodesic equations. This curve is necessarily the projection of an abnormal extremal. The distribution on $\mathbf{R}^3$ is bracket generating. The curve is smooth.

The assertion that every minimizer is the projection of a normal extremal, that is, that every minimizing geodesics satisfies the sub-Riemannian geodesic equations, has occurred numerous times in the literature (see, for example, Rayner [21], Strichartz [22], or Taylor [24]). Strichartz later retracted his claim [23]. Our result came out of attempts to understand this (false) assertion. Gaveau [7] proposed a counterexample to this assertion in 1977 but Brockett [4] found a fundamental flaw in his reasoning. For completeness we have included an appendix containing Gaveau's example and Brockett's refutation of it.

A basic property of our curve is that it is rigid in the sense it admits no piecewise-C^1 endpoint-preserving variations through D-curves. Having no variations, such a curve is automatically an extremal (in the C^1 topology) for any functional. Because this extremality has no relation to the function being minimized (length) there is no reason that the curve should solve its Euler–Lagrange equations, which are the sub-Riemannian geodesic equations. Our main difficulty will be to show that our curve is in fact a minimum.

One of our main tools is a partial dictionary between sub-Riemannian geometry and gauge theory begun in [16]. This dictionary tells us that the sub-Riemannian geodesic equations for our example are identical to the equations of a charged particle traveling in the plane under the influence of a magnetic field determined by the distribution.

The paper is organized as follows. Section 2 contains basic definitions and notation. Section 3 describes the counterexample and some of its properties. Section 4 gives a heuristic proof that it is a minimizing geodesic. Sections 5 and 6 contain the proof; the core of the proof is in §§5.3 and 5.4. In §7 we show that our example is stable.

Since this article was first written (1991), several other proofs of minimality have appeared. One is due to I. Kupka [13] of the University of Toronto and is similar in spirit to ours. Another is due to W.-S. Liu and H. Sussmann [14] and is based on an inequality.

2. Preliminaries. A sub-Riemannian (sR) structure on a smooth manifold Q consists of a nonintegrable distribution $\mathcal{D} \subset TQ$ together with a smoothly varying inner product $\langle \cdot, \cdot \rangle_q$ on the fibers $\mathcal{D}_q$ of this distribution. In this section we describe the basic notions and notations of sR geometry (for more detailed treatments, see Strichartz [22], Brockett [3], [4], Pansu [18], Hamenstädt [10], Rayner [21], Hermann [11], [12], or Vershik and Gershkovich [9]). Sub-Riemannian structures have also been called "Carnot–Caratheodory metrics," "singular Riemannian metrics," and "nonholonomic Riemannian metrics."

A path in Q is said to be horizontal if it is locally rectifiable (for example, piecewise differentiable) and if its derivatives, whenever defined, lie in $\mathcal{D}$. The *length* of a horizontal path γ is the integral over t of $\sqrt{\langle d\gamma/dt,\, d\gamma/dt \rangle_{\gamma(t)}}\, dt$. The *distance* $d(q_0, q_1)$ between two points q_0, q_1 in Q is the infimum of the lengths of the horizontal paths joining them.

DEFINITION 1. *A path $\gamma : [a, b] \to Q$ is called a minimizing geodesic if it is the shortest horizontal path joining its endpoints.*

DEFINITION 2. *A path $\gamma : I \subset \mathbf{R} \to Q$ is called a geodesic if it is locally a minimizing geodesic. In other words, each $t_0 \in I$ is contained in a nontrivial closed subinterval $J \subset I$ such that γ restricted to J is a minimizing geodesic. If, in addition, $d(\gamma(t_2), \gamma(t_1)) = |t_2 - t_1|$ whenever $|t_2 - t_1|$ is sufficiently small, then we say that γ is a unit speed geodesic.*

When do minimizing sub-Riemannian geodesics exist? The distribution is said to be

bracket generating (or to satisfy Hörmander's condition) if every point q in Q has a neighborhood on which there is a local frame field $X_i, i = 1, \ldots,$ rank$\mathcal{D}$ for $\mathcal{D}$ such that the X_i together with all of their iterated Lie brackets $[X_i, X_j], [X_i, [X_j, X_k]], \ldots$, span the tangent space at every point of this neighborhood. (This condition is independent of the choice of local frame.) A corollary to a classical theorem of Chow [6] and Rashevski [20] asserts that if $\mathcal{D}$ is bracket generating and Q is connected, then any two points of Q can be joined by a horizontal path. In particular, it follows from this and the Arzela–Ascoli theorem that if $\mathcal{D}$ is bracket generating and if the inner product on $\mathcal{D}$ is the restriction of a complete Riemannian metric, then any two points of Q can be joined by a minimizing sub-Riemannian geodesic.

A formal application of the method of Lagrange multipliers yields differential equations for the geodesics. They can also be derived from the Pontryagin maximum principle [19], in which case they are the equations for the normal extremals. We call them the sub-Riemannian geodesic equations. To define them, set $a_{ij}(q) = \langle X_i, X_j \rangle_q$, where X_i is a frame for $\mathcal{D}$ as before. Let a^{ij} be the inverse matrix. Then

$$g = \sum \alpha^{ij} X_i X_j$$

is independent of the choice of frame field and defines a fiber-quadratic form on the cotangent bundle T^*Q, that is, a symmetric covariant tensor of type (2,0). (We can define g intrinsically by saying that it is a vector bundle map $T^*Q \to TQ$ which satisfies $g(q)^* = g(q)$, image$(g) = \mathcal{D}$, and $p(w) = \langle g(q)(p), w \rangle$ whenever $p \in T_q^*Q$, and $w \in \mathcal{D}_q$.) The function $H : T^*Q \to \mathbf{R}$ defined by

$$H(q,p) = \frac{1}{2} g(q)(p,p)$$

is the Hamiltonian that defines the geodesic equations. (In the Riemannian case, $\mathcal{D}$ is the entire tangent bundle, and this is the usual kinetic energy that generates geodesic flow.) If we choose coordinates q^μ on Q and the corresponding induced coordinates q^μ, p_ν on T^*Q, then

$$H = \frac{1}{2} g^{\mu\nu}(q) p_\mu p_\mu,$$

where $g^{\mu\nu}(q)$ is a symmetric matrix of rank equal to the rank of $\mathcal{D}$ that represents the (2,0) tensor. And the sub-Reimannian geodesic equations are

$$\dot{q}^\mu = \sum g^{\mu\nu}(q) p_\nu; \qquad \dot{p}_\mu = -\frac{1}{2} \sum \frac{\partial g^{\alpha\beta}}{\partial q^\mu} p_\alpha p_\beta. \tag{1}$$

By a slight abuse of language, we call a curve $q(t)$ in Q a solution to the sub-Riemannian geodesic equations if it is the projection of a solution $(q(t), p(t))$ to this system of first-order ordinary differential equations on T^*Q.

LEMMA 1 (Rayner [21, Cor. 2.2], Hamenstädt [10, Cor. 5.9]). *Every solution to the sub-Riemannian geodesic equations,* (1) *above, is a geodesic.*

Our counterexample is to the converse of this lemma.

Remark 1. The lemma does not require any assumptions on the distribution. In particular, it need not be bracket generating. Rayner's proof proceeds by showing that in a neighborhood of any non-self-intersecting solution γ to the sR geodesic equation, there is an extension of the distribution's inner product to a Riemannian metric such that γ is a Riemannian geodesic relative to that extension.

3. The counterexample. Take Q to be $\mathbf{R}^3$ for our example. Let $\mathcal{D}$ be the distribution annihilated by a smooth one-form of the type

$$\Theta = dz + A(r)d\theta,$$

where (r, θ, z) are standard cylindrical coordinates on Q. This means that a vector $(\dot{r}, \dot{\theta}, \dot{z})$ is in $\mathcal{D}(r, \theta, z)$ if and only if $\dot{z} = -A(r)\dot{\theta}$. We require that A has a single nondegenerate maximum at $r = 1$. Thus, $dA/dr\,|_{r=1} = 0$ and $d^2A/dr^2\,|_{r=1} < 0$. For example,

$$A = \frac{1}{2}r^2 - \frac{1}{4}r^4 \tag{2}$$

is a good choice and defines a smooth one-form on all of $\mathbf{R}^3$. We refer to this choice as the "model case."

Let us check the bracket-generating condition for $\mathcal{D}$. $X_1 = \partial/\partial r$ and $X_2 = \partial/\partial\theta - A(r)(\partial/\partial z)$ form a local nonorthonormal frame field for $\mathcal{D}$. $[Y_1, Y_2] = -(dA/dr)(\partial/\partial z)$ so that $X_1, X_2, [X_1, X_2]$ span $\mathbf{R}^3$ everywhere except at points on the unit cylinder $r = 1$. But here $[X_1, [X_1, X_2]] = -(d^2A/dr^2)(\partial/\partial z) \neq 0$ and so $Y_1, Y_2, [Y_1, [Y_1, Y_2]]$ span $\mathbf{R}^3$, and the bracket-generating condition holds everywhere.

The fiber metric for our example is the restriction of the form $dx^2 + dy^2$ to $\mathcal{D}$. Here (x, y, z) are the usual Cartesian coordinates, so that $dx^2 + dy^2 = dr^2 + r^2d\theta^2$. In other words, the length of a horizontal curve is equal to the Euclidean length of its projection to the x, y plane.

The horizontal curves that provide the counterexample are any horizontal curves lying on the unit cylinder $r = 1$. Thus, they are helices with pitch $-A(1)$. One such curve is parameterized according to $(r, \theta, z) = (1, \theta, -A(1)\theta)$. We call this curve $\hat{C}$. Its projection to the xy plane is the unit circle C.

Our main result follows.

THEOREM 1. *The above-defined distribution $\mathcal{D}$ is bracket generating. The helix $\hat{C}$ or any subarc thereof is a geodesic. It does not satisfy the sub-Riemannian geodesic equations*, (1), *above. In particular, there are minimizing sub-Riemannian geodesics that do not satisfy the sub-Riemannian geodesic equations.*

Analytically speaking, this theorem says that for some sufficiently small positive number θ_0, the arc of the helix $\hat{C}$ from $\hat{C}(0)$ to $\hat{C}(\theta_0)$ is the shortest horizontal curve among all horizontal curves that join $\hat{C}(0)$ to $\hat{C}(\theta_0)$.

Remark 2. Bär [1] gives quite a different kind of a counterexample. His curve γ is also a geodesic that does not satisfy the geodesic equations. However, every sufficiently short arc of his geodesic does satisfy the geodesic equations, whereas *no* subarc of our geodesic satisfies the equations. In his example, normal extremals $\xi : I \to T^*Q$ that project to subarcs of γ can be found, but they cannot not be chosen consistently so as to be continous over all of γ. It follows from the maximum principle that his curve is not a minimizing geodesic.

4. The helix does not satisfy the geodesic equations. An orthonormal framing of D is provided by

$$X_1 = \frac{\partial}{\partial r}$$

and

$$X_2 = \frac{1}{r}\left\{\frac{\partial}{\partial\theta} - A(r)\frac{\partial}{\partial z}\right\}.$$

It follows that the normal Pontryagin Hamiltonian (cf. the equation immediately preceding (1)) is

$$H(r,\theta,z,p_r,p_\theta,p_z)=\frac{1}{2}\left\{p_r^2+\frac{1}{r^2}(p_\theta-p_zA(r))^2\right\}. \tag{3}$$

In Cartesian coordinates,

$$H(x,y,z,p_x,p_y,p_z)=\frac{1}{2}\{(p_x-p_zA_x(x,y))^2+(p_y-p_zA_y(x,y))^2\}.$$

Here $A_x\,dx+A_y\,dy=A(r)d\theta$ and the Cartesian p's are related to the cylindrical p's in the usual way, that is, through the formula $p_x\,dx+p_y\,dy=p_r\,dr+p_\theta\,d\theta$.

This is exactly the Hamiltonian for a particle with charge

$$e=p_z$$

traveling in the xy plane under the influence of a magnetic field normal to the plane and with strength

$$B=\frac{1}{r}\frac{dA}{dr}.$$

It is important in this regard that H is independent of z, for it follows from this that one of the geodesic equations is

$$\dot{p}_z=0$$

and thus the "charge" is constant along solutions.

To write these equations, let

$$\pi:\mathbf{R}^3\to\mathbf{R}^2;\qquad \pi(x,y,z)=(x,y)$$

be the map that projects out the vertical coordinate. If γ is a solution to the geodesic equations, write

$$c=\pi\circ\gamma$$

for its planar projection. Now write c in terms of complex variables $c=x+iy$. Then by straightforward calculation, c satisfies

$$\ddot{c}=-ieB(c)\dot{c}, \tag{4}$$

where $B(x,y)=d\Theta/dx\wedge dy=\partial A_y/\partial x-\partial A_x/\partial y$. In the particular case where A is a function of r alone, we easily calculate that B is given by the previous formula. If γ is parameterized by arclength, or what is the same, if $H=\frac{1}{2}$ along γ, then the above differential equation can be rewritten as

$$\kappa(s)=-eB(c(s)), \tag{5}$$

where κ is the curvature of the unit speed curve c. In our example, $B=0$ on the unit circle C. But C has curvature 1 and is the projection of our helix $\hat{C}$. Thus, $\hat{C}$ cannot satisfy the sub-Riemannian geodesic equations.

Remark 3. This connection between the equations of motion of a particle in a magnetic field and the sub-Riemannian geodesic equations was first pointed out in [16]. If the codimension of D is larger than one, then the equations are related to those of a "classical quark" in a non-Abelian gauge field. The relation only holds if the sub-Riemannian structure admits a Lie group of symmetries acting transversely to its distribution.

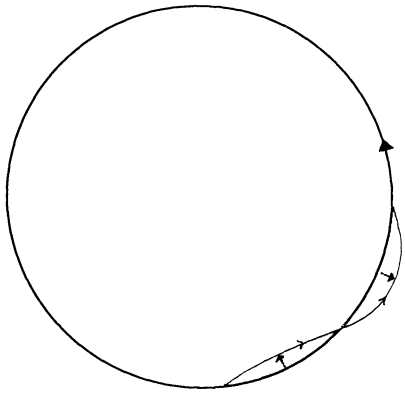


FIG. 1

5. Rigidity: A heuristic proof of the theorem. It remains to show that sufficiently small subarcs of $\hat{C}$ are minimizing geodesics. In this section we give an incomplete but illuminating argument in support of this fact.

Consider a horizontal curve γ in $\mathbf{R}^3$ that connects the point $(x_0, y_0, 0)$ to the point (x_0, y_0, z_1). Thus, the projected curve $c \subset \mathbf{R}^2$ is closed. By Stoke's theorem and the fact that $dz = -Ad\theta$ along γ,

$$z_1 = -\int\int d\Theta = -\int\int B(r) r\, dr\, d\theta, \tag{6}$$

where the last integral is over the (oriented) region in the plane enclosed by c. Following the theory of magnetism, we call the integral $\int\int B(r) r\, dr\, d\theta$ the *flux* enclosed by c. Recall that the (Euclidean) length of c is the same as the sub-Riemannian length of γ. It follows that γ is a minimizing geodesic if and only if its projection to the xy plane minimizes length among all loops based at (x_0, y_0) and enclosing an amount of flux equal to $-z_1$.

Now consider our helix $\hat{C}$. The flux enclosed by going once around $C = \pi \circ \hat{C}$ is $2\pi A(1) = \int\int_D B\, dx\, dy$, where D is the unit disc. Imagine perturbing $\hat{C}$ to form the horizontal curve γ and consequently perturbing its projection C to form c. If we push part of C into the interior of the unit disc, we have subtracted flux because B is positive in the interior. On the other hand, if we push part of C to the exterior of the disc, then we add negative flux, that is, we also subtract flux (Fig. 1). No matter how we perturb C, we decrease the flux and hence increase the height difference z_1 and thus violate the endpoint conditions. We conclude that there are no allowable variations of $\hat{C}$, and consequently it is a local minimum for our constrained variational problem. This heuristic argument can easily be turned into a rigorous proof of the following.

Assertion 1. There are no piecewise C^1 endpoint-preserving variations of $\hat{C}$ through horizontal curves except those whose variation field is tangential to $\hat{C}$.

This shows that $\hat{C}$ is an isolated point in the space of all horizontal piecewise C^1 unparameterized curves that join $\hat{C}(0)$ to $\hat{C}(2\pi)$ with the (piecewise) C^1 topology on this path space. Being an isolated point, it is automatically a local minimum for any function on this space and in particular for the length functional.

Remark 4. Extremals of this type have been a source of major difficulties to practitioners of "classical" (one independent variable) calculus of variations. Such extremals are called "rigid" by Young [26], "abnormal" by Bliss [2], and of "maximal class" by Carathéodory [5]. If the Pontryagin maximum principle [19] is applied to the problem of finding sub-Riemannian geodesics, we find that they are precisely the abnormal extremals and that these are also the

singular extremals in this case. We especially recommend the treatment in the beginning of volume 2 of Young's book regarding these extremals. Also of interest is the treatment of Morse and Mayers [17].

$\hat{C}$ is not isolated in the C^0 or even the Sobolev H^1 topology on the space of all horizontal curves with its endpoints. We can find arbitrarily C^0- or H^1-close curves to $\hat{C}$ that $\hat{C}$'s endpoints by making tiny kinks in chords to the circle C. One might then wonder whether it is possible to make sufficiently small kinks in C so as to shorten its length while keeping its flux the same. The theorem says that this is not possible, at least for small enough arcs of $\hat{C}$. Our work lies in showing this.

6. Proof.

6.1. The proof from two propositions. Theorem 1 follows immediately from the following two propositions.

PROPOSITION 1. *Any minimizer not lying in the cylinder must satisfy the sR geodesic equations.*

PROPOSITION 2. *There is a positive number θ_* such that for all $\theta \leq \theta_*$, every solution γ to the sub-Riemannian geodesic equations with endpoints $\hat{C}(0)$ and $\hat{C}(\theta)$ has length $(\gamma) > \theta$.*

Proof of Theorem 1. Combine the two propositions, recalling that θ is the sub-Riemannian arclength along $\hat{C}$. □

6.2. Proof of Proposition 1. The salient fact is that the unit cylinder is precisely where the distribution D fails to be a contact distribution.

DEFINITION 3. *A rank-two distribution on a three-manifold is said to be contact at a point q if for some (and hence any) frame field X_1, X_2 of D defined near q the vectors $X_1(q), X_2(q)$ together with their Lie bracket $[X_1, X_2](q)$ form a basis for the tangent space at q.*

With this in mind, we can find various proofs of Proposition 1 in the literature (see for example, Zhong [8], Hamenstädt [10], Strichartz [22], or Hermann [11]). We give another proof based on the maximum principle.

Proof of Proposition 1. According to that principle, every minimizer is a Pontryagin extremal and all such extremals can be divided into the abnormal and the normal ones. The normal ones satisfy, by definition, the sR geodesic equations. The proposition now follows from the following lemma.

LEMMA 2. *Every nonconstant abnormal Pontryagin extremal projects to a D-curve that lies on the locus of points where D fails to be contact.*

Proof of Lemma. Let X_1, X_2 be the frame for D used earlier and let X_3 be their Lie bracket. Let P_i be the corresponding "power functions" as above $P_i(q,p) = p(X_i(q))$. The maximum principle tells us to introduce the multipliers $(p(t), \lambda_0) \in (\mathbf{R}^2)^* \oplus \mathbf{R}$, $(p(t), \lambda_0) \neq (0,0)$. For fixed (q, p, λ_0), the controls $u(t)$ leading to an extremal must maximize

$$H(q, u, p, \lambda_0) = u_1 P_1 + u_2 P_2 - (\lambda_0)\frac{1}{2}(u_1^2 + u_2^2).$$

For an abnormal extremal $\lambda_0 = 0$ so the only possibility is that $P_1 = P_2 = 0$, $p \neq 0$ along the extremal. If $u(t)$ is the control inducing the abnormal extremal, then the extremal is an integral curve of the Hamiltonian flow of the time-dependent Hamiltonian $H_*(q,p,t) = u_1(t)P_1 + u_2(t)P_2$. Because $\{P_1, P_2\} = -P_3$, we find that $\dot{P}_1 = -u_2 P_3, \dot{P}_2 = u_1 P_3$. Thus, we require that $P_1 = P_2 = P_3 = 0$ along an abnormal. (The u_i cannot both be zero because the extremal is not a single point.) But these are the components of the covector p on the vectors X_1, X_2, X_3. Consequently, if these vectors form a basis at a point q we must have $p = 0$, which is not allowed. Therefore, the curve $q(t)$ must lie on the locus of points where X_1, X_2, and X_3 become linearly dependent, which is to say where D fails to be contact.

Remark 5. From the point of view of differential topology, Proposition 1 is a statement about the regularity of the endpoint map. This is the map that assigns to each D-path γ beginning at a point q_0 its endpoint $\text{end}(\gamma) = q(1)$. (Alternatively, end is the input-output map.) The proposition states that if D is contact at one point of the curve γ, then end is a submersion at γ. Hence, by the implicit function theorem the subset $\text{end}^{-1}(q_1)$ of D-paths that join q_0 to q_1 forms a smooth submanifold of the space of all horizontal paths based at q_0. It is then legitimate to use the standard method of Lagrange multipliers applied to the functional (length) $+\lambda$(end) to find the equations characterizing extremals. These are the sub-Riemannian geodesic equations.

7. Proof of Proposition 2. The proof is by contradiction. Suppose it is false. Then there must exist a sequence of lengths s_i decreasing to zero and a sequence $\gamma_i : [0, t_i] \to \mathbf{R}^3$ of minimal solutions to the sR geodesic equations, which have the same endpoints as $\hat{C}[0, s_i]$, are parameterized by arclength, and are shorter than s_i. Thus, $\gamma_i(0) = \hat{C}(0)$ and $\gamma_i(t_i) = \hat{C}(s_i)$. Also $\|dc_i/dt\| = 1$, where we write $c_i = \pi \circ \gamma_i$ and finally

$$\sin(s_i) \leq t_i \leq s_i.$$

The first inequality expresses the fact that the Euclidean distance between the endpoints of $C[0, s_i]$ is $\sin(s_i)$.

Let

$$m_{*i} = \max |1 - r_i(t)|,$$

where

$$\gamma_i(t) = (r_i(t), \theta_i(t), z_i(t))$$

is the expression for γ_i in cylindrical coordinates. Then we also must have

$$m_{*i} \leq s_i, \tag{7}$$

for otherwise the length of γ_i would be greater than s_i.

The proof is based on an analysis of the geodesic equations as $i \to \infty$. The salient fact is that the "charges" e_i must go to infinity as i does. To perform the analysis we need the equations in both their geometric form, $\kappa(s) = -eB(c(s))$, and their Hamiltonian form,

$$\dot{\theta} = \frac{1}{r^2}(p_\theta - eA(r)), \tag{8}$$

$$\dot{r} = p_r,$$

$$\dot{z} = -A(r)\dot{\theta},$$

$$\dot{p}_\theta = 0,$$

$$\dot{p}_z = 0.$$

The differential equation for p_r that we use is found by solving for $\dot{r}$ in the formula for the Hamiltonian

$$\frac{1}{2} = H = \frac{1}{2}\left\{\dot{r}^2 + \frac{1}{r^2}(p_\theta - eA(r))^2\right\}. \tag{9}$$

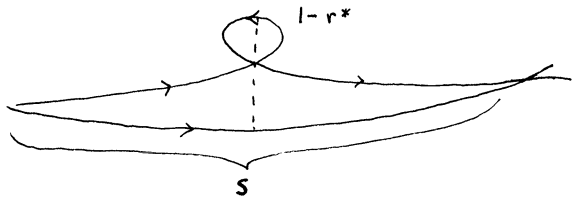


FIG. 2

The following facts regarding solutions will be needed.

Fact 1. p_θ and $p_z = e$ are constant along solutions.

Fact 2. Suppose that the projection $c = \pi \circ \gamma$ of a solution is tangent to a circle $r = \text{const}$ at some point $c_* = c(t_*)$. Then c is symmetric with respect to reflection about the radial line through this point. Analytically, this reads $c(t_* + s) = R(c(t_* - s))$, where R is the reflection about the line joining the origin to c_*.

These two facts follow directly from the equations and their invariance with respect to orthogonal transformations. The following facts are direct consequences of Facts 1 and 2 and the equations.

Fact 3. Let $\beta(n)$ denote the angle a projected solution c makes with the unit circle at its nth crossing. Then $\beta(n+1) = -\beta(n)$. In particular, because $\dot\theta(n) = \cos(\beta(n))$ at each crossing, we have $\dot\theta(n) = \dot\theta(n+1)$.

DEFINITION 4. *An arc of solution γ will mean that part lying between two consecutive intersections with the unit cylinder $r = 1$ or the projection to the xy plane of such an arc. Thus, if there are $n+1$ such intersections, then there are n arcs.*

An arc will be called "interior" or "exterior" depending on whether or not it is interior or exterior to the unit disc. By the "corresponding arc" of C we mean the shorter arc of C lying between the endpoints of such an arc. By the "height" or "vertical coordinate" of an arc we mean the difference between the z-coordinates of γ corresponding to the two endpoints of the arc. Thus, the height is $-\int_{\text{arc}} A d\theta$.

Fact 4. Extremals of $B(c(s))$ along any projected solution c coincide with points of tangency to some circle. Because the curve is symmetric, upon reflection about the line through such a point there is exactly one such extremal on any arc and it is a strict maximum for $|B|$ and for $|1 - r|$ along the arc.

Fact 5. For $e \neq 0$ and $r \neq 1$, the sign of the curvature κ is equal to $-\text{sgn}(e)\text{sgn}(1-r)$.

We begin our analysis by eliminating certain solution curves. Solutions with $e = 0$ can be eliminated because they project to straight line segments. If such a curve satisfies the xy endpoint conditions, its projection must be a chord of the circle. So, by the argument of §4 it cannot satisfy the z-endpoint condition. Solutions with $\dot\theta_0 = 0$ can be eliminated because they must remain exterior to the cylinder (see Fact 3). It follows from the convexity of the unit disc that their projections are always longer than the corresponding segment of the circle, and so they cannot be minimizing.

Solutions with $\dot\theta_0 > 0$ and $e > 0$, or with $\dot\theta_0 < 0$ and $e > 0$ can be eliminated because they can never satisfy the z-component of the endpoint conditions. To see this, observe that the projected arcs of such a curve can never have self-intersections. For example, if $0 < \beta_0 < \pi/2$ and $e > 0$, then by Fact 5 the curvature in the arc's interior is strictly negative. Coupled with Fact 2 this shows that the arc can have no self-intersection. (For example, viewed as a graph over its tangent at $t = 0$, it forms a concave function; see Fig. 2.) Upon crossing the cylinder, the sign of the curvature of the arc switches and the same argument applies again. (As a graph it becomes strictly convex.) Therefore, the Green's theorem argument of §4 (the heuristic

proof) is valid, and these curves can never satisfy the vertical (z) endpoint condition.

We are left with analyzing sequences of solutions $\gamma_i(t) = (x_i(t), y_i(t), z_i(t))$ with initial conditions $\dot\theta_{i,0} > 0, e_i < 0$, or $\dot\theta_{i,0} < 0,\ e_i > 0$. Set

$$h_i = z_i(t_i) - (-A(1)s_i)$$

so that our last endpoint condition is

$$h_i = 0.$$

We have

$$h_i = -\int_{c_i} A d\theta + A(1)s_i.$$

Now

$$A(r) = A(1) - \eta(r)^2,$$

where

$$\eta = \eta(r) = b(1-r) + O((1-r)^2); \qquad b > 0$$

is a smooth invertible function of $1 - r$ near $r = 1$. This follows from the fact that A has a nondegenerate maximum at $r = 1$. (For the model case (2), $A(r) = \frac{1}{2}r^2 - \frac{1}{4}r^4$, we find $A = \frac{1}{4} - \eta(r)^2$ with $\eta(r) = (1-r^2)/2 = (1-r) - \frac{1}{2}(1-r)^2$. Thus $b = 1$. We have chosen $b > 0$ so that $\eta > 0$ when $r < 1$. Then $b = +\sqrt{-\frac{1}{2}(d^2A(1)/dr^2)}$.) It follows that

$$h_i = \int_0^{t_i} \eta^2 \dot\theta \, dt. \tag{10}$$

Because η^2 is positive off of the unit cylinder, it follows that somewhere along the solution the sign of $\dot\theta$ must become opposite to its initial sign, $\dot\theta_0$.

We now fix attention on the particular arc for which this sign switch occurs. Later, we show that the sign switch must occur on every subarc of the solution. Most importantly, we show that the length of this solution arc (and all of these other subarcs) is longer than the corresponding subarc of the circle.

Without loss of generality we assume that $\dot\theta_0 > 0$, the argument for the contrary case being identical. The curvature and symmetry conditions guarantee that the sign switch occurs at the midpoint of the arc. The qualitative shape of the arc is indicated in Fig. 2. We use a subscript * to denote any variable evaluated at this midpoint; thus,

$$\dot r_* = 0.$$

Now $H = \frac{1}{2} = \frac{1}{2}\{\dot r^2 + r^2\dot\theta^2\}$ for any unit speed solution. Therefore

$$1 = r_*^2\dot\theta_*^2.$$

As we saw at the beginning of the proof of Proposition 2 (7), $m_* = |1 - r_*| \to 0$ so that

$$r_* \to 1, \qquad \eta_* \to 0,$$

and

$$\dot{\theta}_* \to -1$$

as $i \to \infty$.

Equation (8) at $r = 1$ tells us that the initial angular velocity is

$$\dot{\theta}_{0,i} = p_{\theta,i} - e_i A(1).$$

Because p_θ is constant, we can rewrite (8) as

$$\dot{\theta} = \frac{1}{r^2}\{\dot{\theta}_{0,i} + e_i \eta(r)^2\}. \tag{11}$$

Evaluating this equation at the midpoint and taking the limit as $i \to \infty$ and recalling that $\eta_* \to 0$, we see that for the limit just calculated for $\dot{\theta}_*$ to be valid we must have

$$e_i \to -\infty$$

as $i \to \infty$. More precisely,

$$e_i \eta_* \to -2. \tag{12}$$

Define the small parameter ε by

$$|e| = \frac{1}{\varepsilon^2}.$$

Define the rescaled variable Y through the relation

$$\eta = \varepsilon Y.$$

The above limit (12) becomes

$$Y_{*,i} \to \sqrt{2}. \tag{13}$$

(The +/sign occurs because the arc is interior to the unit disc.) Also, $A d\theta = A(1)d\theta - \varepsilon^2 Y^2\, d\theta$.

Remark 6. It follows from (7) above that

$$\varepsilon_i = O(s_i)$$

because

$$\eta_* = b m_* + O(\eta_*^2).$$

(A refinement of (7) shows that $m_* = o(s_i)$ and so, in fact, $\varepsilon = o(s_i)$.)

Now

$$\dot{\eta}^2 = R_i(Y), \tag{14}$$

where the function $R = R_i$ is obtained by solving (9) for $\dot{r}^2$ and then using our change of variables $r \to \eta \to Y$. We calculate

$$R_i(Y) = R(Y, \varepsilon_i, \dot{\theta}_{i,0}) = b^2[1 - (1 - Y^2)^2] + O(\varepsilon_i) + O(1 - \dot{\theta}_{i,0}).$$

Now η_* is the unique zero of R_i on its arc, because the midpoint of the arc is the only place where $\dot{\eta} = 0$ (see Fact 4 above). Because R_i and the right-hand side of the equation for $\dot{\theta}$ (11) do not depend on time, it follows that on every arc of our solution $\dot{\theta}$ is negative and close to -1 at that arc's midpoint and that the value of η, there, is η_*.

We complete the argument now by showing that for i large enough, each subarc of the solution is longer than the corresponding arc of the circle. (If the arc is exterior to the circle, this follows immediately from convexity. But this observation does not seem to lead to simplification of the proof. For we do not know that a single arc satisfies the endpoint condition $h_i = 0$ and so several may need to be concatenated to satisfy the vertical endpoint condition.) For this purpose, let us change the meaning of the original notation so that now our curves γ_i each consist of single arcs, with lengths t_i and subtending angles s_i: $\theta(t_i) = s_i$. But we do not insist on the vertical endpoint condition. By reflectional symmetry it suffices to show that $t_{*,i} - \theta_{*,i}$ is eventually positive. Now this quantity is the integral of $dt - d\theta$ over half an arc. Thus, we now show that

$$\int_0^{t_{*,i}} (1 - \dot{\theta}_i(t))dt > 0 \tag{15}$$

provided i is sufficiently large. To do this we change variables to Y. Now $dt = (dt/d\eta)d\eta = (1/\eta)(\varepsilon)dY$. It follows that

$$dt = \varepsilon \frac{dY}{\sqrt{R_i(Y)}}.$$

Hamilton's equation for $\dot{\theta}$ in terms of the scaled variable Y is

$$\dot{\theta} = \frac{1}{r^2}(\dot{\theta}_0 - Y^2).$$

Thus,

$$\begin{aligned} 1 - \dot{\theta} &= 1 - \frac{1}{r^2}(\dot{\theta}_o - Y^2) \\ &= (1 - Y^2) - \frac{1}{r^2}[1 - Y^2 + (\dot{\theta}_0 - 1)] + Y^2 \\ &= \left(1 - \frac{1}{r^2}\right)(1 - Y^2) + Y^2 + \frac{1 - \dot{\theta}_0}{r^2}. \end{aligned}$$

Now

$$1 - \frac{1}{r^2} = \frac{r^2 - 1}{r^2} = -\frac{2\varepsilon}{b} Y g(Y, \varepsilon_i),$$

where $g(Y, \varepsilon)$ is a smooth function converging uniformly to 1 on the interval of integration $[0, Y_*]$ as $\varepsilon \to 0$. This follows from the expansion $\eta = b(1 - r) + O(\eta^2)$. (For the special model we have $\eta = 1 - r^2/2 = (1 - r) - \frac{1}{2}(1 - r)^2$ exactly.) Thus, our integrand is

$$(1 - \dot{\theta})dt = \varepsilon_i[Y^2 + \varepsilon_i Y k_i(Y)] \frac{dY}{\sqrt{R_i(Y)}} + \frac{1 - \dot{\theta}_0}{r^2}\, dt.$$

Here $k_i(Y) = -2g(Y, \varepsilon_i)/b$, which is a smooth bounded function uniformly converging to a constant on the interval of integration. The last term in the integrand, the one involving dt,

is always nonnegative because $1 > \dot{\theta}_{0,i}$ for unit speed curves. (In fact, this term is strictly positive for interior arcs.) It follows that

$$\limsup \frac{t_{*,i} - \theta_{*.i}}{\varepsilon_i} \geq \limsup \int_0^{Y_*} [Y^2 + \varepsilon_i Y k_i(Y)] \frac{dY}{\sqrt{R_i(Y)}}. \tag{16}$$

Assuming the validity of interchanging limits and integration, we obtain

$$\limsup \frac{t_{*,i} - s_{*,i}}{\varepsilon_i} \geq \int_0^{\sqrt{2}} \frac{Y^2}{b\sqrt{1-(1-Y^2)^2}}\, dY = \frac{\sqrt{2}}{b} > 0.$$

This completes the proof, modulo showing that we may interchange the limit with the integration.

7.1. Interchange of limit and integration. The only difficulty is that the limiting denominator $\sqrt{R(Y,0,0)}$ has a simple zero at $Y = 0$ in the case of a self-intersecting arc. This pole cancels with a factor of Y in the limiting numerator, yielding an integrable function. However, we cannot blindly apply the dominated convergence theorem. (To see the difficulties, consider, for example, the function $Y^2/|Y - \varepsilon|$, which has a simple pole at ε and whose limiting denominator has a simple zero at zero. Its integral over the unit interval is $+\infty$ for $\varepsilon > 0$ but the integral of its limit is $\frac{1}{2}$.)

Recall that the tangent vector to any smooth planar curve c turns by a total amount $\int_c \kappa(s)ds$ upon traversing the curve. For our self-intersecting arcs c_i, this amount of turning is $2\pi + \theta_i - 2\beta_i$, where θ_i is the angle subtended by the arc and β_i is the angle the arc initially makes with the circle. (Thus, if the curve consists of one arc, we have $\theta_i = s_i$.)

Because our arcs solve the geodesic equations, they have curvature

$$\kappa(t) = eB(t) = \frac{1}{\varepsilon^2} 2\eta(s) b(1 + O(\varepsilon)) = \frac{1}{\varepsilon} 2Y(s) b K_i(Y),$$

where the function $K_i = K_i(Y) = 1 + O(\varepsilon_i)$ is a smooth function on the interval $[0, Y_{*,i}]$ and tends uniformly to 1. (For the model case, $b = 1$ and $K_i = 1$ exactly.) Changing variables from t to Y and using the reflectional symmetry of the arc we obtain

$$\int_0^{Y_{*,i}} \frac{Y b K_i(Y)}{\sqrt{R_i(Y)}}\, dY = \frac{\pi}{2} + \frac{\theta_i}{4} - \frac{\beta_i}{2}.$$

Let g_i denote the integrand

$$g_i = \frac{Y(bK_i(Y))}{\sqrt{R_i(Y)}} = \frac{Y b(1 + O(\varepsilon_i))}{b\sqrt{1-(1-Y^2)^2 + O(\varepsilon_i)}}.$$

Then

$$g_i \to g = \frac{1}{\sqrt{2 - Y^2}},$$

where the convergence is uniform on compact subsets of the open interval $(0, \sqrt{2})$. We calculate directly that

$$\int g(Y) dY = \frac{\pi}{2}$$

so that $\int g_i \to \int g$ as well.

Referring back to the integrand of the desired limit (16), set

$$f_i = Y(1 + \varepsilon_i k_i(Y)).$$

Then the integrand for the desired limit is $f_i g_i$. Also

$$f_i \to f = Y,$$

where the convergence is now uniform on the entire closed interval $[0, \sqrt{2}]$.

The following lemma now validates interchanging the limit with the integral sign.

LEMMA 3. *Let b_i be a sequence of numbers converging to a number b. Let g_i be a sequence of nonnegative continuous functions defined on the open intervals (a, b_i) that converge uniformly on every compact subset $[a + \delta, b - \delta]$, $\delta > 0$ of the limit interval (a, b) to the function g. Suppose that $\int_a^{b_i} g_i \to \int_a^b g$, this integral being finite. Let f_i be a sequence of continuous functions on a neighborhood of the* closed *interval $[a, b]$ that converge uniformly on that interval to a function f. Then $f_a^{b_i} f_i g_i \to \int_a^b fg$.*

The proof of this lemma is a straightforward exercise in real analysis. We do not give the proof, except to say that the only difficulty is that the g_i can have poles at the endpoints. The value of f at these endpoints together with the uniform convergence of the f_i and the boundedness of $\int g$ serve to control the sequence of integrals near the endpoints.

This completes the proof of Proposition 2 and hence Theorem 1.

8. Stability. The germ of a distribution at a point is called *stable* if any other distribution that is sufficiently (Whitney-) close to it is locally diffeomorphic to it. For example, the Darboux theorem states that any contact distribution is stable.

The distribution in our example is stable. To see this, we formalize some properties of that distribution. Suppose that $\mathcal{D}$ is a distribution on a three-manifold that is defined in a neighborhood of a point q by the vanishing of a nonzero one-form Θ. Define the function B by writing

$$\Theta \wedge d\Theta = B d^3x,$$

where d^3x is a locally defined volume form. Suppose that B has a transverse zero at q: $B(q) = 0$ and $dB(q) \neq 0$. These are precisely the properties of our example that we need.

DEFINITION 5. *A distribution with the above properties will be called a simply degenerate contact structure. The surface $\{B = 0\}$ is called the singular surface.*

By a theorem of Martinet [15], if Θ is any one form with the above properties, then there exists a smooth nonzero function g and coordinates x, y, z defined in a neighborhood of q such that

$$g\Theta = dz - y^2\, dx$$

(see also Zhitomirskii [26]). Because Θ and $g\Theta$ define the same distribution, this proves the following lemma.

LEMMA 4. *Distributions of simply degenerate contact type are stable.*

THEOREM 2 (Stable counterexample). *Suppose that Q is a three-dimensional sR manifold whose underlying distribution is a simply degenerate contact structure. Then every horizontal curve on the singular surface is a geodesic.*

For an open dense set of fiber inner products on the distribution, these geodesics do not satisfy the geodesic equations. In particular, the abnormal geodesics of Theorem 2 are stable under C^2 perturbations of the sR structure.

We do not prove this theorem. It follows the same lines as our proof of Theorem 1, but the algebra and analysis is considerably messier. There is now a simple proof due to Liu and Sussmann [14].

Remark 7. The fact that the result is purely differential-topological suggests that there is a differential-topological proof. We have not found one yet.

Acknowledgments. I thank Alan Weinstein for bringing Rayner's paper [21] to my attention and Stanislaw Janeczko for referring me to the papers of Martinet [15] and Zhitomirskii [26]. I also thank Ge Zhong and Danny Goroff for extended discussions while at Mathematics Science Research Institute, University of California, Berkeley. I thank the University of California, Berkeley, the National Science Foundation, and the University of California, Santa Cruz for support.

Note added in proof. Bryant and Hsu [R. Bryant and L. Hsu, *Rigidity of Integral Curves of Rank Two Distributions*, Invent. Math., 114 (1993), pp. 435–461.] used the term C^1*-rigid curve* for curves having the rigidity property of our abnormal curve. They showed how to find a large class of such curves for a typical two-plane field in a space of arbitrary dimension. Liu and Sussmann (already referenced) independently discovered this same class of curves and call them *regular abnormal extremals*. They prove that these curves are locally minimizing curves, independent of the choice of inner product on the two-plane fields.

REFERENCES

[1] C. Bär, *Carnot–Caratheodory–Metriken*, Diplomarbeit, Bonn, 1988.
[2] G. A. Bliss, *Lectures on Calculus of Variations*, University of Chicago Press, Chicago, IL, 1946.
[3] R. W. Brockett, *Nonlinear control theory and differential geometry*, Proc. of the Internat. Congress of Mathematicians, Warszawa, 1983.
[4] ———, *Control theory and singular Riemannian geometry*, in New Directions in Applied Mathematics, P. J. Hilton and G. S. Young, eds., Springer-Verlag, New York, 1981.
[5] C. Carathéodory, *Calculus of Variations and Partial Differential Equations of the First Order*, Vol. 2, Holden–Day, San Francisco, California, 1967.
[6] W. L. Chow, *Uber Systeme van Linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., 117 (1939), pp. 98–105.
[7] B. Gaveau, *Principle de Moindre Action propagation de la chaleur et estimees sons elliptiques sur certains groupes nilpotents*, Acta Math., 139 (1977), pp. 95–153.
[8] Z. Ge, *On a variational problem and the spaces of horizontal paths*, Pacific J. Math., 149 (1991), pp. 61–93.
[9] V. Ya Gershkovich and A. M. Vershik, *Non-holonomic manifolds and nilpotent analysis*, J. Geom. Phys. 5 (1988), pp. 407–451.
[10] U. Hamenstädt, *Some regularity theorems for Carnot–Caratheodory metrics*, J. Differential Geom. 32 (1990), pp. 819–850.
[11] R. Hermann, *Some differential geometric aspects of the Lagrange variational problem*, Indiana Math. J., 6 (1962), pp. 634–673.
[12] ———, *Geodesics of singular Riemannian metrics*, Bull. Amer. Math. Soc., 79 (1973), pp. 780–782.
[13] I. Kupka, preprint, University of Toronto, 1992.
[14] W. S. Liu and H. Sussmann, *Shortest paths for Sub-Riemannian metrics on rank* 2 *distributions*, Trans. Amer. Math. Soc., to appear.
[15] J. Martinet, *Sur les Singularites Des Formes Différentielles*, Ann. Inst. Fourier, 20 (1970), pp. 95–178.
[16] R. Montgomery, *The isoholonomic problem and some applications*, Comm. Math. Phys., 128 (1990), pp. 565–592.
[17] M. Morse and S. Mayers, *The problems of Lagrange and Mayer with variable endpoints*, in Marston Morse: Selected Papers, R. Bott, ed., Springer-Verlag, New York, 1981.
[18] P. Pansu, *Metriques de Carnot–Caratheodory et quasiisometries des espaces symetriques de rang un*, Ann. of Math., 129 (1989), pp. 1–60.
[19] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.
[20] P. K. Rashevski, *About connecting two points of complete nonholonomic space by admissible curve*, Uchen. Zap. Ped. Inst. Libknexta, (1938), pp. 83–94. (In Russian.)

[21] C. B. RAYNER, *The exponential map for the Lagrange problem on differentiable manifolds*, Philos. Trans. Roy. Soc. London Ser. A, 262 (1967), pp. 299–344.
[22] R. STRICHARTZ, *Sub-Riemannian geometry*, J. Differential Geom., 24 (1983), pp. 221–263.
[23] ———, *Corrections to "sub-Riemannian geometry,"* J. Differential Geom., 30 (1989), pp. 595–596.
[24] T. J. S. TAYLOR, *Some aspects of differential geometry associated with hypoelliptic second order operators*, Pacific J. Math., 136 (1989), pp. 355–378.
[25] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, Chelsea, New York, 1980.
[26] M. YA. ZHITOMIRSKII, *Singularities and normal forms of odd-dimensional Pfaff equations*, Funktsional. Anal. Prilozxhen. 23 (1989), pp. 70–71.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1621–1642, November 1994

008

INVERSE FUNCTION THEOREMS AND SHAPE OPTIMIZATION*

LUC DOYEN†

Abstract. In this paper the problem of shape optimization under shape constraints is investigated. Using the shape gradient and shape tangent cones, inverse function theorems are established. With these theorems, the existence of Lagrangian or Kuhn–Tucker multipliers for shape optimization problems with equality or inequality constraints is proved.

Key words. shape optimization, shape gradient, tangent cones, velocity cones, inverse function theorem, Fermat rule, Lagrangian multipliers, Kuhn–Tucker multipliers

AMS subject classification. 49A22

1. Introduction.

Shape analysis and shape optimization (see [14], [15], [6]) deal with problems where the variables are no longer vectors of parameters or functions, but the shape of a geometric domain K contained in a subset E of R^p.

We adapt to the metric space $\mathcal{P}(E)$ of all nonempty closed subsets of a given compact set E, Graves' inverse function theorem, and, more generally, some constrained inverse function theorems (see [1], [11]), which allow us to solve locally a constrained problem of the form

$$\text{Find } K \in \mathcal{K} \subset \mathcal{P}(E) \quad \text{such that } F(K) = y$$

for any right-hand side y in the neighbourhood of a fixed

$$y_0 = F(K_0) \in F(\mathcal{K}),$$

where F is a map from $\mathcal{P}(E)$ into a finite-dimensional space Y.

To do this, we first need to define the gradient $\overset{\circ}{J}(K)$ of a functional J on $\mathcal{P}(E)$; the framework of shape derivatives introduced by many authors (see [3], [16], [5]) is used.

We also propose extensions of the contingent and Clarke's tangent cones, hereafter referred to as the velocity cones, that are defined for a family $\mathcal{K}$ of nonempty closed subsets of E and denoted by $\mathcal{V}_{\mathcal{K}}(K)$ and $\mathcal{U}_{\mathcal{K}}(K)$, respectively.

Once inverse function theorems have been established, they are applied to the calculus of velocity cones of subsets, especially those defined by inequality and equality constraints. This, if a Fermat rule is used, naturally gives rise to the shape Lagrangian and Kuhn–Tucker multipliers for the class of shape optimization problems of the form

$$J(\widehat{K}) = \inf J(K).$$
$$\begin{cases} A_i(K) \le 0, & i = 1, \dots, r, \\ B_j(K) = 0, & j = 1, \dots, s, \\ K \in \mathcal{P}(E), \end{cases}$$

These multipliers $(\widehat{\beta}, \widehat{\lambda})$ satisfy the following equality:

$$\overset{\circ}{J}(\hat{K}) = \sum_{i=1}^{r} \widehat{\lambda}_i \overset{\circ}{A}_i(\widehat{K}) + \sum_{j=1}^{s} \widehat{\beta}_j \overset{\circ}{B}_j(\widehat{K}).$$

* Received by the editors February 19, 1992; accepted for publication (in revised form) July 13, 1993.
† Ceremade, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, France.

2. Shape gradient. Let E be a subset of R^p. The definition of the directional derivative has to be adapted to the family $\mathcal{P}(E)$ of all nonempty closed subsets of E. To do this, the velocity method is used. From [5], it is known that velocity and transformation approaches are equivalent (see [7], [8], [9], [10]).

2.1. Velocity method and shape-directional derivative. Let us consider the following differential equation in R^p:

$$\left\{ \begin{array}{l} x'(t) = f(x(t)), \\ x(0) = x, \end{array} \right. \tag{1}$$

where $f : R^p \to R^p$.

The solution map is denoted by $T_f(t, \cdot)$, namely

$$T_f(t, x) := x(t),$$

where $x(\cdot)$ denotes a fixed solution of (1).

Given an initial domain K, we introduce the set

$$K_t = T_f(t, K) = \{T_f(t, x) \,|\, x \in K\}$$

that represents the transformed domain at time t.

Now, let us consider a subset E of R^p. We denote by $d_E(y)$ the distance between a point $y \in R^p$ and the set E, that is,

$$d_E(y) = \inf_{x \in E} \|x - y\|.$$

Nagumo's theorem (see Appendix 7.1) gives Assumption 1 on f and E, which guarantees that, for any domain $K \subset E$, the transformed set K_t stays in E for any $t \geq 0$.

Assumption 1. The following hold:

$$\left\{ \begin{array}{ll} \text{(i)} & f \text{ is Lipschitzian on } \overline{E}, \\ \text{(ii)} & \forall\, x \in \overline{E}, \qquad f(x) \in T_E(x) \cap -T_E(x), \end{array} \right. \tag{2}$$

where the contingent (or Bouligand) cone $T_E(x)$ is defined for an element x of $\overline{E}$ by

$$v \in T_E(x) \Longleftrightarrow \liminf_{h \to 0^+} \frac{d_E(x + hv)}{h} = 0.$$

The following definition provides the adaptation of directional derivatives for a shape functional.

DEFINITION 2.1 (Shape-directional derivative). *Consider a subset E of R^p and a map $J : \mathcal{P}(E) \to R$. Given a domain K in $\mathcal{P}(E)$, and $f : R^p \to R^p$ satisfying Assumption* 1, *we say that J has an Eulerian semiderivative at K in the direction f, if*

$$\lim_{t \to 0^+} \frac{J(T_f(t, K)) - J(K)}{t}$$

exists and is finite. It is denoted by $DJ(K, f)$.

2.2. Shape gradient. To define the shape gradient, the following spaces of functions are used.

- $\mathcal{D}^l(E, R^p)$ is the space of l-times continuously differentiable functions from R^p into R^p having a compact support in E. As usual, $\mathcal{D}^\infty(E, R^p)$ is written as $\mathcal{D}(E, R^p)$.
- $\mathcal{F}_E^l = \begin{cases} \mathcal{D}^l(E, R^p) & \text{if } 1 \le l \le +\infty, \\ \mathcal{D}^0(E, R^p) \cap \mathrm{Lip}(E, R^p) & \text{if } l = 0, \end{cases}$

where $\mathrm{Lip}(E, R^p)$ denotes the space of Lipschitzian maps from E into R^p.

DEFINITION 2.2 (Shape gradient). *Consider a subset E of R^p, a domain K in $\mathcal{P}(E)$, and a map $J : \mathcal{P}(E) \to R$. The functional J is said to be shape differentiable at K, if the map from $\mathcal{D}(E, R^p)$ into R defined by*

$$f \mapsto DJ(K, f)$$

is linear and continuous on $\mathcal{D}(E, R^p)$. This map defines a vector distribution that is denoted by $\overset{\circ}{J}(K)$ and called the shape gradient of J at K. Furthermore, if this map is continuous for the topology $\mathcal{F}_E^l$, we say that the shape gradient $\overset{\circ}{J}(K)$ is of order l.

Remark. Note that (2) holds true for any function f in $\mathcal{D}^1(E, R^p)$ because for any $x \in \partial E$, we have $f(x) = 0 \in T_E(x)$.

We now extend the notion of shape gradient to a map from $\mathcal{P}(E)$ into R^n.

DEFINITION 2.3. *Let us consider a subset E of R^p, a domain K in $\mathcal{P}(E)$, and a map $J : \mathcal{P}(E) \to R^n$. We denote by $J_i(K)$ the ith coordinate of the vector $J(K)$. The map J is said to be shape differentiable of order l at K if the functionals J_i are shape differentiable of order l at K. We denote by $\overset{\circ}{J}(K)$ the sequence $(\overset{\circ}{J}_1(K), \dots, \overset{\circ}{J}_n(K))$.*

3. Velocity cones. Now we introduce two "velocity" cones that can be regarded as extensions to the metric space $\mathcal{P}(E)$ of the Bouligand and the Clarke tangent cones of normed spaces.

3.1. Tangent cones. Because the contingent cone $T_K(x)$ has already been defined, we now introduce Clarke's cone $C_K(x)$ and mention some properties of both cones. For details and proofs about these tangent cones, we refer to [1].

The Clarke cone of K at x is the set

$$C_K(x) = \left\{ v \in X \mid \lim_{h \to 0+, x' \to_K x} \frac{d_K(x' + hv)}{h} = 0 \right\},$$

where $x' \to_K x$ means that the convergence is in the set K.

PROPOSITION 3.1. *Let K be a subset of a normed space X, and x an element of K. We always have*

(i) $C_K(x) \subset T_K(x)$,
(ii) *$T_K(x)$ is a closed cone,*
(iii) *$C_K(x)$ is a closed convex cone,*
(iv) $C_K(x) = \liminf_{x_n \to_K x} T_K(x_n)$.

We say that K is sleek at x_0 if it is a closed subset and if the cone-valued map

$$K \ni x \rightsquigarrow T_K(x)$$

is lower semicontinuous at x_0. The domain K is called sleek if it is sleek at every point of K.

It can be proved that convex subsets are sleek and that, when K is a sleek subset of a Banach space, the contingent and the Clarke tangent cones to K coincide and, consequently, are convex.

3.2. Velocity cones. To construct tangent cones analogues in $\mathcal{P}(E)$, a distance between domains is needed. The Hausdorff distance between two compact subsets L and K in $\mathcal{P}(E)$ is defined as follows:

$$d\!\!\!/(K, L) = \sup_{x\in K} \inf_{y\in L} \|x - y\| + \sup_{y\in L} \inf_{x\in K} \|x - y\|.$$

We recall that if E is a compact subset of R^p, then $(\mathcal{P}(E), d\!\!\!/)$ is a compact metric space.

Notation. Consider a family $\mathcal{K}$ of compact sets in E, and a compact set L in E. The distance between the domain L and the family $\mathcal{K}$ is defined by

$$d_{\mathcal{K}}(L) = \inf_{K\in\mathcal{K}} d\!\!\!/(L, K).$$

We propose the following extensions of the tangent cones.

DEFINITION 3.2 (Velocity cones). *Let E be a subset of R^p, K an element of a family $\mathcal{K}$ of compact subsets of E, and l an integer satisfying $0 \le l \le \infty$.*

- *The velocity cone of order l to $\mathcal{K}$ at K is the set*

$$\mathcal{V}^l_{\mathcal{K}}(K) = \left\{ f \in \mathcal{F}^l_E \mid \liminf_{h\to 0^+} \frac{d_{\mathcal{K}}(T_f(h, K))}{h} = 0 \right\}.$$

- *The velocity convex cone of order l to $\mathcal{K}$ at K is the set*

$$\mathcal{U}^l_{\mathcal{K}}(K) = \left\{ f \in \mathcal{F}^l_E \mid \lim_{h\to 0^+, K'\to_{\mathcal{K}} K} \frac{d_{\mathcal{K}}(T_f(h, K'))}{h} = 0 \right\}.$$

Remark. We can see at once the characterizations of these two cones as follows.

- $f \in \mathcal{V}^l_{\mathcal{K}}(K)$ if and only if there exists a sequence of reals $t_n \to 0^+$ and there exists a sequence $\widetilde{K_n} \in \mathcal{K}$ satisfying $(1/t_n) d\!\!\!/(\widetilde{K_n}, T_f(t_n, K)) \to 0$.
- $f \in \mathcal{U}^l_{\mathcal{K}}(K)$ if and only if for any sequence $K_n \to_{\mathcal{K}} K$ (in the sense of Hausdorff) and for any sequence of reals $t_n \to 0^+$, there exists a sequence $\widetilde{K_n} \in \mathcal{K}$ satisfying $(1/t_n) d\!\!\!/(\widetilde{K_n}, T_f(t_n, K_n)) \to 0$.

The definition below provides a characterization of regularity of a family $\mathcal{K}$.

DEFINITION 3.3 (Well shaped). *Let E be a subset of R^p, l an integer satisfying $0 \le l \le \infty$, and K be in a closed family $\mathcal{K}$ of compact subsets of E. We say that $\mathcal{K}$ is well shaped at K if*

$$\mathcal{U}^l_{\mathcal{K}}(K) \subset \liminf_{K_n\to_{\mathcal{K}} K} \mathcal{V}^l_{\mathcal{K}}(K_n) \quad \textit{for the } \mathcal{D}^l(E, R^p) \textit{ topology.}$$

3.3. Properties of the velocity cones. The velocity cones (§3.2) verify properties similar to those of tangent cones.

PROPOSITION 3.4. *Let E be a subset of R^p, and K be in a family $\mathcal{K}$ of compact subsets of E. For any l satisfying $0 \le l \le \infty$, we have*

(i) $\mathcal{U}^l_{\mathcal{K}}(K) \subset \mathcal{V}^l_{\mathcal{K}}(K)$,
(ii) $\mathcal{V}^l_{\mathcal{K}}(K)$ *is a cone,*
(iii) $U^l_{\mathcal{K}}(K)$ *is a convex cone,*
(iv) $\mathcal{V}^l_{\mathcal{K}}(K)$ *and* $\mathcal{U}^l_{\mathcal{K}}(K)$ *are closed.*

Proof. (i) This can be easily checked.

(ii) Because $T_f(t, \cdot) = T_{\lambda f}(t/\lambda, \cdot)$, it is clear that $\mathcal{V}^l_{\mathcal{K}}(K)$ and $\mathcal{U}^l_{\mathcal{K}}(K)$ are cones.

(iii) Let f_1 and f_2 be elements of $\mathcal{U}^l_{\mathcal{K}}(K)$ and consider sequences

$$\text{(3)} \qquad \begin{cases} t_n \to 0^+, \\ \mathcal{K} \ni K_n \to K. \end{cases}$$

By definition, there exists $L_n^1 \in \mathcal{K}$ such that

$$\frac{d\!\!\!/(L_n^1, T_{f_1}(t_n, K_n))}{t_n} \to 0. \tag{4}$$

The triangular inequality yields

$$d\!\!\!/(L_n^1, K) \leq d\!\!\!/(L_n^1, T_{f_1}(t_n, K_n)) + d\!\!\!/(T_{f_1}(t_n, K_n), K_n) + d\!\!\!/(K_n, K).$$

Because f_1 is bounded as a continuous function on its compact support, it is easy to check that for any x_n in K_n,

$$\|T_{f_1}(t_n, x_n) - x_n\| \leq t_n \|f_1\|_\infty,$$

where $\|f\|_\infty = \sup_{x \in R^p} \|f(x)\|$ is the uniform norm of f.

Hence

$$d\!\!\!/(T_{f_1}(t_n, K_n), K_n) \to 0,$$

which leads to

$$d\!\!\!/(L_n^1, K) \to 0.$$

Consequently, because f_2 belongs to $\mathcal{U}_{\mathcal{K}}^l(K)$, there exists $L_n^2 \in \mathcal{K}$ such that

$$\frac{d\!\!\!/(L_n^2, T_{f_2}(t_n, L_n^1))}{t_n} \to 0. \tag{5}$$

Furthermore, the triangular inequality leads to

$$\begin{aligned} d\!\!\!/(L_n^2, T_{f_1+f_2}(t_n, K_n)) \leq\ & d\!\!\!/(L_n^2 T_{f_2}(t_n, L_n^1)) \\ & + d\!\!\!/(T_{f_2}(t_n, L_n^1), T_{f_2}(t_n, T_{f_1}(t_n, K_n)) \\ & + d\!\!\!/(T_{f_2}(t_n, T_{f_1}(t_n, K_n)), T_{f_1+f_2}(t_n, K_n)). \end{aligned} \tag{6}$$

In view of (4), (5), and Lemmas 7.4 and 7.5 (below), we can conclude that

$$\frac{1}{t_n} d\!\!\!/(L_n^2, T_{f_1+f_2}(t_n, K_n)) \to 0^+,$$

which implies that

$$f_1 + f_2 \in \mathcal{U}_{\mathcal{K}}^l(K).$$

Therefore $\mathcal{U}_{\mathcal{K}}^l(K)$ is convex.

(iv) Let us consider a sequence of functions f_n of $\mathcal{V}_{\mathcal{K}}^l(K)$ converging for the $\mathcal{F}_E^l$ topology to a function f. Recall that $d_{\mathcal{K}}$ is a 1-Lipschitz function, that is,

$$|d_{\mathcal{K}}(K_1) - d_{\mathcal{K}}(K_2)| \leq d\!\!\!/(K_1, K_2), \qquad \forall\, K_1, K_2.$$

From Lemma 7.6 we obtain that for any n,

$$\liminf_{h \to 0^+} \frac{1}{h} d_{\mathcal{K}}(T_f(h, K)) \leq \liminf_{h \to 0^+} \frac{1}{h} d_{\mathcal{K}}(T_{f_n}(h, K)) + 2\|f_n - f\|_\infty.$$

Because the convergence in $\mathcal{F}_E^l$ implies the uniform convergence, if we let n tend to ∞, we conclude that f belongs to $\mathcal{V}_{\mathcal{K}}^l(K)$. This implies that $\mathcal{V}_{\mathcal{K}}^l(K)$ is closed.

The same arguments are used to prove that $\mathcal{U}^l_K(K)$ is closed. □

PROPOSITION 3.5. *Let E be a compact subset of R^p and K be in $\mathcal{P}(E)$. We have*

$$\mathcal{U}^l_{\mathcal{P}(E)}(K) = \mathcal{V}^l_{\mathcal{P}(E)}(K) = \mathcal{F}^l_E.$$

Consequently, $\mathcal{P}(E)$ is well shaped at any K in $\mathcal{P}(E)$.

Proof. We already know that

$$\mathcal{U}^l_{\mathcal{P}(E)}(K) \subset \mathcal{V}^l_{\mathcal{P}(E)}(K) \subset \mathcal{F}^l_E.$$

It remains to check that

$$\mathcal{U}^l_{\mathcal{P}(E)}(K) \supset \mathcal{V}^l_{\mathcal{P}(E)}(K) \supset \mathcal{F}^l_E.$$

To do this, consider

$$\text{(7)} \qquad \begin{cases} f \in \mathcal{F}^l_E, \\ t_n \to 0^+, \\ K_n \to K \quad \text{in } \mathcal{P}(E). \end{cases}$$

Because $f \in \mathcal{F}^l_E$, Nagumo's theorem (Theorem 7.1) ensures that

$$\widetilde{K_n} = T_f(t_n, K_n) \subset E.$$

Consequently

$$\frac{d\!\!l(T_f(t_n, K_n), \widetilde{K_n})}{t_n} = 0.$$

Therefore, f belongs to $\mathcal{U}^l_{\mathcal{P}(E)}(K)$. □

3.4. Relations with tangent cones. Here we consider the case of a singleton $\{x\}$, and K is identified with the set $\{\{x\}\}_{x\in K} \subset \mathcal{P}(E)$. The following vector distribution $f \to f(x)$ is denoted δ_x.

PROPOSITION 3.6. *Let K be a closed subset of a set E in R^p, and let x be in K. We have*

$$\text{(8)} \qquad \begin{aligned} &\text{(i)} \quad f \in \mathcal{V}^l_K(\{x\}) \Rightarrow \delta_x f \in T_K(x), \\ &\text{(ii)} \quad f \in \mathcal{U}^l_K(\{x\}) \Rightarrow \delta_x f \in C_K(x). \end{aligned}$$

Proof. (i) Let us assume that the assertion $f \in \mathcal{V}^l_K(\{x\})$ is satisfied; thus

$$\text{(9)} \qquad \exists x_n \in K \quad \text{such that } \frac{1}{t_n} d\!\!l(x_n, T_f(t_n, x)) \to 0^+.$$

Hence

$$\frac{1}{t_n} d_K(T_f(t_n, x)) \to 0^+.$$

Moreover, we have

$$\text{(10)} \qquad \begin{aligned} |d_K(x + t_n f(x)) - d_K(T_f(t_n, x))| &\le \|x + t_n f(x) - T_f(t_n, x)\| \\ &\le t_n \varepsilon(t_n). \end{aligned}$$

Consequently,

$$\liminf_{t\to 0^+} \frac{1}{t} d_K(x + tf(x)) = 0,$$

which means that the assertion $\delta_x f \in T_K(x)$ is satisfied.

(ii) Let us assume that $f \in \mathcal{U}_K^l(\{x\})$. Consider a sequence x_n converging to x in K and a sequence of reals t_n converging to 0^+. This yields

$$\exists y_n \in K \quad \text{such that } \frac{1}{t_n} d(y_n, T_f(t_n, x_n)) \to 0^+. \tag{11}$$

Hence

$$\frac{1}{t_n} d_K(T_f(t_n, x_n)) \to 0^+.$$

Consequently, using (10) again, we obtain

$$\lim_{\substack{y \to_K x \\ t\to 0^+}} \frac{1}{t} d_K(y + tf(x)) = 0.$$

This proves that $\delta_x f \in C_K(x)$. □

4. Constrained inverse function theorem. Inverse function theorems are known to be efficient tools in analysis of a wide range of problems. Let us recall that Graves' theorem states that, if a continuously differentiably map $f : X \to Y$ between two Banach spaces has a surjective derivative $f'(\hat{x})$ at some $\hat{x} \in X$, then the inverse image $f^{-1}(\cdot)$ enjoys a Lipschitzian behaviour around $f(\hat{x})$.

Here, the constrained inverse function theorem and, consequently, Graves' inverse function theorem are extended to shape maps. More generally, the theorems are stated for maps where the variables are both vectors and domains.

4.1. Pseudo-Lipschitz set-valued map. First, let us make how we can characterize the Lipschitzian behaviour of a set-valued map precise.

DEFINITION 4.1 (Pseudo-Lipschitz). *When X and Y are metric spaces, we say that the set-valued map $F : X \rightsquigarrow Y$ is pseudo-Lipschitz around $(x, y) \in \mathrm{Graph}(F)$, if there exists a positive constant l and neighbourhoods $V_x \subset \mathrm{Dom}(F)$ of x and V_y of y, such that*

$$\forall x_1, x_2 \in V_x, \qquad \forall y_1 \in F(x_1) \cap V_y,$$

$$\exists y_2 \in F(x_2) \quad \textit{such that } d_Y(y_1, y_2) \le l d_X(x_1, x_2).$$

4.2. Constrained inverse function theorem. In what follows, we use the following notations.

- Consider two maps $F : \mathcal{P}(E) \to Y$ and $f : X \to Y$ where X, Y are normed spaces. The map $F \oplus f$ is defined by

$$(F \oplus f)(K, x) = F(K) + f(x),$$

where K belongs to $\mathcal{P}(E)$ and x belongs to X.

- The unit ball for the Euclidian norm on $Y = R^n$ is denoted B_Y and the unit ball of the space of bounded functions on R^p for the uniform norm is denoted by B_∞.

THEOREM 4.2 (Constrained inverse shape function). *Let X be a Banach space; Y, a finite-dimensional space; E, a compact set of R^p; $\mathcal{K}$, a closed subset of $\mathcal{P}(E)$; and K_0 be in $\mathcal{K}$. Let M be a closed subset of X and x_0 an element of M. Let us consider a shape function $F : \mathcal{P}(E) \to Y$ that is continuous (in the Hausdorff metric) on $\mathcal{K}$ and a continuous function $f : X \to Y$. Let us assume that the map F, in a neighbourhood of K_0, is Lipschitzian (in the Hausdorff metric) and shape differentiable of order l $(0 \leq l \leq \infty)$ and that the map f, in a neighbourhood of x_0, is differentiable. Furthermore, let us assume that the following transversality condition holds true:*

$$(12)\qquad \begin{cases} \exists c, c', \eta > 0, \qquad \exists \alpha \in [0,1[\quad \textit{satisfying}, \\ \forall K \in \mathcal{K} \cap B(K_0, \eta), \qquad \forall x \in M \cap B(x_0, \eta), \\ B_Y \subset \overset{\circ}{F}(K)(\mathcal{V}^l_{\mathcal{K}}(K) \cap cB_\infty) + f'(x)(C_M(x) \cap c'B_X) + \alpha B_Y. \end{cases}$$

Then the set-valued map $y \rightsquigarrow (F \oplus f)^{-1}(y) \cap (\mathcal{K} \times M)$ is pseudo-Lipschitz in a neighbourhood of $(F(K_0) + f(x_0), (K_0, x_0))$.

Remark. The above theorem still holds true with $T^\flat_M(x)$[1] instead of $C_M(x)$.

If f is taken to be 0, then the preceding theorem deals only with map of sets and becomes Theorem 4.3.

THEOREM 4.3. *Let Y be a finite-dimensional space; E, a compact set of R^p; $\mathcal{K}$, a closed subset of $\mathcal{P}(E)$; and K_0 be in $\mathcal{K}$. Let us consider a shape function $F : \mathcal{P}(E) \to Y$ continuous (in the Hausdorff metric) on $\mathcal{K}$. Let us assume that the map F, in a neighbourhood of K_0, is Lipschitzian (in the Hausdorff metric) and shape differentiable of order l $(0 \leq l \leq \infty)$. Furthermore, let us assume that*

$$(13)\qquad \begin{cases} \exists c, \eta > 0, \qquad \exists \alpha \in [0,1[\quad \textit{satisfying}, \\ \forall K \in \mathcal{K} \cap B(K_0, \eta), \qquad B_Y \subset \overset{\circ}{F}(K)(\mathcal{V}^l_{\mathcal{K}}(K) \cap cB_\infty) + \alpha B_Y \end{cases}$$

Then the set-valued map $y \rightsquigarrow F^{-1}(y) \cap \mathcal{K}$ is pseudo-Lipschitz in a neighbourhood of $(F(K_0), K_0)$.

Proof of Theorem 4.2. Let us consider

$$(14)\qquad \begin{cases} K \in \mathcal{K} & \text{such that } d(K, K_0) < \dfrac{\eta}{3}, \\ x \in M & \text{such that } \|x - x_0\| < \dfrac{\eta}{3}, \\ \rho > 0 & \text{such that } \dfrac{3\rho}{\eta} < \dfrac{1-\alpha}{2c+c'}, \\ y \in Y & \text{such that } \|y - F(K) - f(x)\| < \rho, \\ \varepsilon > 0 & \text{such that } \dfrac{3\rho}{\eta} < \varepsilon < \dfrac{1-\alpha}{2c+c'}, \end{cases}$$

and introduce the function

$$\mathcal{K} \times M \ni (K, x) \rightsquigarrow V(K, x) = \|y - F(K) - f(x)\|.$$

[1] The adjacent cone to K at x is the set defined by

$$T^\flat_K(x) = \left\{ v \in X \mid \lim_{h \to 0+} \frac{d_K(x + hv)}{h} = 0 \right\}.$$

Because E is a compact set of R^p, E is a complete metric space, which implies that $\mathcal{P}(E)$ is a complete metric space for the Hausdorff distance. Furthermore, $\mathcal{K}$ being closed, $\mathcal{K}$ is complete.

On the other hand, M is complete as a closed subset of the Banach space X. Consequently, the product set $\mathcal{K} \times M$ is complete for the distance $d((K,x),(H,y) = \|x-y\| + d\!\!\!l(K,H)$.

Because V is positive and lower semicontinuous, we can apply Ekeland's theorem[2] to V, that is,

(15)
$$\begin{cases} \exists(\widetilde{K},\tilde{x}) \in \mathcal{K}\times M, & V(\widetilde{K},\tilde{x}) + \varepsilon d\!\!\!l(\widetilde{K},K) + \varepsilon\|\tilde{x}-x\| \le V(K,x), \\ \forall (H,y) \neq (\widetilde{K},\tilde{x}), & V(\widetilde{K},\tilde{x}) < \varepsilon d\!\!\!l(H,\widetilde{K}) + \varepsilon\|\tilde{x}-y\|, (H,y) \in \mathcal{K}\times M + V(H,y). \end{cases}$$

It can be easily deduced that $d\!\!\!l(K_0,\widetilde{K}) \le 2\eta/3$, and that $\|x_0-\tilde{x}\| \le 2\eta/3$, which leads to

(16)
$$\begin{cases} \widetilde{K} \in B(K_0,\eta) \cap \mathcal{K}, \\ \tilde{x} \in B(x_0,\eta) \cap M. \end{cases}$$

Consequently, using the transversality condition (12), we have

$$B_Y \subset \overset{\circ}{F}(\widetilde{K})(\mathcal{V}^l_{\mathcal{K}}(\widetilde{K}) \cap cB_\infty) + f'(\tilde{x})(C_m(\tilde{x}) \cap c'B_X) + \alpha B_Y.$$

Therefore we have

(17)
$$y - F(\widetilde{K}) - f(\tilde{x}) = \overset{\circ}{F}(\widetilde{K})g + f'(\tilde{x})v + w,$$

where

(18)
$$\begin{cases} g \in \mathcal{V}^l_{\mathcal{K}}(\widetilde{K}), \qquad v \in C_M(\tilde{x}), \\ \|g\|_\infty \le c\|y - F(\widetilde{K}) - f(\tilde{x})\|, \\ \|w\| \le \alpha\|y - F(\widetilde{K}) - f(\tilde{x})\|, \\ \|v\| \le c'\|y - F(\widetilde{K}) - f(\tilde{x})\|. \end{cases}$$

Using the definition of $\mathcal{V}^l_{\mathcal{K}}(\widetilde{K})$, we know that there exist $t_n \to 0^+$ and $\widetilde{K_n} \in \mathcal{K}$ such that

(19)
$$\frac{1}{t_n} d\!\!\!l(\widetilde{K_n}, K_n) \to 0^+, \quad \text{where } K_n = T_g(t_n,\widetilde{K}).$$

Similarly, using the definition of $C_M(\tilde{x})$, we know that there exists a sequence $\widetilde{\varepsilon_n} \to 0$ such that

$$\widetilde{x_n} = \tilde{x} + t_n(v + \widetilde{\varepsilon_n}) \in M.$$

[2] See, for instance, [1] for details and proof.

EKELAND'S THEOREM. *Consider $V : E \to R \cup \{+\infty\}$ a strict, lower semicontinuous and positive function, where (E,d) is a complete metric space; then, given $\varepsilon > 0$ and $x_0 \in \mathrm{Dom}V$,*

$$\begin{cases} \exists \tilde{x} \in E & V(\tilde{x}) + \varepsilon d(\tilde{x},x_0) \le V(x_0), \\ \forall x \neq \tilde{x} & V(\tilde{x}) < \varepsilon d(x,\tilde{x}) + V(x). \end{cases}$$

From (15), we obtain that

$$\|y - F(\widetilde{K_n}) - f(\widetilde{x_n})\| + \varepsilon\|\widetilde{x_n} - \tilde{x}\| + \varepsilon d\!\!\!^{-}(\widetilde{K_n}, \widetilde{K}) \geq \|y - F(\widetilde{K}) - f(\tilde{x})\|. \tag{20}$$

Because F is Lipschitzian in a neighbourhood of K_0, we have

$$\|F(\widetilde{K_n}) - F(K_n)\| \leq k\, d\!\!\!^{-}(\widetilde{K_n}, K_n). \tag{21}$$

Hence, by virtue of (19),

$$F(\widetilde{K_n}) - F(K_n) = t_n\widehat{\varepsilon_n} \quad \text{with } \|\hat{\varepsilon}_n\| \to 0^+.$$

Using (21) and the definition of $\overset{\circ}{F}(\widetilde{K})$, we see that

$$F(\widetilde{K_n}) = F(\widetilde{K}) + t_n(\overset{\circ}{F}(\widetilde{K})g + \varepsilon_n^1).$$

Because f is differentiable around $\tilde{x}$, we can write

$$f(\widetilde{x_n}) = f(\tilde{x}) + t_n(f'(\tilde{x})v + \varepsilon_n^2).$$

We thus obtain

$$y - F(\widetilde{K_n}) - f(\widetilde{x_n}) = y - F(\widetilde{K}) - f(\tilde{x}) - t_n(\overset{\circ}{F}(\widetilde{K})g + f'(\tilde{x})v + \varepsilon_n^1 + \varepsilon_n^2).$$

In view of (17), we can write

$$y - F(\widetilde{K_n}) - f(\widetilde{x_n}) = (y - F(\widetilde{K}) - f(\tilde{x}))(1 - t_n) - t_n(\varepsilon_n^1 + \varepsilon_n^2 - w).$$

From (20), we obtain

$$\|y - F(\widetilde{K}) - f(\tilde{x})\| \leq \|\varepsilon_n\| + \|w\| + \frac{\varepsilon}{t_n}(d\!\!\!^{-}(\widetilde{K_n}, \widetilde{K}) + \|\widetilde{x_n} - \tilde{x}\|). \tag{22}$$

Moreover, we have

$$\begin{aligned} \sup_{z_n \in T_g(t_n, \tilde{K})} \inf_{z \in K} \|z_n - z\| &\leq \sup_{z \in \tilde{K}} \|T_g(t_n, z) - z\| \\ &\leq t_n\|g\|_\infty. \end{aligned} \tag{23}$$

Thus

$$\frac{d\!\!\!^{-}(K_n, \widetilde{K})}{t_n} \leq 2\|g\|_\infty. \tag{24}$$

Consequently, from (19) and the triangular inequality, we deduce that

$$\frac{d\!\!\!^{-}(\widetilde{K_n}, \widetilde{K})}{t_n} \leq 2\|g\|_\infty + \varepsilon_n. \tag{25}$$

Furthermore,

$$\frac{\|\widetilde{x_n} - \tilde{x}\|}{t_n} \leq \|v\| + \widetilde{\varepsilon_n}.$$

Consequently, using (18) and letting n tend to ∞, (22) becomes

$$\|y - F(\widetilde{K}) - f(\tilde{x})\| \leq (\varepsilon(2c + c') + \alpha)\|y - F(\widetilde{K}) - f(\tilde{x})\|.$$

Because $\varepsilon(2c+c')+\alpha<1$, it can be deduced that $\|y-F(\widetilde{K})-f(\tilde{x})\|=0$. Therefore

$$y=(F\oplus f)(\widetilde{K},\tilde{x}) \quad \text{with} \begin{cases} \widetilde{K}\in\mathcal{K}, \qquad \tilde{x}\in M, \\ d((K,x),(\widetilde{K},\tilde{x}))\le \dfrac{1}{\varepsilon}\|y-(F\oplus f)(K,x)\|. \end{cases}$$

This proves the theorem. □

Now we give a version of Theorem 4.2 using the convex cone $\mathcal{U}_{\mathcal{K}}(K)$ and stronger regularity assumptions on the maps and the family $\mathcal{K}$.

THEOREM 4.4. *Let X be a Banach space; Y, a finite-dimensional space; E, a compact set of R^p; and K_0 be in a subset $\mathcal{K}$ of $\mathcal{P}(E)$. Let x_0 be an element of a subset M of X. Consider a shape function $F:\mathcal{P}(E)\to Y$ that is continuous on $\mathcal{K}$ and a continuous function $f:X\to Y$. We assume that M is sleek at x_0, that $\mathcal{K}$ is well shaped at K_0, that the map F, in a neighbourhood of K_0, is Lipschitzian (in the Hausdorff metric) and continuously (weakly) shape differentiable of order l $(1\le l<\infty)$ and that the map f, in a neighbourhood of x_0, is continuously differentiable. Let us assume that*

$$\overset{\circ}{F}(K_0)\mathcal{U}^l_{\mathcal{K}}(K_0)+f'(x_0)C_M(x_0)=Y. \tag{26}$$

Then the set-valued map $y\rightsquigarrow(F\oplus f)^{-1}(y)\cap(\mathcal{K}\times M)$ is pseudo-Lipschitz in a neighbourhood of $((F\oplus f)(K_0,x_0),(K_0,x_0))$.

If f is taken to be 0, then the previous theorem deals only with maps of sets and becomes Theorem 4.5.

THEOREM 4.5. *Let Y be a finite-dimensional space; E, a compact set of R^p; $\mathcal{K}$, a closed subset of $\mathcal{P}(E)$; and let K_0 be in $\mathcal{K}$. Let us consider a shape function $F:\mathcal{P}(E)\to Y$ that is continuous on $\mathcal{K}$. Assume that the family $\mathcal{K}$ is well shaped at K_0, that the map F, in a neighbourhood of K_0, is Lipschitzian (in the Hausdorff metric) and continuously (weakly) shape differentiable of order l $(1\le l<\infty)$. Furthermore, let us assume that the following surjectivity assumption holds true:*

$$\overset{\circ}{F}(K_0)\mathcal{U}^l_{\mathcal{K}}(K_0)=Y. \tag{27}$$

Then the set-valued map $y\rightsquigarrow F^{-1}(y)\cap\mathcal{K}$ is pseudo-Lipschitz in a neighbourhood of $(F(K_0),K_0)$.

Proof of Theorem 4.4. We establish that the transversality requirement of Theorem 4.2 is satisfied. Let y_i be an element of S_Y, the unit sphere of Y. Because E is compact and $l<\infty$, $\mathcal{D}^l_E$ is a Banach space for the norm

$$\|f\|=\sum_{i=0}^{l}\|D^if\|_\infty.$$

Because $\mathcal{U}^l_{\mathcal{K}}(K_0)$ and $C_M(x_0)$ are closed $(l\ge1)$ and convex, we can apply the Robinson–Ursescu theorem[3] to the linear and continuous function

$$\mathcal{U}^l_{\mathcal{K}}(K_0)\times C_M(x_0)\ni(g,v)\mapsto\overset{\circ}{F}(K_0)g+f'(x_0)v.$$

[3] We refer, for instance, to [1] for the details and complete proof.

THE ROBINSON–URSESCU THEOREM. *Let X, Y be Banach spaces. Let us consider a continuous linear operator $A\in\mathcal{L}(X,Y)$ and a closed convex subset K of X. Suppose that Ax_0 belongs to the interior of the image of A. Then there exist positive constants l and β such that for any $y\in y_0+\beta B$, there exists a solution $x\in K$ to the equation $y=Ax$ satisfying*

$$\|x-x_0\|\le l\|y-Ax_0\|.$$

Hence there exists $c > 0$, g_0^i in $\mathcal{U}_{\mathcal{K}}(K_0)$ and v_0^i in $C_M(x_0)$ such that

$$\begin{cases} \overset{\circ}{F}(K_0)g_0^i + f'(x_0)v_0^i = y_i, \\ \|g_0^i\| + \|v_0^i\| \leq c\|y_i\| = c. \end{cases} \tag{28}$$

Let us consider $\varepsilon > 0$. From Proposition 3.1(iv), we know that, for every y^i, there exists $\eta_i > 0$ such that, for all $x \in B(x_0, \eta_i)$, we can find $v^i \in T_M(x)$ satisfying

$$\|v^i - v_0^i\| \leq c\varepsilon. \tag{29}$$

Because the family $\mathcal{K}$ is well shaped at K_0, we know that, for all y^i, there exists $\eta_i' > 0$ such that, for all $K \in B(K_0, \eta_i')$, there exists $g_i \in \mathcal{V}_{\mathcal{K}}(K)$ satisfying

$$\|g_i - g_i^0\|_\infty \leq c\varepsilon.$$

On the other hand, the continuity of $\overset{\circ}{F}$ and f' implies the existence of reals $\eta_0, \eta_0' > 0$ such that, for all $(K, x) \in B(K_0, \eta_0) \times B(x_0, \eta_0')$,

$$\begin{cases} \|\overset{\circ}{F}(K)\varphi - \overset{\circ}{F}(K_0)\varphi\| < \varepsilon \quad \text{for all } \varphi \in \mathcal{D}^l(E, R^p), \\ \|f'(x) - f'(x_0)\| < \varepsilon. \end{cases} \tag{30}$$

Let us now use the fact that S_Y is compact to claim that there exist p balls $B(y_i, \varepsilon)$ such that

$$S_Y \subset \bigcup_{i=1}^{p} B(y_i, \varepsilon). \tag{31}$$

Setting $\eta = \min(\min_{i=0,\dots,p} \eta_i, \min_{i=0,\dots,p} \eta_i')$, we have, for all $y \in S_Y$, for all $K \in B(K_0, \eta)$, and for all $x \in B(x_0, \eta)$,

$$\begin{cases} y &= \overset{\circ}{F}(K)g^i + f'(x)v^i + w_i, \\ w_i &= \overset{\circ}{F}(K_0)(g_0^i - g^i) + (\overset{\circ}{F}(K_0)g^i - \overset{\circ}{F}(K)g^i) \\ &\quad + f'(x_0)(v_0^i - v^i) + (f'(x_0) - f'(x))v^i + y - y_i. \end{cases} \tag{32}$$

In view of (30) and (31), we obtain

$$\|w_i\| \leq \varepsilon(2 + c\|f'(x_0)\| + c\|\overset{\circ}{F}(K_0)\| + c(1 + \varepsilon)) = \alpha. \tag{33}$$

Choosing ε such that $\alpha \in [0, 1[$ and setting $c' = c(1 + \varepsilon)$, we can conclude that

$$\exists c', \eta > 0, \qquad \exists \alpha \in [0, 1[\quad \text{satisfying}$$

$$\forall K \in \mathcal{K} \cap B(K_0, \eta), \qquad \forall x \in M \cap B(x_0, \eta), \tag{34}$$

$$B_Y \subset \overset{\circ}{F}(K)(\mathcal{V}_{\mathcal{K}}(K) \cap c'B_\infty) + f'(x)(C_M(x) \cap c'B_X) + \alpha B_Y.$$

This completes the proof of Theorem 4.5. □

4.3. Inverse function theorem. Using Theorem 4.5 and Proposition 3.5, we derive the following adaptation of Graves' theorem.

THEOREM 4.6 (Inverse shape function). *Let Y be a finite-dimensional space; E, a compact set of R^p; and K_0 be in $\mathcal{P}(E)$. Let us consider a continuous shape function $F : \mathcal{P}(E) \to Y$. Assume that the map F, in a neighbourhood of K_0, is Lipschitzian (in the Hausdorff metric) and continuously shape differentiable of order l $(1 \leq l < \infty)$, and that*

$$\overset{\circ}{F}(K_0) \quad \text{is surjective.} \tag{35}$$

Then the set-valued map $y \rightsquigarrow F^{-1}(y)$ is pseudo-Lipschitz in a neighbourhood of $(F(K_0), K_0)$.

5. Calculus of velocity cones. From Theorem 4.5, we can obtain a simple expression of the velocity cone when $\mathcal{K}$ is defined by equality and inequality constraints

$$\begin{cases} K \in \mathcal{P}(E), \\ A_i(K) \leq 0, \qquad i = 1, \ldots, r, \\ B_j(K) = 0, \qquad j = 1, \ldots, s, \end{cases} \tag{36}$$

where A_i, B_j are shape functionals.

5.1. Calculus of velocity cones of inverse image. More generally, we study the case of a family $\mathcal{K}$ defined as an inverse image of a set M in a finite-dimensional space Y, that is,

$$\mathcal{K} = F^{-1}(M).$$

PROPOSITION 5.1. *Let Y be a finite-dimensional space; E, a compact set of R^p; $\mathcal{L}$ a well shaped subset of $\mathcal{P}(E)$; and M, a sleek subset of Y. Let us consider a shape function $F : \mathcal{P}(E) \to Y$ that is continuous on $\mathcal{L}$. Assume that the map F, in a neighbourhood of $K \in \mathcal{L} \cap F^{-1}(M)$, is Lipschitzian (in the Hausdorff metric) and continuously shape differentiable of order l $(1 \leq l < \infty)$ and that the following transversality condition holds true:*

$$\overset{\circ}{F}(K)\mathcal{U}^l_{\mathcal{L}}(K) - C_M(F(K)) = Y. \tag{37}$$

Then

$$\mathcal{V}^l_{\mathcal{L}\cap F^{-1}(M)}(K) = \mathcal{V}^l_{\mathcal{L}}(K) \cap \overset{\circ}{F}(K)^{-1} T_M(F(K)).$$

If $\mathcal{L}$ is equal to the whole space $\mathcal{P}(E)$, we obtain Corollary 5.2.

COROLLARY 5.2. *Let E be a compact set of R^p, and M be a sleek subset of a finite-dimensional space Y. Let us consider a continuous map $F : \mathcal{P}(E) \to Y$. Consider a subset $K \in F^{-1}(M)$ and assume that the map F, in a neighbourhood of K, is Lipschitzian (in the Hausdorff metric) and continuously shape differentiable of order l $(1 \leq l < \infty)$. If the transversality condition*

$$\overset{\circ}{F}(K)\mathcal{D}^l(E, R^p) - C_M(F(K)) = Y \tag{38}$$

is satisfied, then

$$\mathcal{V}^l_{F^{-1}(M)}(K) = \overset{\circ}{F}(K)^{-1} T_M(F(K)).$$

Proof of 5.1. Let f be an element of $\mathcal{V}^l_{\mathcal{L}\cap F^{-1}(M)}(K)$. By definition, there exist $t_n \to 0^+$ and $K_n \in \mathcal{L} \cap F^{-1}(M)$ such that

$$\frac{1}{t_n} d\!\!\!l(K_n, T_f(t_n, K)) \to 0^+.$$

Obviously $f \in \mathcal{V}^l_{\mathcal{L}}(K)$.

Because F is shape differentiable of order l and Lipschitzian in a neighbourhood of K, we have

$$F(K_n) = F(K) + t_n(\overset{\circ}{F}(K)f + \varepsilon(t_n)). \tag{39}$$

Because $F(K_n)$ belongs to M, we have, by definition of the contingent cone,

$$\overset{\circ}{F}(K)f \in T_M(F(K)),$$

which states that

$$f \in \overset{\circ}{F}(K)^{-1}(T_M(F(K)).$$

Consequently,

$$\mathcal{V}^l_{\mathcal{L}\cap F^{-1}(M)}(K) \subset \mathcal{V}^l_{\mathcal{L}}(K) \cap \overset{\circ}{F}(K)^{-1} T_M(F(K)).$$

Conversely, let us consider $f \in \mathcal{V}^l_{\mathcal{L}}(K)$ such that $\overset{\circ}{F}(K)f \in T_M(F(K))$. Then, because M is sleek, there exist sequences $t_n \to 0^+$, $v_n \to \overset{\circ}{F}(K)f$, and $K_n \in \mathcal{L}$ such that

$$\begin{cases} \dfrac{1}{t_n} d\!\!\!l(K_n, T_f(t_n, K)) \to 0, \\ y_n = F(K) + t_n v_n \in M. \end{cases} \tag{40}$$

Let us introduce the map $F \ominus 1$ as follows:

$$\begin{aligned} F \ominus 1 : \mathcal{P}(E) \times Y &\to Y \\ (K, Y) &\mapsto F(K) - y. \end{aligned} \tag{41}$$

Theorem 4.4 and the assumption

$$\overset{\circ}{F}(K)\mathcal{U}^l_{\mathcal{L}}(K) - C_M(F(K)) = Y \tag{42}$$

imply that the map $\alpha \rightsquigarrow (F \ominus 1)^{-1}(\alpha) \cap (\mathcal{L} \times M)$ is pseudo-Lipschitz around $((K, F(K)), 0)$.

Because F is continuous, we have

$$(F \ominus 1)(K_n, y_n) = F(K_n) - F(K) - t_n(\overset{\circ}{F}(K)f + \varepsilon(t_n)) \to 0. \tag{43}$$

This yields

$$\exists \widehat{K_n} \in \mathcal{L}, \qquad \exists \widehat{y_n} \in M \quad \text{such that } (F \ominus 1)(\widehat{K_n}, \widehat{y_n}) = 0 \tag{44}$$

and

$$d\!\!\!l(\widehat{K_n}, K_n) \leq k \| F(K_n) - y_n - (F(\widehat{K_n}) - \widehat{y_n}) \|. \tag{45}$$

Using (39), it can be deduced that

$$\mathrm{dl}(\widehat{K_n}, K_n) \leq k t_n \|\varepsilon_n\|. \tag{46}$$

In view of (40), we obtain

$$\frac{1}{t_n} \mathrm{dl}(\widehat{K_n}, T_f(t_n, K)) \to 0. \tag{47}$$

Because $\widehat{K_n} \in \mathcal{L} \cap F^{-1}(M)$, we can conclude that $f \in \mathcal{V}^l_{\mathcal{L}\cap F^{-1}(M)}(K)$. □

5.2. The case of inequality and equality constraints. Let us consider the case of inequality and equality constraints. Namely,

$$\mathcal{K} = \left\{ K \in \mathcal{P}(E) \;\middle|\; \begin{array}{l} A_i(K) \geq 0, \ \forall i \in \{1, \ldots, r\} \\ B_j(K) = 0, \ \forall j \in \{1, \ldots, s\} \end{array} \right\}, \tag{48}$$

where B_i and A_j are functions from $\mathcal{P}(E)$ into R.

The maps A and B represent the vectors of shape functionals $(A_1, \ldots, A_r)$ and $(B_1, \ldots, B_s)$, respectively. We denote by $I(K) = \{i \mid A_i(K) = 0\}$ the set of active constraints.

PROPOSITION 5.3. *Let E be a compact subset of R^p. Let us assume that the shape functions A_i and B_j are continuous on $\mathcal{P}(E)$, Lipschitzian in the Hausdorff metric, and continuously shape differentiable of order l $(1 \leq l < \infty)$ in a neighbourhood of $K \in \mathcal{K}$ (defined in* (48)*) and that*

$$\left\{ \begin{array}{ll} \text{(i)} \ \overset{\circ}{B}(K)\mathcal{D}^l(E, R^p) = R^s, & \\ \text{(ii)} \ \exists g_0 \in \mathcal{D}^l(E, R^p) \quad \textit{such that} & \overset{\circ}{B}(K)(g_0) = 0, \\ & \overset{\circ}{A}_i(K)(g_0) > 0, \qquad \forall i \in I(K). \end{array} \right. \tag{49}$$

Then

$$\mathcal{V}^l_{\mathcal{K}}(K) = \{f \in \mathcal{F}^l_E \mid \overset{\circ}{B}(K)f = 0 \textit{ and } \overset{\circ}{A}_i(K)f \geq 0, \ \forall i \in I(K)\}.$$

Proof. It is only necessary to check that the transversality assumption of Corollary 5.2 holds true with $Y = R^{r+s}$, $F = (A, B)$, and $M = M^r_+ \times \{0\}^s$. Indeed, take $(y, z) \in R^r \times R^s$. From the first assumption, there exists $g \in \mathcal{D}^l(E, R^s)$ such that $\overset{\circ}{B}(K)g = z$. Fix

$$\alpha = \min_{i \in I(K)} \overset{\circ}{A}_i(K)g_0$$

and

$$\lambda = \max(0, y_1 - \overset{\circ}{A}_1(K)g, \ldots, y_r - \overset{\circ}{A}_r(K)g)/\alpha.$$

Let us set

$$v_i = \overset{\circ}{A}_i(K)(g + \lambda g_0) - y_i.$$

By construction, $v_i \neq 0$ for any $i \in I(K)$, so that $v = (v_1, \ldots, v_r)$ belongs to the tangent cone $C_{R^r}(A(K))$. Therefore $f = \lambda g_0 + g$ belongs to $\mathcal{D}^l(E, R^p)$ and satisfies

$$(y, z) = \overset{\circ}{F}(K)f - (v, 0).$$

Therefore

$$Y = \overset{\circ}{F}(K)\mathcal{D}^l(E, R^p) - C_M(F(K)),$$

and we can apply Corollary 5.2. □

5.3. Example. Let E be a compact set of R^p. Let $\mathcal{K}$ be the family of all closed connected subsets K of $\mathrm{Int}(E)$ satisfying the ε-cone property[4], with volume $V(K)$ equal to α, where

$$V(K) = \int_K dx.$$

We observe the following.

- V is shape differentiable of order 1 on any $K \in \mathcal{K}$ and for any $f \in \mathcal{D}^1_E$ and we have

$$\overset{\circ}{V}(K)f = \int_K \mathrm{div}(f(x))dx.$$

- It can be checked that V is continuous (Hausdorff) on $\mathcal{K}$, because it is continuous on $\mathcal{E} = \{K \subset E \text{ satisfying the } \varepsilon\text{-cone property}\}$.
- From the Steiner formula (see, for instance, [13]), we obtain, for any convex set $K \in \mathcal{K}$,

$$V(K \oplus B(0,r)) = V(K) + rM + r^2\varepsilon(r).$$

Thus, for any sequence K_n such that $d\!l(K_n, K) \to 0$, we have

$$V(K_n) \leq V(K \oplus B(0, d\!l(K_n, K))) = V(K) + M'd(K_n, K),$$

which implies that V is Lipschitzian in a neighbourhood of any convex set $K \in \mathcal{K}$.

- $\overset{\circ}{V}(K)$ is (weakly) continuous in a neighbourhood of any convex set $K \in \mathcal{K}$ because, for any $f \in \mathcal{D}^1(E, R^p)$,

$$|\overset{\circ}{V}(K_n)f - \overset{\circ}{V}(K)f| \leq \|\mathrm{div}(f)\|.|V(K_n) - V(K)|.$$

- We can write

$$\mathcal{K} = \mathcal{L} \cap V^{-1}(\alpha),$$

where $\mathcal{L}$ is the family of closed-connected subsets of $\mathrm{Int}(E)$ satisfying the ε-cone property. Note that

$$\mathcal{V}^1_{\mathcal{L}}(K) = \mathcal{D}^1_E(K)$$

because the image of a connected set by a continuous function is connected and because $T_f(t, \cdot)$ is Lipschitz.

- The transversality condition can be easily established by taking $f \in \mathcal{D}^1_E$ such that

$$\mathrm{div}(f)_{|K} = \frac{y}{\alpha}.$$

Thus we obtain, in the case when K is convex,

$$\mathcal{V}^1_{\mathcal{K}}(K) = \left\{ f \in \mathcal{D}^1_E \mid \int_K \mathrm{div}(f(x))dx = 0 \right\}.$$

[4] See [4]. Denote by $C(x, v, \varepsilon)$ the truncated cone of angle ε, direction v and vertex x intersected with the ball $B(x, \varepsilon)$. We say that K has the ε-cone property if, for any $x \in \partial K$, there exists a direction v such that

$$C(y, v, \varepsilon) \subset K, \quad \forall\, y \in B(x, \varepsilon) \cap \bar{K}.$$

6. Fermat rule and shape Lagrangian. We now consider the constrained optimization problem on $\mathcal{P}(E)$,

$$J(\widehat{K}) = \inf_{K \in \mathcal{K}} J(K), \tag{50}$$

where E is a subset of R^p, $\mathcal{K}$ a subset of $\mathcal{P}(E)$, and $J : \mathcal{P}(E) \to R$.

6.1. Fermat rule. First, let us recall the definition of the negative polar cone L^- of a subset L of a normed space X as follows:

$$p \in L^- \Longleftrightarrow \forall\, x \in L, \qquad \langle p, x \rangle \leq 0.$$

The Fermat rule is extended to the shape optimization case as follows.

THEOREM 6.1 (Fermat). *Let E be a subset of R^p and $\mathcal{K}$ be a family of compact subsets of E. Let us assume that the map $J : \mathcal{P}(E) \to R$ is Lipschitzian and shape differentiable of order l $(0 \leq l \leq \infty)$ in the neighbourhood of a solution $\widehat{K}$ of the constrained optimization problem* (50). *Then*

$$-\overset{\circ}{J}(\widehat{K}) \in \mathcal{V}_{\mathcal{K}}^{l}(\widehat{K})^-.$$

Proof. Let f be an element of $\mathcal{V}_{\mathcal{K}}^{l}(\widehat{K})$. Then

$$\begin{cases} \exists t_n \to 0^+, \\ \exists K_n \in \mathcal{K} \quad \text{such that } \varepsilon_n = \dfrac{1}{t_n} d(K_n, T_f(t_n, \widehat{K})) \to 0^+. \end{cases} \tag{51}$$

Because the functional J is Lipschitz, we have

$$\|J(K_n) - J(T_f(t_n, \widehat{K}))\| \leq c t_n \varepsilon_n.$$

Moreover, the definition of $\overset{\circ}{J}(\widehat{K})$ leads to

$$J(T_f(t_n, \widehat{K})) = J(\widehat{K}) + t_n(\overset{\circ}{J}(\widehat{K})(f) + \varepsilon(t_n)). \tag{52}$$

Consequently,

$$J(K_n) \leq J(\widehat{K}) + t_n \varepsilon_n + t_n(\overset{\circ}{J}(\widehat{K})(f) + \varepsilon(t_n)).$$

Because $J(K_n) \geq J(\widehat{K})$, dividing by t_n and letting n converge to ∞, we can conclude that

$$\overset{\circ}{J}(\widehat{K})(f) \geq 0. \qquad \square$$

The Euler's optimality condition then appears as a consequence of Fermat's rule (Theorem 6.1) and of Proposition 3.5.

THEOREM 6.2. *Let E be a compact subset of R^p. Let us assume that the shape map $J : \mathcal{P}(E) \to R$ is Lipschitzian and shape differentiable of order l $(0 \leq l \leq \infty)$ in the neighbourhood of a set solution $\widehat{K}$ of the constrained optimization problem* (50). *Then*

$$\overset{\circ}{J}(\widehat{K}) = 0.$$

6.2. Shape Lagrangian. Now let us consider the particular case of shape inequality and equality constraints, that is,

$$J(\widehat{K}) = \inf J(K). \quad \begin{cases} A_i(K) \le 0, & i = 1, \dots, r, \\ B_i(K) = 0, & i = 1, \dots, s, \\ K \in \mathcal{P}(E) \end{cases} \tag{53}$$

The Lagrangian of the shape optimization problem (53) is the map

$$\mathcal{P}(E) \times R^r \times R^s \to R$$
$$(K, \lambda, \beta) \mapsto L(K, \lambda, \beta) = J(K) - \langle \lambda, A(K) \rangle - \langle \beta, B(K) \rangle.$$

THEOREM 6.3. *Let E be a compact subset of R^p and $\widehat{K}$ be a solution of* (53). *Let us assume that $\widehat{K}, B, A$ satisfy* (49) *of Proposition* 5.3 *and that the map J verifies the assumptions of Theorem* 6.1. *Then*

$$\exists (\widehat{\lambda}, \widehat{\beta}) \in R^r_+ \times R^s \quad \textit{such that} \quad \begin{cases} \overset{\circ}{L}(\widehat{K}, \widehat{\lambda}, \widehat{\beta}) = 0, \\ \widehat{\lambda}_i A_i(\widehat{K}) = 0, \qquad \forall i \in \{1, \dots, r\}, \end{cases}$$

and $\widehat{\beta}$ and $\widehat{\lambda}$ are called the Lagrangian multipliers.

Proof. Set $F = (A, B)$ and $M = R^r_+ \times \{0\}^s$. We know from Proposition 5.1 that

$$\overset{\circ}{F}(\widehat{K}) \mathcal{V}^l_{F^{-1}(M)}(\widehat{K}) = T_M(F(\widehat{K})),$$

with $\overset{\circ}{F}(\widehat{K}) \in \mathcal{L}(\mathcal{D}^l(E, R^p), R^n)$.

Observing that $\mathcal{D}^l(E, R^p)$ is a Banach space (because E is compact and $k < \infty$), we use the bipolar theorem[5] to obtain

$$\mathcal{V}^l_{F^{-1}(M)}(\widehat{K})^- = \overset{\circ}{F}(\widehat{K})^* T_M(F(\widehat{K}))^-.$$

Because $T_{\{0\}}(0)^- = X$ and $T_{R^n_+}(x)^- = \{p \,|\, p_i \le 0, p_i = 0 \text{ if } x_i > 0\}$, we conclude the proof. □

6.3. Example. Consider the problem

$$\min \int_K h(x)dx \quad \begin{cases} K \in \mathcal{P}(E) \\ K \text{ connected}, \\ K \, \varepsilon\text{-cone}, \\ V(K) \ge 1, \end{cases} \tag{54}$$

[5] See [1].

BIPOLAR THEOREM. *Let X and Y be two Banach spaces; L, a subset of X; and F, a linear and continuous operator from X into Y. Then*

$$F(L)^- = (F^*)^{-1}(L^-).$$

where h is $\mathcal{C}^1(R^p, R)$ and E is a compact set of R^p. The solution domain $\widehat{K}$ is assumed to be convex.

As in the example in §5.3, the regularity assumptions on V can be checked. The transversality condition on V holds if g_0 is taken such that div $g_0 > 0$ on $\widehat{K}$. The functional $J(K) = \int_K h(x)dx$ is shape differentiable of order 1 and, for any $f \in \mathcal{D}^1_E$, we have

$$\overset{\circ}{J}(K)f = \int_K \operatorname{div}(h(x)f(x))dx.$$

Furthermore, J is Lipschitzian around the convex domain $\widehat{K}$ because

$$J(\widehat{K}) \leq V(\widehat{K}) \sup_{x\in E} \|h(x)\|$$

and because V is Lipschitzian in the neighbourhood of $\widehat{K}$ (see Proposition 5.3).

The Lagrangian optimality condition yields

$$\begin{cases} \widehat{\lambda}(V(\widehat{K}) - 1) = 0, \\ \displaystyle\int_{\widehat{K}} \operatorname{div}(h(x)f(x))dx - \widehat{\lambda}\int_{\widehat{K}} \operatorname{div}(f(x))dx = 0, \quad \forall f \in \mathcal{D}^1(E, R^p). \end{cases}$$

In particular, when $p = 1$, setting $E = [a, b]$ $(b \geq a + 1)$, the solutions of Theorem 6.3 are

$$\begin{cases} \widehat{K_1} = [a, a + 1], & \widehat{\lambda_1} = 1, \\ \widehat{K_2} = [b - 1, b], & \widehat{\lambda_2} = b - 1. \end{cases} \tag{55}$$

7. Appendix. We recall below Nagumo's theorem [2], whereby we can define shape directional derivatives and mention some properties of the solution map $T_f(t, \cdot)$.

THEOREM 7.1 (Nagumo). *Let E be a subset of R^p and f a map from R^p into R^p satisfying*

$$\begin{cases} \text{(i)} \ f \text{ is Lipschitzian on } \bar{E}, \\ \text{(ii)} \ \forall x \in \bar{E}, \qquad f(x) \in T_E(x) \cap -T_E(x). \end{cases} \tag{56}$$

Then, for all $t > 0$,

$$\begin{cases} \text{(a)} \ T_f(t, \cdot) \text{ is Lipschitzian on } \bar{E}, \\ \text{(b)} \ T_f(t, \cdot) \text{ is a bijection from } \bar{E} \text{ into } \bar{E}, \\ \text{(c)} \ T_f(t, \cdot)^{-1} \text{ is Lipschitzian on } \bar{E}. \end{cases} \tag{57}$$

LEMMA 7.2. *Let E be a subset of R^p and f be in $\mathcal{F}^k_E$ $(k \geq 0)$. Then there exists $T > 0$ such that*

$$\forall t \in]0, T[, \qquad \frac{1}{t}(T_f(t, \cdot) - I) \in \mathcal{D}^k(E, R^p). \tag{58}$$

Proof. By standard arguments (see, for example, [12]), we can establish that, for any $t > 0$,

$$T_f(t, \cdot) \in C^k(E, R^p).$$

Consequently, for any $t > 0$,

$$\frac{1}{t}(T_f(t, \cdot) - I) \in C^k(E, R^p).$$

Now let us consider $x \notin D = \text{Supp}(f) = \overline{\{y \in E \mid f(y) \neq 0\}} \subset E$. Because D is closed, there exists $\eta > 0$ such that $B(x, \eta) \cap D = \emptyset$. Consider a solution $x(\cdot)$ of (1). By continuity of $x(\cdot)$ around zero, we know that there exists $T > 0$ such that

$$|t| < T \Rightarrow \|x(t) - x\| < \eta.$$

Consequently, for any $t \in]0, T]$, we can write

$$(T_f(t, \cdot) - I)(x) = \int_0^t f(x(s))ds = 0.$$

Therefore, for any $t \in]0, T[$,

$$\text{Supp}(T_f(t, \cdot) - I) \subset \text{Supp}(f) \subset E,$$

which completes the proof. □

LEMMA 7.3. *Let K be a compact subset of $E \subset R^p$ and $f : E \to R^p$ be a function satisfying the assumptions* (56) *of Theorem* (7.1). *Then*

$$\forall\, t_n \to 0^+, \qquad \frac{T_f(t_n, \cdot) - I}{t_n} \to f \quad \textit{uniformly on } K.$$

Proof. First, we have that, for any t converging to 0^+ and for any x in K,

$$\frac{1}{t}(T_f(t, x) - x) = \frac{1}{t}\int_0^t f(x(s))ds \to f(x),$$

which implies the pointwise convergency.

Using Gronwall's lemma, we can establish that

$$\|(T_f(t, \cdot) - I)(x) - (T_f(t, \cdot) - I)(y)\| \leq (e^{kt} - 1)\|x - y\|,$$

where k is the Lipschitz constant of f.

Let us consider any sequence t_n converging to 0^+, and the function α defined by

$$\alpha(t) = \frac{1}{t}(e^{kt} - 1), \qquad \alpha(0) = k.$$

Because this sequence is bounded and α is continuous around zero, the sequence $\alpha(t_n)$ is bounded.

Therefore $\{(1/t_n)(T_f(t_n, \cdot) - I)\}_{n \in N}$ is an equicontinuous family of continuous and bounded functions on K.

Consequently, because K is compact, we can apply the Ascoli theorem and deduce that $\frac{1}{t}(T_f(t, \cdot) - I) \to f$ uniformly on K. □

LEMMA 7.4. *Let E be a subset of R^p and K be a compact subset of E. Let f and g be two applications belonging to $\mathcal{F}_E^0$. Then*

$$\lim_{h \to 0^+} \frac{1}{h} d(T_f(h, T_g(h, K)), T_{f+g}(h, K)) = 0.$$

Proof. For any $x \in K$, we have

$$\begin{aligned}\frac{1}{h}(T_f[h, T_g(h, x)] - T_{f+g}(h, x)) = &\left(\frac{T_f(h, T_g(h, x)) - T_g(h, x)}{h} - f(T_g(h, x))\right) \\ &+ \left(\frac{T_{f+g}(h, x) - x}{h} - (g + f)(x)\right) \\ &+ \left(\frac{T_g(h, x) - x}{h} - g(x)\right) + f(T_g(h, x)) - f(x).\end{aligned}$$

Now let us denote

$$\begin{cases} i_1^h = \dfrac{T_{f+g}(h,\cdot) - I}{h} - (f+g), \\ g_1^h = \dfrac{T_g(h,\cdot) - I}{h} - g, \\ g_1^h = \dfrac{T_f(h,\cdot) - I}{h} - f. \end{cases}$$

From Lemma 7.2 we can write

(59)

$$\frac{1}{h}\sup_{x\in K}\|T_f[h,T_g(h,x)] - T_{f+g}(h,x)\| \le \sup_{x\in \mathrm{Supp}(f_1)}\|f_1(x)\| + \sup_{x\in \mathrm{Supp}(g_1^h)}\|g_1^h(x)\| + \sup_{x\in \mathrm{Supp}(i_1^h)}\|i_1^h(x)\| + \sup_{x\in E}\|f(T_g(h,x)) - f(x)\|.$$

Because g is continuous on the compact set $\mathrm{Supp}(f)$, we have

$$\|g\| = \sup_{x\in E}\|g(x)\| \le M < \infty.$$

Thus, because f is Lipschitzian, we have

(60)

$$\begin{aligned} \|f(T_g(h,x)) - f(x)\| &\le c\|T_g(h,x) - x\| \\ &\le c\left\| \int_0^h g(x(s))ds \right\| \\ &\le hMc. \end{aligned}$$

By virtue of (60) and Theorem 7.3 we can conclude that

$$\frac{1}{h}d\!\!\!d(T_f[h,T_g(h,K)],T_{f+g}(h,K)) \to 0. \qquad \square$$

LEMMA 7.5. *Let E be a subset of R^p and K_1, K_2 be two compact subsets of E. Consider $f \in \mathcal{F}_E^0$. Then*

$$\exists T > 0, \quad \exists k > 0, \quad \forall h \in [0,T], \quad d\!\!\!d(T_f(h,K_1),T_f(h,K_2)) \le k d\!\!\!d(K_1,K_2).$$

Proof. Because f is Lipschitzian, we know from Theorem 7.1 that $T_f(h,\cdot)$ is also Lipschitzian. Thus

$$\sup_{x\in K_2}\inf_{y\in K_1}\|T_f(h,x) - T_f(h,y)\| \le k \sup_{x\in K_2}\inf_{y\in K_1}\|x-y\|,$$

and the lemma is established. $\square$

LEMMA 7.6. *Let f and g be two functions that belong to $\mathcal{F}_E^0$, and let K_1 be a compact subset of E. Then*

$$\lim_{h\to 0+}\frac{1}{h}d\!\!\!d(T_g(h,K_1),T_f(h,K_1)) \le 2\|f-g\|_\infty.$$

Proof. For any $x \in K_1$, we have

$$\|T_g(h,x) - T_f(h,x)\| \le h(\|g_1^h(x)\| + \|g(x) - f(x)\| + \|f_1^h(x)\|),$$

where

$$\begin{cases} g_1^h = \dfrac{T_g(h,\cdot) - I}{h} - g, \\ f_1^h = \dfrac{T_f(h,\cdot) - I}{h} - f. \end{cases}$$

We know from Lemma 7.2 that f_1^h and g_1^h belong to $\mathcal{D}^0(E, R^p)$. Consequently,

$$\sup_{x\in K_1} \|T_g(h,x) - T_f(h,x)\| \leq h\left(\sup_{x\in \mathrm{Supp}(g_1^h)} \|g_1^h(x)\| + \sup_{x\in E} \|g(x) - f(x)\| + \sup_{x\in \mathrm{Supp}(f_1^h)} \|f_1^h(x)\|\right).$$

Hence, using Lemma 7.3 and the fact that the supports of g_1^k and f_1^h are compact, we obtain

$$\frac{1}{h} d_H(T_g(h,K_1), T_f(h,K_1)) \leq \varepsilon_1(h) + 2\|f - g\|_\infty + \varepsilon_2(h).$$

This completes the proof of the lemma. □

REFERENCES

[1] J.-P. Aubin and H. Frankowska (1990), *Set-valued analysis*, Birkhäuser, Boston, Basel, Berlin.

[2] J.-P. Aubin (1991), *Viability theory*, Birkhäuser, Boston, Basel, Berlin.

[3] J. Céa (1981), *Problems of shape optimal design*, Optimization of Distributed Parameter Structures, Vols. I and II, E. J. Haug and J. Céa, eds., Sijhoff and Noorhoff, Alphen aan den Rijn, the Netherlands, pp. 1005–1087.

[4] D. Chenais (1975), *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52, pp. 189–289.

[5] M. C. Delfour and J. P. Zolesio (1990), *Structure of shape derivatives for nonsmooth domains*, Report CRM-1669, Centre de Recherche Mathématiques de l'université de Montréal, Canada.

[6] ——— (1988), *Shape sensitivity analysis via minmax differentiability*, SIAM J. Control Optim., 26, pp. 834–862.

[7] ——— (1989), *Anatomy of the shape hessian*, Ann. Mat. Pura Appl., 153, CLIII.

[8] ——— (1989), *Computation of the shape hessian by a lagrangian method*, in Fifth Symp. on Control of Distributed Parameter Systems, A. El Jai and M. Amouroux, eds., Pergamon Press, pp. 85–90.

[9] ——— (1989), *Shape hessian by the velocity method, a lagrangian approach*, Stabilization of Flexible Structures, J. P. Zolesio, ed., Springer-Verlag, Berlin.

[10] ——— (1991), *Velocity method and lagrangian formulation for the computation of the shape Hessian*, SIAM J. Control Optim., 29, pp. 1414–1442.

[11] H. Frankowska (1990), *Some inverse mapping theorems*, Ann. Inst. Henri Poincaré, Analyse non linéaire, 3, pp. 183–234.

[12] J. K. Hale (1969), *Ordinary differential equations*, Interscience Series on Pure and Applied Mathematics, Vol. 21, John Wiley, New York.

[13] G. Matheron (1975), *Random sets and integral geometry*, John Wiley, New York.

[14] O. Pironneau (1983), *Optimal shape design for elliptic systems*, Computational Physics, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo.

[15] J. Sokolowski and J. P. Zolesio (1992), *Introduction to shape optimization and shape analysis*, Springer Series in Computational Mathematics.

[16] J. P. Zolesio (1979), *Identification de domaines par déformation*, Thèse de doctorat d'état, Université de Nice, France.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1643–1674, November 1994

009

ESTIMATION OF INTERFACES FROM BOUNDARY MEASUREMENTS*

KARL KUNISCH† AND XIAOSU PAN‡

Abstract. The determination of an interface in an elliptic differential equation from Dirichlet and Neumann boundary data is investigated. Uniqueness of the function characterizing the interface is proved. For numerical purposes the problem is formulated as an optimization problem involving constraints that are based on potential theory. An augmented Lagrangian formulation is used for the solution of the optimization problem. Its convergence is investigated and numerical experiments are described.

Key words. inverse problems, elliptic differential equation, augmented Lagrangian techniques, interface problem

AMS subject classifications. 35R30, 65M30, 31A25

1. Introduction.

The aim of this contribution is the study of the reconstruction of an interface from boundary measurements. We are given a domain Ω in $\mathbf{R}^2$ and a curve a connecting two points of the boundary of Ω and dividing Ω into two connected subdomains Ω_1 and Ω_2. The governing equation is given by

$$-\Delta u = f_a \quad \text{in } \Omega, \tag{1.1}$$

where

$$f_a(x,y) = \begin{cases} \rho_1 & \text{for } (x,y) \in \Omega_1, \\ \rho_2 & \text{for } (x,y) \in \Omega_2, \end{cases}$$

with ρ_1 and ρ_2 known constants. The problem consists of determining a from measurements (z_1, z_2) of the boundary data $(u|\Gamma, \partial u/\partial n|\Gamma)$, with Γ the boundary of Ω. We also address the question when the data are available only on part of the boundary.

The motivation for this study originates from problems in the geophysical sciences. The earth is a stratified body made of layers with high density contrasts. On a macroscopic scale one distinguishes between the earth's core, its mantle, and its crust. The interfaces between these layers are referred to as the core-mantle boundary and the Mohorovičić discontinuity, respectively. While the densities within these layers are fairly well known, the precise location of the layers is not. A similar situation occurs in the crust itself. Over wide areas it consists of distinct layers made of material with high density differences. The problem consists of determining the location of the interfaces. To achieve this goal one uses measurements of the earth's potential on the surface of the earth and its low order derivatives, both measured on the surface and eventually in space. Such measurements are available with high data density and with high precision. The mathematical model that combines the quantities density and potential is given by the potential equation

$$-\Delta u = f \quad \text{in } \Omega,$$

* Received by the editors February 18, 1992; accepted for publication (in revised form) April 23, 1993.

† Fachbereich Mathematik, Technische Universität Berlin, Strasze des 17, Juni 136, D-10623 Berlin, Germany. The work of this author was supported in part by a fund from the Bundesministerium für Wissenschaft und Forschung, Austria.

‡ Department of Mathematics, Nanjing Aeronautical Institute, Nanjing 210016, People's Republic of China.

where u denotes potential and f stands for density. The density itself depends on the location of the layers. The domain Ω represents all or part of the solid earth. Data are available for the potential u and its normal derivative $\partial u/\partial n$ on the surface of the earth. Here n denotes the unit outer normal to the surface of the earth. In the case where Ω represents only part of the solid earth, data will at first only be available on a part of the boundary Γ of Ω. The remaining parts of the boundary lie inside the earth. As a specific case let Ω represents a mountain arising above the geoid (Fig. 1). In this case primary data are available on the surface of the mountain but not on the geoid. One can use geophysical continuation methods, supplemented by mass remove-restore procedures, to obtain secondary (derived) data on that part of the boundary of Ω which represents the bottom of the mountain. As references one may consult [L], [M]. In the mathematical analysis we consider the two dimensional cross section, which is again denoted by Ω.

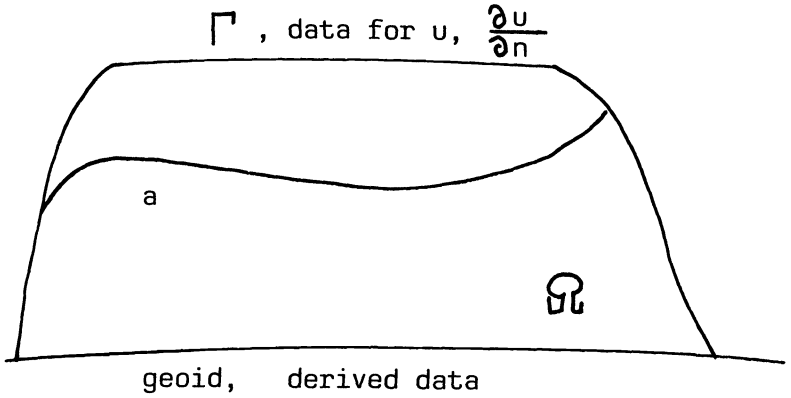


FIG. 1

We first establish conditions that ensure uniqueness of the function a from the measurements. Then Green's formula is used to establish certain compatibility conditions. These compatibility conditions play an essential part in the numerical solution of the estimation problem, which is based on an optimization theoretic formulation. We propose a least-squares formulation with the potential theoretic information in the form of the compatibility conditions as constraints. Numerically, the compatibility conditions also serve the purpose of providing a good start-up value for the unknown interface. During the iteration the cost functional combines a regularized least squares term, an equation error term, and the compatibility conditions. An additional distinct feature of our approach is that the unknown function a and the state variable u of (1.1) are both considered as independent variables related by the equation $\Delta u + f_a = 0$, which is also realized as a constraint. The optimization problem is solved by an augmented Lagrangian technique that is shown to converge provided that a second-order sufficient optimality condition holds. This second-order condition is investigated independently and can be shown to hold in specific cases. The paper ends with a description of numerical experience with the proposed algorithm. For the presentation of our results and specifically for the numerical code we chose Ω to be a rectangle. The choice of this specific domain allows certain simplifications, but it also implies some difficulties due to the lack of high smoothness of the boundary.

The contents of the following sections is best characterized by their titles.

1. Identifiability.
2. The compatibility conditions.

3. Formulation of the inverse problem as a constrained minimization problem.
4. Existence of Lagrange multiplier.
5. The augmented Lagrangian algorithm.
6. The augmentability condition.
7. A numerical example.
8. References.

2. Identifiability. In this section we prove the identifiability of the function a characterizing the interface from boundary measurements. For convenience we recall the problem under consideration as follows:

$$\begin{cases} -\Delta u = f_a & \text{in } \Omega, \\ u|\Gamma = z_1, \\ \partial u/\partial n|\Gamma = z_2, \end{cases} \tag{2.1}$$

where $\Omega = \{(x,y) \in \mathbf{R}^2 : 0 < x < 2, 0 < y < 1\}$; n; denotes the outward normal to Ω, Γ stands for the boundary of Ω;

$$f_a(x,y) = \begin{cases} \rho_1 & \text{if } y > a(x), \\ \rho_2 & \text{if } y < a(x), \end{cases}$$

with ρ_1 and ρ_2 fixed values in $\mathbf{R}$; and $a : [0,2] \to [0,1]$. The boundary data z_1 and z_2 are assumed to be known. If the function a was known as well, then (2.1) would be an overdetermined system.

However, a is unknown and we seek to determine it from knowledge of the boundary data z_1 and z_2. From Theorem 2.1 below it follows that the boundary data z_1 and z_2 determine a uniquely. For $a : [0,2] \to [0,1]$, we denote by $u(a)$ a solution of $-\Delta u = f_a$. Further, $\Gamma_i, i = 1, ..., 4$, denote the four edges of Ω, enumerated in a counterclockwise manner, starting with Γ_1 as the lower edge. The Dirichlet and Neuman trace operators on Γ are denoted by $\tau_0 u = u|\Gamma$ and $\tau_1 u = \partial u/\partial n|\Gamma$, respectively. In this and the following section we only require the restriction of τ_0 and τ_1 to elements $u \in H^2(\Omega)$. In this case $\tau_0 u \in H^{1/2}(\Gamma), \tau_1 u$ is interpreted as $\tau_1 u = \Pi_{j=1}^4 \partial u/\partial n|\Gamma_j$, and Green's formula holds (see Lemmas A.1 and A.3).

THEOREM 2.1. *Let a and $\tilde{a}$ be piecewise $C^{1,1}$-functions mapping $[0,2]$ into $[0,1]$, and let $u(a)$ and $u(\tilde{a})$ be solutions in $H^2(\Omega)$ of $-\Delta u = f_a$ and $-\Delta u = f_{\tilde{a}}$, respectively. Then $\rho_1 \neq \rho_2$ and*

$$(\tau_0 u(a), \tau_1 u(a)) = (\tau_0 u(\tilde{a}), \tau_1 u(\tilde{a})) \tag{2.2}$$

imply $a = \tilde{a}$.

Proof. Let us assume that $a \neq \tilde{a}$. Set $w = u(a) - u(\tilde{a})$ and note that $w \in H^2(\Omega) \subset C^{0,1}(\bar{\Omega})$ by Lemma A.4. The trace operators in (2.2) are well defined (see Lemma A.1). We denote by Ω_1 and Ω_2 the subdomains of Ω which lie above, respectively below, $a(x)$. More precisely,

$$\Omega_1 = \{(x,y) : 0 < x < 2, a(x) < y < 1\},$$

$$\Omega_2 = \{(x,y) : 0 < x < 2, 0 < y < a(x)\}.$$

The subdomains $\tilde{\Omega}_1$ and $\tilde{\Omega}_2$ are defined analogously with a replaced by $\tilde{a}$. Further we put

$$U = (\Omega_1 \cap \tilde{\Omega}_1) \cup (\Omega_2 \cap \tilde{\Omega}_2).$$

This is the union of the two domains which are above and below both a and $\tilde{a}$. We find that

$$-\Delta w = 0 \quad \text{in } U.$$

Since Δ is an analytic hypoelliptic operator [Tr], it follows that w is analytic in U. Due to (2.2), w satisfies homogeneous Cauchy data on Γ_3. The Cauchy–Kowalewsky theorem and analytic continuation [Mi] can now be used to argue that $w = 0$ on $\Omega_1 \cap \tilde{\Omega}_1$. An analogous argument implies that $w = 0$ on $\Omega_2 \cap \tilde{\Omega}_2$. Since $w \in C^{0,1}(\bar{\Omega})$ it follows that

$$w = 0 \quad \text{in } \bar{U}. \tag{2.3}$$

Next we turn to the region *between* a and $\tilde{a}$, and define

$$B = \{x \in [0,2] : a(x) = \tilde{a}(x), \text{ or } x = 0, \text{ or } x = 2\}.$$

Due to continuity of a and $\tilde{a}$, B is closed and $A := [0,2]\backslash B$ is open in $\mathbf{R}$. Since $a \neq \tilde{a}$, the set A is nonempty and it can be represented as $A = \cup_{i=1}^{\infty} I_i$, with I_i nonempty open pairwise disjoint intervals in $[0,2]$. (The union could of course be finite.) Denoting by α_i and β_i the left and right endpoints of I_i, we have α_i and $\beta_i \in B$ for all i. We further put

$$S_i = \{(x,y) : x \in I_i, m(x) < y < M(x)\},$$

where $m(x) = \min(a(x), \tilde{a}(x)), M(x) = \max(a(x), \tilde{a}(x))$. By construction, $m(x) < M(x)$ for all $x \in I_i$, and therefore

$$\text{meas}(S_i) = \iint_{S_i} dxdy = \int_{\alpha_i}^{\beta_i} (M(x) - m(x))dx > 0 \tag{2.4}$$

for every i. We now consider two cases. First, let $\alpha_i \neq 0$ and $\beta_i \neq 2$. Then

$$-\Delta w = \pm(\rho_1 - \rho_2) \quad \text{in } S_i, \tag{2.5}$$

and the transmissivity condition (see Lemma A.2) and (2.3) imply

$$\frac{\partial w}{\partial n}|\partial S_i = 0. \tag{2.6}$$

Here n denotes the unit normal to the boundary ∂S_i of S_i and points into the exterior of S_i. Due to the regularity assumptions on a and $\tilde{a}$, the boundary ∂S_i of S_i is piecewise $C^{1,1}$ and it is in this sense that (2.6) is well defined. Since S_i is a curvilinear polygon with a $C^{1,1}$ boundary, Green's formula (with $v = 1$, Lemma A.3) together with (2.5), (2.6) can be applied to induce

$$(\rho_1 - \rho_2)\ \text{meas}(S_i) = 0,$$

which contradicts $\rho_1 \neq \rho_2$ and meas $(S_i) > 0$. In the second case $\alpha_i = 0$ or $\beta_i = 2$, or both. We then find

$$\partial S_i \subset (\bar{S}_i \cap \overline{\Omega_1 \cap \tilde{\Omega}_1}) \cup (\bar{S}_i \cap \overline{\Omega_2 \cap \tilde{\Omega}_2}) \cup (\bar{S}_i \cap \Gamma).$$

As before, $\partial w/\partial n|(\bar{S}_i \cap \overline{\Omega_1 \cap \tilde{\Omega}_1}) = \partial w/\partial n|(\bar{S}_i \cap \overline{\Omega_2 \cap \tilde{\Omega}_2}) = 0$ and $\partial w/\partial n|(\bar{S}_i \cap \Gamma) = 0$. Again we can use Greens's formula to obtain a contradiction. Hence we have shown that $a = \tilde{a}$. □

Remark 2.2. From the proof it follows that the assumptions on knowledge of the boundary data can be weakened. In fact the conclusion of Theorem 2.1 remains correct, if (2.2) is replaced by

$$\left(u(a)\left|\gamma_1 \cup \gamma_3, \frac{\partial u(a)}{\partial n}\right|\tilde{\Gamma}\right) = \left(u(\tilde{a})\left|\gamma_1 \cup \gamma_3, \frac{\partial u(\tilde{a})}{\partial n}\right|\tilde{\Gamma}\right), \tag{2.7}$$

where $\tilde{\Gamma} = \Gamma_2 \cup \Gamma_4 \cup \gamma_1 \cup \gamma_3$ and γ_1, γ_3 are nonempty open subsets of Γ_1, Γ_3, respectively.

3. The compatibility properties. If, for a function a mapping $[0,2]$ into $[0,1]$, (2.1) has a solution $u = u(a) \in H^2(\Omega)$, then a must satisfy certain compatibility properties which we could refer to also as moment problems. We derive them next. They will be used for the formulation of determining the interface a numerically from knowledge of data z_1, z_2. The following assumptions will be used:

$$\rho_1 \neq \rho_2, \tag{3.1}$$

$$a : [0,2] \to [0,1] \quad \text{is measurable,} \tag{3.2}$$

$$u = u(a) \in H^2(\Omega) \quad \text{satisfies (2.1).} \tag{3.3}$$

We note that, due to trace theorems, (3.3) implies that $z_1 = \tau_0 u \in C(\Gamma) \cap \Pi_{i=1}^4 H^{3/2}(\Gamma_i)$ and $z_2 = \tau_1 u \in \Pi_{j=1}^4 H^{1/2}(\Gamma_j)$.

Let w be any harmonic function in Ω, i.e. $-\Delta w = 0$ in Ω. Due to Green's formula,

$$\iint_\Omega (u\Delta w - w\Delta u)d\Omega = \int_\Gamma \left(z_1 \frac{\partial w}{\partial n} - wz_2\right)d\Gamma,$$

and consequently

$$\iint_\Omega w f_a d\Omega = \int_\Gamma \left(z_1 \frac{\partial w}{\partial n} - wz_2\right)d\Gamma.$$

This further implies

$$\iint_{y<a(x)} w\, dxdy = \frac{1}{\rho_2 - \rho_1}\left[\int_\Gamma \left(z_1 \frac{\partial w}{\partial n} - wz_2\right)d\Gamma - \rho_1 \iint_\Omega w d\Omega\right]. \tag{3.4}$$

The boundary integrations along Γ have to be interpreted as the sum of the boundary integrations along the four edges Γ_j. Let χ_a be the characteristic function of the domain $\{(x,y) \in \Omega : y < a(x)\}$. Then (3.4) becomes

$$\langle w, \chi_a \rangle_{L^2(\Omega)} = A(w, z_1, z_2), \tag{3.5}$$

where

$$A(w, z_1, z_2) = \frac{1}{\rho_2 - \rho_1}\left[\int_\Gamma \left(z_1 \frac{\partial w}{\partial n} - wz_2\right)d\Gamma - \rho_1 \iint_\Gamma w d\Omega\right].$$

Note that A is independent of a. We point out that the compatibility properties (3.5) are well defined for $(z_1, z_2) \in H^{1/2}(\Gamma) \times H^{1/2}(\Gamma)^*$. Since (3.5) has to hold for any harmonic function, we have derived infinitely many compatibility conditions. In practice we use at most four. These are characterized by the functions

$$w_1(x,y) = 1, \quad w_2(x,y) = 2y, \quad w_3(x,y) = x, \quad w_4(x,y) = 2xy.$$

This leads to four basic compatibility properties which we abbreviate by

$$G_i(a) = A_i. \tag{3.6}$$

The following formulas for $G_i(a) = \langle w_i, \chi_a \rangle_{L^2(\Omega)}$ are given for later reference:

$$G_1(a) = \int_0^2 a \, dx, \qquad G_2(a) = \int_0^2 a^2 dx,$$

$$G_3(a) = \int_0^2 xa \, dx, \qquad G_4(a) = \int_0^2 xa^2 dx.$$

4. Formulation of the inverse problem as a constrained minimization problem. For the purpose of numerical determination of a from z_1 and z_2, the following optimization problem is introduced:

$$\min J(a,u) \text{ over } (a,u) \in Q_{ad} \times H^1(\Omega) \quad \text{with } e(a,u) = 0, \tag{4.1}$$

where

$$J(a,u) = \frac{1}{2}|\mathcal{D}(\tau_0 u - z_1)|^2_{H^1(\Omega)} + \frac{1}{2}|\mathcal{N}(\tau_1 u - z_2)|^2_{H^1(\Omega)} + \frac{1}{2}\beta |a_x|^2_{L^2(0,2)}, \tag{4.2}$$

$$e(a,u) = N(\Delta u + f_a), \tag{4.3}$$

and

$$Q_{ad} = \{a \in H^1(0,2) : a(x) \in [0,1], G_i(a) = A_i, i = 1, ..., 4\}. \tag{4.4}$$

In (4.2) the regularization parameter β is chosen in $[0, \infty)$. Throughout it is assumed that $(z_1, z_2) \in H^{1/2}(\Gamma) \times H^{1/2}(\Gamma)^*$ and that Q_{ad} is not empty. The operators $\mathcal{D} : H^{1/2}(\Gamma) \to H^1(\Omega)$, $\mathcal{N} : H^{1/2}(\Gamma)^* \to H^1(\Omega)$, and $N : H^1(\Omega)^* \to H^1(\Omega)$ are Dirichlet and Neumann solution operators given by

$$\mathcal{D}g = u \quad \text{with } -\Delta u = 0, \quad \tau_0 u = g,$$

$$\mathcal{N}g = u \quad \text{with } -\Delta u + u = 0, \quad \tau_1 u = g,$$

$$\begin{aligned} Nf &= u \quad \text{with } u \text{ the solution of } \langle \nabla u, \nabla v \rangle_{L^2} + \langle u, v \rangle_{L^2} \\ &= \langle f, v \rangle_{H^1, {}^*H^1} \quad \text{for all } v \in H^1(\Omega). \end{aligned}$$

They are all isomorphisms (see, e.g., [DL], [GR]).

The data $(z_1, z_2) \in H^{1/2}(\Gamma) \times H^{1/2}(\Gamma)^*$ are called *attainable* if there exists $(\hat{a}, \hat{u}) \in Q_{ad} \times H(\Delta, \Omega)$, such that

$$e(\hat{a}, \hat{u}) = 0, \quad \tau_0 \hat{u} = z_1 \quad \text{in } H^{1/2}(\Gamma) \quad \text{and} \quad \tau_1 \hat{u} = z_2 \quad \text{in } H^{1/2}(\Gamma)^*.$$

In this case $(\hat{a}, \hat{u})$ is a solution of the unregularized problem, i.e., $\beta = 0$ in (4.2), and $J(\hat{a}, \hat{u}) = 0$.

The equality constraints $G_i(a) = A_i$ are referred to as compatibility conditions. For attainable data they are automatically satisfied since they coincide with the compability properties derived in §3. The assumption of nonemptiness of Q_{ad} is not a severe one. In fact, a simple implicit function theorem argument implies that for every observation $(\hat{z}_1, \hat{z}_2)$ which is attainable by some pair $(\hat{a}, \hat{u})$, there exists a neighborhood $V(\hat{z}_1, \hat{z}_2)$ in $H^{1/2}(\Gamma) \times H^{1/2}(\Gamma)^*$ such that for every $(z_1, z_2) \in V(\hat{z}_1, \hat{z}_2)$, the set Q_{ad} is not empty, provided that $\hat{a}$ is not identically 0 or 1 and that $L := \{1, \hat{a}, x, \hat{a} \cdot x\}$ is not linearly dependent on some subintervall of $(0, 2)$. If only a subset of all four compatibility conditions is used for the definition of Q_{ad}, then the last statement remains correct if the set L is defined as the set of corresponding elements $1, \hat{a}, x$, or $\hat{a} \cdot x$.

Let us comment on (4.1)–(4.4). This is a least squares formulation with the equation $-\Delta u = f_a$ realized as an explicit constraint. The state variable u and the parameter a are both independent variables. If $\beta > 0$, then $\beta |a_x|^2_{L^2(0,2)}$ represents a Tikhonov regularization term. The specific choice of the operators $\mathcal{D}, \mathcal{N}$, and N guarantee that the Dirichlet as well as the Neumann boundary data residuals and the equation constraint $e(a, u) = 0$ are all uniformly represented as H^1-functions. The operators $\mathcal{D}$ and $\mathcal{N}$ in the definition of J can also be interpreted as preconditioning [IKK]. The use of a regularization term like $\beta |a_x|_{L^2}$ is well known for nonlinear inverse problems, we refer to [EKN], [IK2], [KS] and the literature cited there. We approach (4.1)–(4.4) by a Lagrangean framework, which requires the Hilbert space structure of Q_{ad}. The constraints $G_i(a) = A_i, i = 1, \ldots, 4$, in the definition of Q_{ad} were derived in §3. Numerically they play the role of providing good initial guesses and of stabilizing the algorithm. While we have used all four compatibility conditions in the formulation of problem (4.1)–(4.4), all the theoretical results—except for Theorem 4.1(ii), where $G_1(a) = A_1$ is used—remain correct without these conditions and their proofs become easier. To avoid index notation for subsets, we chose to put all four conditions as explicit constraints.

The cost functional $J(a, u)$ in (4.2) could equivalently be expressed as

$$J(a, u) = \frac{1}{2} |\tau_0 u - z_1|^2_{H^{1/2}(\Gamma)} + \frac{1}{2} |\tau_1 u - z_2|^2_{H^{1/2}(\Gamma)^*} + \frac{1}{2} \beta |a_x|^2_{L^2(0,2)}.$$

Here we chose the form given in (4.2) since it also describes the technique by which the $H^{1/2}(\Gamma)$ and the $H^{1/2}(\Gamma)^*$ norms were implemented numerically. The formulation (4.1)–(4.4) for the estimation of a from boundary measurements is motivated by previous work on the estimation of the coefficient a in $-\operatorname{div}(a \operatorname{grad} u) = f$ from data z corresponding to u (see [IK2], [IKK]) and from an optimization theoretic formulation of the impedance computed tomography problem [IJ]. There are several alternatives to formulating our estimation problem as an optimization problem. One of them is given by

$$\min_{\substack{a \in Q_{ad} \\ -\Delta u(a) = f_a \\ \tau_0 u(a) = z_1}} |\tau_1 u(a) - z_2|^2_{H^{1/2}(\Gamma)^*} + \beta |a_x|^2_{L^2(0,2)}. \tag{4.5}$$

This is a regularized least-squares formulation where u is treated as a dependent variable. The numerical success with (4.1)–(4.4) justifies giving priority to the more involved formulations over simpler ones like (4.5), which we expect to be less successful numerically. An alternative to (4.5) would be to enforce the Neumann boundary condition $\tau_1 u(a) = z_2$ and to use the Dirichlet condition in the cost functional. In either case the formulation would be asymetric and it would not be balanced with respect to errors in z_1 and z_2.

Before we commence our study of (4.1) it is necessary to address a technical issue that will have consequences throughout the remainder of this paper. We are referring to the fact that the Neumann boundary operator cannot be extended from the set of test functions $\mathcal{D}(\bar{\Omega})$ to a continuous linear operator from $H^1(\Omega)$ to $H^{1/2}(\Gamma)^*$. Additional regularity for the elements in the domain is required. One possibility is given by considering the space $H(\Delta, \Omega) = \{u \in H^1(\Omega) : \Delta u \in L^2(\Omega)\}$ [DL], [GR], where more precisely $\Delta u \in L^2(\Omega)$ means that the distribution $v \to \langle \Delta u, v \rangle_{L^2(\Omega)}$ for $v \in H_0^1(\Omega)$ can be indentified with an element of $L^2(\Omega)$. The regularity provided by $H(\Delta, \Omega)$ is sufficient but not necessary for a well-defined Neumann operator $H^{1/2}(\Gamma)^* \to H^1(\Omega)$. We shall return to this comment at the beginning of §5. For the moment it suffices to note that $\tau_1 u$ in (4.2) is well defined due to the fact that the constraint $e(a, u) = 0$ implies that $u \in H(\Delta, \Omega)$.

The existence problem for (4.1)–(4.4) is addressed next.

THEOREM 4.1. (i) *If $\beta > 0$, then* (4.1)–(4.4) *has a solution $(a^*, u^*) \in Q_{ad} \times H^1(\Gamma)$.*

(ii) *Any solution (a^*, u^*) of* (4.1)–(4.4) *satisfies*

$$\int_\Omega \mathcal{D}(\tau_0 u^*) d\Omega = \int_\Omega \mathcal{D} z_1 d\Omega$$

and

$$\begin{cases} |u^* - \tilde{u}|_{H^1(\Omega)} \le C|\tau_1 u^* - z_2|_{H^{1/2}(\Gamma)^*} \le c|\mathcal{D}(\tau_0 \tilde{u} - z_1)|_{H^1(\Omega)}, \\ J(a^*, u^*) \le \dfrac{1}{2}|\mathcal{D}(\tau_0 \tilde{u} - z_1)|^2_{H^1(\Omega)} + \dfrac{\beta}{2}|a_x^*|^2_{L^2(0,2)}, \end{cases}$$

where c and C are independent of (a^*, u^*) and $\tilde{u} = \int_\Omega \mathcal{D} z_1 d\Omega / \mathrm{meas}(\Omega) + \tilde{u}_2$, with $\tilde{u}_2$ the unique solution of

$$\begin{cases} -\Delta \tilde{u}_2 = f_{a^*}, \\ \tau_1 \tilde{u}_2 = z_2, \\ \int_\Omega \mathcal{D}(\tau_0 \tilde{u}_2) d\Omega = 0. \end{cases}$$

(iii) *If the data (z_1, z_2) are attainable and if in addition $z_1 \in C(\Gamma) \cap \Pi_{j=1}^4 H^{3/2}(\Gamma_j)$, then the solution (a^*, u^*) of* (4.1)–(4.4) *with $\beta = 0$ is unique in the class $(Q_{ad} \cap \{a$ piecewise $C^{1,1}\}) \times H^1(\Omega)$.*

To interpret the estimate involving u^* and $\tilde{u}$ in (ii), observe that the distance between u^* and $\tilde{u}$ is bounded by the amount that $\tau_1 u^*$ fails to match the data z_2, respectively, $\tau_0 \tilde{u}$ fails to match data z_1. The function $\tilde{u}$ satisfies the differential equation, the Neumann boundary condition, and the Dirichlet boundary condition in an averaged sense.

Proof. (i) Throughout, c denotes a generic constant which depends on embedding constants, Poincaré-type inequalities, and meas(Ω). Let (a_k, u_k) be a minimizing sequence for (4.1)–(4.4).

By Lemma A.6 there exists a constant c such that

$$\begin{cases} |\tau_0 u_k|_{H^{1/2}(\Gamma)} \le c, \\ |\tau_1 u_k|_{H^{1/2}(\Gamma)^*} \le c, \\ |a_k|_{H^1(0,2)} \le c, \end{cases} \tag{4.6}$$

for all k. Since $H^1(0,2)$ is compactly embedded in $C(0,2)$ it follows that there exists a subsequence of $\{a_k\}$ again denoted by the same symbol, and $a^* \in H^1(0,2)$ such that $\lim a_k = a^*$ strongly in $C(0,2)$. It is simple to argue that $a^* \in Q_{ad}$. Next it is shown that $\{u_k\}$ is a bounded subset of $H^1(\Omega)$. Consider the linear functional on H^1 given by

$$l(u) = \int_\Gamma u \, d\Gamma / \operatorname{meas}(\Gamma).$$

It determines a decomposition of $H^1(\Omega)$ as a direct sum of $\ker l = \{u : \int_\Gamma u \, d\Gamma = 0\}$ and $\{\lambda \cdot 1 : \lambda \in \mathbf{R}\}$, with 1 the constant function with value 1. Any $u \in H^1(\Omega)$ can be written as $u = u^{(1)} + u^{(2)}$, where $u^{(1)} = \int_\Gamma u d\Gamma / \operatorname{meas} \Gamma$ and $l(u_2) = 0$. Any u_k is decomposed as

$$u_k = u_k^{(1)} + u_k^{(2)},$$

where

$$u_k^{(1)} = \int_\Gamma u_k \, d\Gamma / \operatorname{meas}(\Gamma), \quad \int_\Gamma u_k^{(2)} \, d\Gamma = 0, \quad \tau_1 u_k = \tau_1 u_k^{(2)}.$$

Since $e(a_k, u_k) = 0$ for all k, we have

$$-\Delta u_k^{(2)} = f_{a_k} \quad \text{in } L^2(\Omega),$$

and therefore by Green's formula (Lemma A.3),

$$\langle f_{a_k}, v\rangle_{L^2(\Omega)} = \langle \nabla u_k^{(2)}, \nabla v\rangle_{L^2(\Omega)} - \langle \tau_1 u_k^{(2)}, v\rangle_{H^{1/2}(\Gamma)^*, H^{1/2}(\Gamma)} \quad \text{for all } v \in H^1. \tag{4.7}$$

Taking $v = u_k^{(2)}$ in (4.7) gives

$$|\nabla u_k^{(2)}|^2_{L^2(\Omega)} \le |\tau_1 u_k^{(2)}|_{H^{1/2}(\Gamma)^*} |u_k^{(2)}|_{H^{1/2}(\Gamma)} + |f_{a_k}|_{L^2(\Omega)} |u_k^{(2)}|_{L^2(\Omega)},$$

and Lemma A.6 and (4.6) imply the existence of c such that $|\nabla u_k^{(2)}|_{L^2(\Omega)} \le c$. From Lemma A.6(iv) we further have $|u_k|^2_{H^1(\Omega)} \le \sigma_3^2 [c^2 + |\int_\Gamma u_k d\Gamma|^2]$.

Applying (4.6) once again shows that $\{u_k\}$ is a bounded subset of $H(\Delta, \Omega)$. Hence there exists a subsequence, again denoted by $\{u_k\}$, and $u^* \in H(\Delta, \Omega)$, such that

$$\begin{cases} u_k \rightharpoonup u^* & \text{in } H^1(\Omega), \\ \tau_0 u_k \rightharpoonup \tau_0 u^* & \text{in } H^{1/2}(\Gamma), \\ \tau_1 u_k \rightharpoonup \tau_1 u^* & \text{in } H^{1/2}(\Gamma)^*. \end{cases} \tag{4.8}$$

For every $w \in H^1(\Omega)$ we find

$$\begin{aligned} \langle \Delta u^* + f_{a^*}, w\rangle_{L^2(\Omega)} &= \langle \Delta(u^* - u_k), w\rangle_{L^2(\Omega)} + \langle f_{a^*} - f_{a_k}, w\rangle_{L^2(\Omega)} \\ &= -\langle \nabla(u^* - u_k), \nabla w\rangle_{L^2(\Omega)} + \langle \tau_1(u^* - u_k), w\rangle_{H^{1/2}(\Gamma)^*, H^{1/2}(\Gamma)} \\ &\quad + \langle f_{a^*} - f_{a_k}, w\rangle_{L^2(\Omega)}. \end{aligned}$$

It is simple to argue that $\lim\langle f_{a_k} - f_{a^*}, w\rangle_{L^2(\Omega)} = 0$. This, together with (4.8), implies that $-\Delta u^* + f_{a^*} = 0$ in $H^1(\Omega)^*$, and therefore $e(a^*, u^*) = 0$. Weak lower semicontinuity of norms, the fact that bounded linear operators map weakly convergent sequences into weakly convergent sequences, and (4.8) finally imply that (a^*, u^*) is a minimizer of J. This proves (i).

(ii) Let k be the linear functional on $H(\Delta, \Omega)$ given by

$$k(u) = \int_\Omega \mathcal{D}(\tau_0 u)\, d\Omega \,/\mathrm{meas}(\Omega).$$

It determines a decomposition of $H(\Delta, \Omega)$ as the direct sum of $\ker k = \{u : \int_\Omega \mathcal{D}(\tau_0 u) d\Omega = 0\}$ and $\{\lambda \cdot 1 : \lambda \in \mathbf{R}\}$. Any $u \in H(\Delta, \Omega)$ can be represented as $u = u_1 + u_2$, where

$$u_1 = \int_\Omega \mathcal{D}(\tau_0 u) d\Omega / \mathrm{meas}(\Omega) \quad \text{and} \quad \int_\Omega \mathcal{D}(\tau_0 u_2) d\Omega = 0. \tag{4.9}$$

For any $u = u_1 + u_2$ we obtain,

$$\begin{aligned} |\mathcal{D}(\tau_0 u - z_1)|^2_{H^1(\Omega)} &= |\mathcal{D}(\tau_0 u_2 - z_1)|^2_{H^1(\Omega)} + |u_1|^2_{H^1(\Omega)} + 2\langle \mathcal{D}(\tau_0 u_2 - z_1), u_1\rangle_{L^2(\Omega)} \\ &= |\mathcal{D}(\tau_0 u_2 - z_1)|^2_{H^1(\Omega)} + |u_1|^2 \,\mathrm{meas}(\Omega) - |2u_1|\langle \mathcal{D}z_1, 1\rangle_{L^2(\Omega)}, \end{aligned}$$

where (4.9) was used. We find for $(a, u) \in H^1(0,2) \times H(\Delta, \Omega)$,

$$J(a, u) = J_1(a, u_1) + J_2(a, u_2),$$

where

$$J_1(a, u_1) = \frac{1}{2}|u_1|^2 \,\mathrm{meas}(\Omega) - u_1\langle \mathcal{D}z_1, 1\rangle_{L^2(\Omega)} + \frac{1}{2}\beta |a_x|_{L^2(0,2)},$$

and

$$J_2(a, u_2) = \frac{1}{2}|\mathcal{D}(\tau_0 u_2 - z_1)|^2_{H^1(\Omega)} + \frac{1}{2}|\mathcal{N}(\tau_1 u_2 - z_2)|^2_{H^1(\Omega)}.$$

Let (a^*, u^*) be a solution of (4.1)–(4.4). We decompose $u^* = u_1^* + u_2^*$, and claim that

$$J_1(a^*, u_1^*) = \min_{u_1 \in \mathbf{R}} J_1(a^*, u_1), \tag{4.10}$$

$$J_2(a^*, u_2^*) = \min_{u_2 \in \mathcal{J}} J_2(a^*, u_2), \tag{4.11}$$

where $\mathcal{J} = \{u_2 \in H^1(\Omega) : \int \mathcal{D}(\tau_0 u_2) d\Omega = 0, e(a^*, u_2) = 0\}$. If (4.10) were false, then $J_1(a^*, u_1^*) > \min_{u_1 \in \mathbf{R}} J_1(a^*, u_1)$, and further

$$\begin{aligned} J(a^*, u^*) = J_1(a^*, u_1^*) + J_2(a^*, u_2^*) &> \min_{u_1 \in \mathbf{R}} J_1(a^*, u_1) + \min_{u_1 \in \mathcal{J}} J_2(a^*, u_2) \\ &= \min_{\substack{u \in H^1 \\ e(a^*, u) = 0}} J(a^*, u) = J(a^*, u^*), \end{aligned}$$

which is a contradiction and proves (4.10). The verification of (4.11) is similar. From a short calculation using (4.10) we deduce that

$$u_1^* = \frac{< \mathcal{D}z_1, 1 >_{L^2(\Omega)}}{\mathrm{meas}(\Omega)},$$

and from (4.11) we have

$$J_2(a^*, u_2^*) \leq J_2(a^*, \tilde{u}_2),$$

where $\tilde{u}_2$ is the unique element in $\mathcal{J} \cap \{\tau_1 u_2 = z_2\}$. Equivalently, $\tilde{u}_2$ is the unique solution in $H^1(\Omega)$ of

$$\left\{ \begin{array}{r} -\Delta \tilde{u}_2 = f_{a^*}, \\ \tau_1 \tilde{u}_2 = z_2, \\ \int_\Omega \mathcal{D}(\tau_0 \tilde{u}_2) d\Omega = 0. \end{array} \right. \tag{4.12}$$

To argue the existence of a unique solution to (4.12), consider

$$\left\{ \begin{array}{r} -\Delta v = f_{a^*}, \\ \tau_1 v = z_2, \\ \int_\Omega v d\Omega = 0. \end{array} \right. \tag{4.13}$$

The constraint $G_1(a^*) = A_1$ implies $\langle 1, f_{a^*} \rangle_{L^2(\Omega)} = \langle 1, -z_2 \rangle_{L^2(\Gamma)}$, which is the compatibility condition for (4.13). It is now well known that (4.13) has a unique solution $v \in H^1(\Omega)$, and it follows that

$$\tilde{u}_2 = v + \gamma, \quad \text{where } \gamma = -\int_\Omega \mathcal{D}(\tau_0 v) d\Omega / \text{ meas}(\Omega)$$

is the desired unique solution in $H(\Delta, \Omega)$ of (4.12).

Let us define

$$\tilde{u} = u_1^* + \tilde{u}_2.$$

We derive the estimate of $u^* - \tilde{u}$ in terms of $\tau_1 u^* - z_2$ next. Let $w = u_2^* - \tilde{u}_2$ and $g = \tau_1 u^* - z_2$, and observe that $\tilde{u}_2 \in \mathcal{J}, u_2^* \in \mathcal{J}$, and $\tau_1 u_2^* = \tau_1 u^*$ in $H^{1/2}(\Gamma)^*$. We find that w is the unique solution in $H(\Delta, \Omega)$ of

$$\left\{ \begin{array}{r} -\Delta w = 0, \\ \tau_1 w = g, \\ \int_\Omega \mathcal{D}(\tau_0 w) d\Omega = 0, \end{array} \right. \tag{4.14}$$

and consequently

$$| \bigtriangledown w|^2_{L^2(\Omega)} \leq \sigma_1 |g|_{H^{1/2}(\Gamma)^*} |w|_{H^1(\Omega)},$$

with σ_1 from Lemma A.6. Moreover, we have the Poincaré-type inequality

$$|v|_{H^1(\Omega)} \leq C| \bigtriangledown v|_{L^2(\Omega)} \quad \text{for all } v \in H^1(\Omega) \quad \text{with } \int_\Omega \mathcal{D}(\tau_0 v) d\Omega = 0.$$

It follows that

$$|w|_{H^1(\Omega)} \leq c|g|_{H^{1/2}(\Gamma)^*},$$

and thus

$$|u^* - \tilde{u}|_{H^1(\Omega)} = |u_2^* - \tilde{u}_2|_{H^1(\Omega)} = |w|_{H^1(\Omega)} \leq c|\tau_1 u^* - z_2|_{H^{1/2}(\Gamma)^*}. \tag{4.15}$$

Finally, observe that

$$J(a^*,u^*) = \frac{1}{2}|\mathcal{D}(\tau_0 u^* - z_1)|^2_{H^1(\Omega)} + \frac{1}{2}|\mathcal{N}(\tau_1 u^* - z_2)|^2_{H^1(\Omega)} + \frac{\beta}{2}|a^*_x|^2_{L^2(0,2)}$$
$$\leq J(a^*,\tilde{u}) = \frac{1}{2}|\mathcal{D}(\tau_0 \tilde{u} - z_1)|^2_{H^1(\Omega)} + \frac{\beta}{2}|a^*_x|^2_{L^2(0,2)}.$$

Together with Lemma A.6 this inequality gives

$$|\tau_1 u^* - z_2|_{H^{1/2}(\Gamma)^*} \leq \sigma_1 |\mathcal{N}(\tau_1 u^* - z_2)|_{H^1(\Omega)} \leq \sigma_1 |\mathcal{D}(\tau_0 \tilde{u} - z_1)|_{H^1(\Omega)}.$$

The asserted bounds in (ii) follow from these estimates.

(iii) The attainability assumption implies the existence of $(\hat{a},\hat{u}) \in Q_{ad} \times H^1(\Omega)$, where $\hat{u}$ satisfies

$$\begin{aligned} -\Delta \hat{u} &= f_{\hat{a}} \quad \text{in } H^1(\Omega)^*, \\ \tau_0 \hat{u} &= z_1 \quad \text{in } H^{1/2}(\Gamma), \\ \tau_1 \hat{u} &= z_2 \quad \text{in } H^{1/2}(\Gamma)^*. \end{aligned} \tag{4.16}$$

Due to the regularity assumptions on the data z_1 and by Lemma A.5, it follows that $\hat{u} \in H^2(\Omega)$. If (4.1)–(4.4) with $\beta = 0$ has another solution (a^*,u^*), then (a^*,u^*) satisfies (4.16) with $(\hat{a},\hat{u})$ replaced by (a^*,u^*). If both a^* and $\hat{a}$ are piecewise $C^{1,1}$ functions, then Theorem 3.1 implies that $\hat{a} = a^*$, and $\hat{u} = u^*$ by (4.16). This is the desired uniqueness in the case of attainability. □

Remark 4.2. The conclusions of Theorem 4.1 (i) and (iii) remain correct without the compatibility conditions. In (ii) of Theorem 4.1, only the first compatibility condition $G_1(a) = A_1$ is used.

5. Existence of a Lagrange multiplier. In this section we discuss the existence and the properties of a Langrange multiplier associated with the constrained optimization problem (4.1)–(4.4). We start with the following lemma.

LEMMA 5.1. *The Fréchet derivative of* $a \to f_a$ *from* $H^1(0,2)$ *to* $H^1(\Omega)^*$ *at* a *in direction* h *is given by*

$$f'_a(a)h = (\rho_2 - \rho_1)(h\tau_a),$$

where $(h\tau_a) \in H^1(\Omega)^*$ *is defined by*

$$\langle h\tau_a, \phi\rangle_{H^1(\Omega)^*,H^1(\Omega)} = \int_0^2 \phi(x,a(x))h(x)dx.$$

Proof. Let $\phi \in H^1(\Omega), \alpha \in \mathbf{R}$, and a and h in $H^1(0,2)$. Then

$$\langle f_{a+\alpha h}, \phi\rangle_{L^2(\Omega)'} = \int_0^2 \left[\rho_2 \int_0^{a+\alpha h} \phi(x,y)dy + \rho_1 \int_{a+\alpha h}^1 \phi(x,y)dy\right]dx,$$

provided that it is well defined. The Gateaux derivative is now easily found to be

$$\begin{aligned} \frac{d}{d\alpha}\langle f_{a+\alpha h}, \phi\rangle|_{\alpha=0} &= (\rho_2 - \rho_1)\int_0^2 \phi(x,a(x))h(x)dx \\ &= \langle(\rho_2 - \rho_1)h\tau_a, \phi\rangle_{H^1(\Omega)^*,H^1(\Omega)}. \end{aligned} \tag{5.1}$$

The mapping $h \to (h\tau_a)$ for $a \in H^1(0,2)$ is an element of $\mathcal{L}(H^1(0,2), H^1(\Omega)^*)$. To verify that (5.1) gives in fact the Fréchet derivative of $a \to f_a$, we need to show the continuity of $a \to (h \to (h\tau_a))$ from $H^1(0,2)$ to $\mathcal{L}(H^1(0,2), H^1(\Omega)^*)$. Let a and $\bar{a}$ be in $H^1(0,2)$. Then

$$\begin{aligned}
\sup_{|h|_{H^1}=1} |(h\tau_a) - (h\tau_{\bar{a}})|_{H^1(\Omega)^*} &= \sup_{\substack{|h|_{H^1}=1 \\ |\phi|_{H^1}=1}} \Big| \int_0^2 [\phi(x,a(x)) - \phi(x,\bar{a}(x))]h(x)dx \Big| \\
&\le c \sup_{|\phi|_{H^1}=1} \int_0^2 \left| \int_{a(x)}^{\bar{a}(x)} \phi_y(x,s)ds \right| dx \\
&\le c \sup_{|\phi|_{H^1}=1} \int_0^2 |\bar{a}(x) - a(x)|^{1/2} \Big(\int_0^1 \phi_y^2(x,y)dy \Big)^{1/2} dx \\
&\le c^2 |\bar{a} - a|^{1/2}_{L^\infty(0,2)} \sqrt{2} \sup_{|\phi|_{H^1}=1} \left[\int_0^2 \int_0^1 \phi_y^2(x,y)dydx \right]^{1/2} \\
&\le \sqrt{2} c^2 |\bar{a} - a|^{1/2}_{L^\infty(0,2)},
\end{aligned}$$

where c is independent of h and ϕ. □

In the previous section the constraint $\Delta u + f_a = 0$ served two purposes: it induced additional regulariy for $u \in H^1(\Omega)$, namely $\Delta u \in L^2$, so that $\tau_1 u$ was well defined, and it enforced the equation itself. For iterative numerical methods, where at each step of the iteration the variables are not required to be feasible (i.e., they do not necessarily satisfy the equality constraint), u must be chosen in a smaller space than $H^1(\Omega)$, so that $\tau_1 u$ is well defined. We may be tempted to choose $H^1(0,2) \times H(\Delta, \Omega)$ as the space for the optimization problem. For the numerical implementation, however, we do not want to exclude using linear finite elements that are not contained in $H(\Delta, \Omega)$. Moreover, due to the fact that the Fréchet derivative of $a \to f_a$ is only in $H^1(\Omega)^*$, we cannot show the existence of a Lagrange multiplier associated with the constraint $e(a,u) = 0$ if u is chosen in $H(\Delta, \Omega)$. We therefore introduce the space $H(\Gamma, \Omega)$ as the closure of the space of all infinitely differentiable functions on Ω such that all the derivatives have continuous extension up to Γ in the norm

$$|u|^2_{H(\Gamma,\Omega)} = |u|^2_{H^1(\Omega)} + |\tau_1 u|^2_{H^{1/2}(\Gamma)^*}$$

(see, e.g., [Az]). Let us point out that due to Lemma A.3, we have $H(\Delta, \Omega) \subset H(\Gamma, \Omega)$ and hence any solution of (4.1)–(4.4) satisfies $(a^*, u^*) \in H^1(0,2) \times H(\Gamma, \Omega)$.

To obtain an understanding of the space $H(\Gamma, \Omega)$, we consider the mapping

$$S : H(\Gamma, \Omega) \to H^1(\Omega) \times H^{1/2}(\Gamma) \times H^{1/2}(\Gamma)^*,$$

defined for an infinitely differentiable function u on Ω with all derivatives having continuous extensions up to Γ by

$$S\ u = (u, \tau_0 u, \tau_1 u),$$

and extended by continuity to any $H(\Gamma, \Omega)$. Then

$$\begin{aligned}
SH(\Gamma,\Omega) = \{(u, u_0, u_1) : u \in H^1, u_0 = \tau_0 u \in H^{1/2}(\Gamma), \\
u_1 \in H^{1/2}(\Gamma)^* \text{ arbitrary}\}
\end{aligned} \tag{5.2}$$

[Az, p. 66]. Thus, to be precise, we should denote any element of $H(\Gamma,\Omega)$ by a pair of elements $(u,u_1)\in H^1(\Omega)\times H^{1/2}(\Gamma)^*$. To avoid additional notation we refrain from doing so, but we stress that for $u\in H(\Gamma,\Omega)$, $\tau_1 u$ is only a symbol denoting the element u_1.

Some additional notation will be used, as follows:

$$X = H^1(0,2)\times H(\Gamma,\Omega),$$

$$C = \{(a,u)\in X : a(x)\in[0,1]\},$$

$$Y = H^1(\Omega)\times \mathbf{R}^4,$$

$$g : X\to Y,$$

$$g(a,u) = \Big(e(a,u), G_1(a)-A_1, G_2(a)-A_2, G_3(a)-A_3, G_4(a)-A_4\Big),$$

where $e : X\to H^1(\Omega)$ is given by

$$e(a,u) = N(\Delta u + f_a).$$

Here $\Delta u\in H^1(\Omega)^*$ must be interpreted as

$$\langle \Delta u, v\rangle_{H^1(\Omega)^*,H^1(\Omega)} = -\langle \nabla u,\nabla v\rangle_{L^2(\Omega)} + \langle \tau_1 u,\tau_0 v\rangle_{L^2(\Gamma)}$$

for all $v\in H^1(\Omega)$. We note that C is a closed convex subset of X.

Problem (4.1)–(4.4) can be equivalently expressed as

$$(P^\beta)\quad \min_{\substack{(a,u)\in C\\ g(a,u)=0}} J(a,u) = \frac{1}{2}|\mathcal{D}(\tau_0 u - z_1)|^2_{H^1(\Omega)} + \frac{1}{2}|\mathcal{N}(\tau_1 u - z_2)|^2_{H^1(\Omega)} + \frac{\beta}{2}|a_x|^2_{L^2(0,2)}.$$

The Fréchet derivatives of J and g will be needed. The functional J is quadratic in a and u and hence the derivative is obvious. The Fréchet derivatives of G_i are simple to calculate and the Fréchet derivative of $a\to f_a$ from $H^1(0,2)$ to $H^1(\Omega)^*$ was calculated in Lemma 5.1.

Henceforth we fix a solution $x^* = (a^*,u^*)\in X$ of $(P^\beta), \beta\geq 0$. A functional $\Lambda^* = (\lambda^*,\mu_1^*,\dots,\mu_4^*)\in Y^*$ is called Lagrange multiplier if

$$J'(x^*) + \Lambda^* g'(x^*)\in C(x^*)^+, \tag{5.3}$$

where

$$C(x^*)^+ = \{\eta\in X^* :< \eta, x - x^* >_{X^*,X}\geq 0,\ \text{for all } x\in C(x^*)\}$$

and

$$C(x^*) = \{\lambda(x-x^*) : x\in C, \lambda\in \mathbf{R}^+\}.$$

Some comments are in order. The primes in (5.3) denote Fréchet derivatives with respect to $x=(a,u)$. Henceforth we identify the functional $\Lambda^*\in Y^*$ with its Riesz

representation $\Lambda^* \in Y$ and analogously for $\eta \in X$. Due to the special form of C and g, (5.3) is equivalent to

$$\begin{cases} J_a(a^*, u^*) + \Lambda^* g_a(a^*, u^*) \in C(a^*)^+, \\ J_u(a^*, u^*) + \Lambda^* g_u(a^*, u^*) = 0, \end{cases} \tag{5.4}$$

where

$$C(a^*)^+ = \{\eta_1 \in H^1(0,2) : \langle \eta_1, a \rangle_{H^1(0,2)} \geq 0 \text{ for all } a \in C(a^*)\}$$

and

$$C(a^*) = \{\lambda(a - a^*) : 0 \leq a(x) \leq 1, \lambda \in \mathbf{R}^+\},$$

and the subscripts with J and g denote partial Fréchet derivatives. If a^* satisfies $0 < a^*(x) < 1$ on $[0,2]$, then $C(a^*)^+ = \{0\}$ and the first equation in (5.4) becomes

$$J_a(a^*, u^*) + \Lambda^* g_a(a^*, u^*) = 0. \tag{5.5}$$

The existence of a Lagrange multiplier is implied by certain regular point conditions for the minimizer (a^*, u^*). Here we employ the condition

$$0 \in \operatorname{int}\{g'(a,u)(C - (a,u))\} \tag{5.6}$$

for $(a,u) \in C$, where the expression on the right-hand side of (5.6) denotes the interior of the set $\{g'(a,u)(c - (a,u)) : c \in C\}$. If (5.6) holds with $(a,u) = (a^*, u^*)$, then it is known to imply the existence of Λ^* satisfying (5.3) (see, e.g., [ZK]). To guarantee (5.6) an additional condition is necessary. To state this condition we first note that

$$(G_1'(a)h, \ldots, G_4'(a)h) = (\langle l_1, h \rangle_{L^2}, \ldots, \langle l_4, h \rangle) \quad \text{for } a \text{ and } h \in H^1(0,2),$$

where

$$l_1(x) = 1, \quad l_2(x) = 2a(x), \quad l_3(x) = x, \quad l_4(x) = 2xa(x).$$

For $a \in H^1$ with $0 \leq a \leq 1$, let

$$\vec{a} = \operatorname{col}(\langle l_1, a \rangle_{L^2}, \ldots, \langle l_4, a \rangle).$$

The condition announced above is given by

$$\begin{gathered} \text{there exist vectors } \{k_j\}_{j=1}^4 \text{ in } C(0,2), \\ \text{such that the matrix } M \in \mathbf{R}^{4\times 4} \text{ defined by} \\ M_{ij} = \langle l_i, k_j \rangle_{L^2(0,2)} \\ \text{is nonsingular and} \\ 0 < b = \sum_{i=1}^4 b_i k_i(x) < 1, \\ \text{with } \{b_i\}_{i=1}^4 \text{ the coordinates of } \vec{b} = M^{-1}\vec{a}. \end{gathered} \tag{5.7}$$

If a is not a linear function and satisfies $0 < a < 1$, then (5.7) is satisfied with $k_i = l_i$ and $b = a$. Further discussion of (5.7) follows after the next theorem.

THEOREM 5.2. *Condition* (5.7) *implies* (5.6). *In particular, if* (5.7) *holds for* a^*, *where* (a^*, u^*) *is a solution of* (P^β), *then there exists an associated Lagrange multiplier* Λ^*.

Proof. The assertion follows from the fact that there exists $\eta > 0$ such that

$$(5.8)\begin{pmatrix} w \\ \vec{r} \end{pmatrix} = \begin{pmatrix} N(\Delta v + (\rho_2 - \rho_1)(h\tau_{a^*})) \\ \langle l_1, h\rangle_{L^2} \\ \vdots \\ \langle l_4, h\rangle_{L^2} \end{pmatrix} - \begin{pmatrix} N(\Delta u^* + (\rho_2 - \rho_1)(a^*\tau_{a^*})) \\ \\ \vec{a}^* \\ \\ \end{pmatrix}$$

has a solution $(h, v) \in C$ for all $(w, \vec{r}) \in Y = H^1(\Omega) \times \mathbf{R}^4$, provided that $|\vec{r}|_{\mathbf{R}^4} \leq \eta$. We verify below the existence of a solution $h \in H^1(0, 2)$ to

$$(5.9) \qquad \vec{r} = \begin{pmatrix} \langle l_1, h\rangle_{L^2} \\ \vdots \\ \langle l_4, h\rangle_{L^2} \end{pmatrix} - \vec{a}^*$$

for all sufficiently small $\vec{r} \in \mathbf{R}^4$. Once h is determined, we turn to the first equation in (5.8), which has to be solved for $v \in H(\Gamma, \Omega)$.

This equation is equivalent to solving

$$(5.10) \qquad w - Nf = N(\Delta v)$$

for $v \in H(\Gamma, \Omega)$, where $f \in H^1(\Omega)^*$ is given by $f = (\rho_2 - \rho_1)(h - a^*)\tau_a^* - \Delta u^*$. Since N is an isomorphism from $H^1(\Omega)$ to $H^1(\Omega)^*$, there exists a unique $\tilde{v} \in H^1(\Omega)^*$ such that

$$w - Nf = N\tilde{v},$$

and it suffices to solve $\Delta v = \tilde{v}$ for $v \in H(\Gamma, \Omega)$. Let g be any element in $L^2(\Gamma)$ satisfying

$$\langle g, 1\rangle_{L^2(\Gamma)} = \langle \tilde{v}, 1\rangle_{H^1(\Omega)^*, H^1(\Omega)}.$$

Then there exists $v \in H(\Gamma, \Omega)$ satisfying

$$-\langle \nabla v, \nabla \varphi\rangle + \langle g, \varphi\rangle_{L^2(\Gamma)} = \langle \tilde{v}, \varphi\rangle_{H^1(\Omega)^*, H^1(\Omega)}$$

for all $\varphi \in H^1(\Omega)$ (c.f. [Az, p. 62]), and an element v satisfying the desired properties is found.

Returning to (5.9) we look for a solution of the form $h = \sum_{j=1}^4 h_j k_j$. Inserting this expression in (5.9) gives

$$\vec{h} = M^{-1}(\vec{a} + \vec{r}), \quad \text{with } \vec{h} = \operatorname{col}(h_1, \ldots, h_4).$$

Assumption (5.7) implies the existence of $\tilde{\eta} > 0$ such that $0 \leq h \leq 1$ for all $\vec{r} \in \mathbf{R}^4$ with $|\vec{r}|_{\mathbf{R}^4} \leq \eta$. □

The following result summarizes different situations in which (5.7) holds. Some of them require that not all compatibility conditions $G_i(a) = A_i$ are considered simultaneously. As a consequence the corresponding components in the definition of g and Λ^* have to be dropped. No additional notation will be introduced for these cases.

PROPOSITION 5.3. *Let* $a \in H^1(0, 1)$ *satisfy* $0 \leq a \leq 1$.

(i) *If only one of the constraints* $\{G_i(a) - A_i = 0\}_{i=1}^4$ *is considered, then* (5.7) *holds provided that* $a \not\equiv 0$ *and* $a \not\equiv 1$.

(ii) *If only the constraints* $G_i(a) - A_i = 0, i = 1, 3$ *are considered, then* (5.7) *holds if* $0 < \epsilon \leq P_2 a \leq 1 - \epsilon < 1$, *where* P_2 *is the orthogonal projector onto* $\operatorname{span}\{l_1, l_3\}$ *in the* L^2*-norm.*

(iii) *If only the constraints* $G_i(a) - A_i = 0, i = 2, 4$ *are present, then* $0 < \epsilon \leq a \leq 1 - \epsilon < 1$ *implies* (5.7).

(iv) *If all constraints* $\{G_i(a) - A_i = 0\}_{i=1}^4$ *are considered, then* (5.7) *holds if* $0 < \epsilon \leq a \leq 1 - \epsilon < 1$ *and* a *is such that* $\{1, a, x, a \cdot x\}$ *are not linearly dependent as elements in* $H^1(0, 2)$.

Proof. (i) In this case we choose $k_i = 1$. Then $b = \langle l_i, a\rangle_{L^2} / \langle l_i, 1\rangle_{L^2}$ and $M = \langle l_i, 1\rangle_{L^2}$. Since $a \not\equiv 0$ and $a \not\equiv 1$, it follows that $0 < b < 1$ as desired.

(ii) We choose $k_1 = l_1, k_3 = l_3$. Then

$$M = \begin{pmatrix} 2 & 2 \\ 2 & \frac{8}{3} \end{pmatrix}$$

is nonsingular and $b = P_2 a$, which satisfies $0 < \epsilon \leq b \leq 1 - \epsilon$, by assumption.

(iii) In this case we take $k_2 = l_2, k_4 = l_4$. Since $a \not\equiv 0$, the functions l_2 and l_4 are linearly independent and therefore

$$M = \begin{pmatrix} \langle l_2, l_2\rangle \langle l_2, l_4\rangle \\ \langle l_4, l_2\rangle \langle l_4, l_4\rangle \end{pmatrix}$$

is nonsingular. Since $a \in \operatorname{span}\{l_2, l_4\}$, it follows that $b = P_2 a = a$ and by assumption, $0 < \epsilon \leq b \leq 1 - \epsilon < 1$.

(iv) We take $k_i = l_i$ for $i = 1, \ldots, 4$. Due to the linear independence of $\{1, a, x, a \cdot x\}$, the matrix M with $(M)_{ij} = \langle l_i, l_j\rangle$ is nonsingular. Again we find $b = P_2 a = a$ and $0 < \epsilon \leq b \leq 1 - \epsilon < 1$, as desired. □

Remark 5.4. Proposition 5.3 does not cover the case when the first and second compatibility condition are present simultaneously, unless $0 < \epsilon \leq a \leq 1 - \epsilon < 1$. But this situation is not excluded by (5.7), as the following example shows. We consider the first two compatibilities $G_i(a) - A_i = 0, i = 1, 2$ with $a = x/2$. In this case we choose $k_1 = 1, k_2(x) = \cos(\pi/2)x$. We find

$$M = \begin{pmatrix} 2 & 0 \\ 2 & -\frac{8}{\pi^2} \end{pmatrix} \quad \text{and} \quad b = \frac{1}{2} - \frac{\pi^2}{24} \cos \frac{\pi}{2} x,$$

so that (5.7) clearly holds.

We turn to a discussion of the implications of (5.3) for the problem (P^β). The next result will be given under the assumption of existence of a Lagrange multiplier Λ^* for all four compatibility conditions. The modifications for the case where only a subset of these conditions is enforced are straightforward. For convenience of notation we put $G(a) = (G_1(a), \ldots, G_4(a))$.

THEOREM 5.5. *Let* (5.6) *hold and let* $(a^*, u^*) \in X$ *be a solution of* (P^β) *with Lagrange multiplier* $\Lambda^* = (\lambda^*, \mu_1^*, \ldots, \mu_4^*) \in Y$. *Then* λ^* *is the unique solution of*

$$\begin{aligned} \langle \nabla \lambda^*, \nabla v\rangle_{L^2(\Omega)} &= \langle \mathcal{D}(\tau_0 u^* - z_1), \mathcal{D}(\tau_0 v)\rangle_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega) \\ \tau_0 \lambda^* &= \tau_0 \mathcal{N}(z_2 - \tau_1 u^*) \quad \text{in } H^{1/2}(\Gamma). \end{aligned} \tag{5.11}$$

Moreover, there exists $\eta^* \in C(a^*)^+$ *such that*

$$\begin{aligned} &\langle \mu^*, G'(a^*)h\rangle_{\mathbf{R}^4} + (\rho_2 - \rho_1)\langle h\tau_{a^*}, \lambda^*\rangle_{L^2(\Omega)} + \beta\langle a_x^*, h_x\rangle_{L^2(0,2)} \\ &\quad = \langle \eta^*, h\rangle_{H^1(0,2)} \quad \text{for all } h \in H^1(0, 2). \end{aligned} \tag{5.12}$$

If $\beta > 0$, and $\rho_2 \neq \rho_1$ or at least one of the compatibility conditions is present, this implies the regularity property $a^* \in H^2(I)$, where $I = \{x \in [0,2] : 0 < a^*(x) < 1\}$.

Proof. We employ the necessary optimality condition (5.4). It follows that there exists $\eta^* \in C(a^*)^+$ such that

$$\begin{aligned}\beta\langle a_x^*, h_x\rangle_{L^2(0,2)} + (\rho_2 - \rho_1)\langle h\tau_{a^*}, \lambda^*\rangle_{H^1(\Omega)^*,H^1(\Omega)} + \langle \mu^*, G'(a^*)h\rangle_{\mathbf{R}^4} \\ = \langle \eta^*, h\rangle_{H^1(0,2)} \quad \text{for all } h \in H^1(0,2) \quad \text{and}\end{aligned}$$

$$\begin{aligned}\langle \mathcal{D}(\tau_0 u^* - z_1), \mathcal{D}(\tau_0 v)\rangle_{H^1(\Omega)} + \langle \mathcal{N}(\tau_1 u^* - z_2), \mathcal{N}(\tau_1 v)\rangle_{H^1(\Omega)} \\ -\langle \nabla\lambda^*, \nabla v\rangle_{L^2(\Omega)} + \langle \tau_0\lambda^*, \tau_1 v\rangle_{H^{1/2}(\Gamma),H^{1/2}(\Gamma)^*} = 0\end{aligned} \tag{5.13}$$

for all $v \in H(\Gamma, \Omega)$. We remind the reader that in (5.13), $\tau_1 v$ is only the symbol for the second coordinate of $v \in H(\Gamma, \Omega)$, unless v has additional regularity. The first of these equation is (5.12). Using the definition of $\mathcal{N}$ in the second we obtain

$$\begin{aligned}\langle \mathcal{D}(\tau_0 u^* - z_1), \mathcal{D}(\tau_0 v)\rangle_{H^1(\Omega)} - \langle \nabla\lambda^*, \nabla v\rangle_{L^2(\Omega)} \\ + \langle \tau_0\lambda^* + \tau_0\mathcal{N}(\tau_1 u^* - z_2), \tau_1 v\rangle_{H^{1/2}(\Gamma),H^{1/2}(\Gamma)^*} = 0.\end{aligned} \tag{5.14}$$

Recalling (5.2) and evaluating (5.14) for v in the set $\{v \in H(\Gamma, \Omega) : \tau_1 v = 0\}$ we find

$$\langle \nabla\lambda^*, \nabla v\rangle_{L^2(\Omega)} = \langle \mathcal{D}(\tau_0 u^* - z_1), \mathcal{D}(\tau_0 v)\rangle_{H^1(\Omega)},$$

for all $v \in H^1(\Omega)$.

The first equality in (5.12) now follows. Reconsidering (5.14) we find $\langle \tau_0\lambda^* + \tau_0\mathcal{N}(\tau_1 u^* - z_1), \tau_1 v\rangle_{H^{1/2}(\Gamma),H^{1/2}(\Gamma)^*} = 0$ for all $\tau_1 v \in H^{1/2}(\Gamma)^*$ and thus (5.11) is verified.

We turn to the proof of the regularity property. Since I is open relative to [0,2], it can be expressed as $I = \bigcup_{j=1}^{\infty} I_j$, with I_j pairwise disjoint intervals open relative to $[0,2]$. On each I_j, the functional η^* is zero, i.e., $\langle \eta^*, h\rangle_{H^1(I_j)} = 0$ for all $h \in H^1(I_j)$. From (5.12) it follows that

$$\mu^* G'(a^*) + (\rho_2 - \rho_1)\lambda^*(\cdot, a^*(\cdot)) = \beta a_{xx}^* \quad \text{in } H^1(I_j)^*. \tag{5.15}$$

Since the left-hand side of (5.15) is in $L^2(I_j)$ and $\beta > 0$ by assumption, it follows that a_{xx}^* can be identified with an element in $L^2(I_j)$ and $a^* \in H^2(I_j)$. It follows that $a^* \in H^2(I)$ and the proof is finished. □

Remark 5.6. From (5.13) we deduce that $|\lambda^*|_{H^1}$ is small provided that $|\mathcal{D}(\tau_0 u^* - z_1)|_{H^1}$ and $|\mathcal{N}(\tau_1 u^* - z_2)|_{H^1}$ are small. Moreover $\lambda^* = 0$, if $\tau_0 u^* = z_1$ and $\tau_1 u^* = z_2$, which can be expected to hold only in the case that the data (z_1, z_2) are attainable and $\beta = 0$.

6. The augmented Lagrangian algorithm. In this section we describe the algorithm that we propose for the solution of (P^β). Conceptually we can distinguish three stages. In the first stage the equality constraint $g(a,u) = 0$ is eliminated from the set of explicit constraints and only the simple constraints $0 \leq a(x) \leq 1$ are left as explicit constraints. The second and third stages consist of discretizing the resulting infinite-dimensional optimization problem with only simple constraints and of choosing an appropriate algorithm to solve the finite-dimensional optimization problems. While the specific choices that are made in these last two stages as well as the order in which these two stages are carried out are important, we do not study these aspects

here. The focus in this and the following section is on the analysis of the first stage. An augmented Lagrangian functional is used to eliminate the constraint $g(a,u)=0$. The resulting optimization problems are quadratic in u for fixed a and "completely" nonlinear in a. A conjugate gradient algorithm is used to solve the optimization problems. The discretization of the variables (a,u) is carried out by finite elements.

The augmented Lagrangian function

$$\mathcal{L}_c : X \times Y = H^1(0,2) \times H(\Gamma,\Omega) \times H^1(\Omega) \times \mathbf{R}^4 \to \mathbf{R} \quad \text{is defined by}$$

$$\mathcal{L}_c(x,\Lambda) = J(x) + \langle \Lambda, g(x)\rangle_Y + \frac{c}{2}|g(x)|_Y^2,$$

where $c \in \mathbf{R}^+$. In this and the following section it is convenient to use interchangably the notation (a,u) and x to denote an element in X. Let us assume that x^* is a local solution of (P^β) with associated Lagrange multiplier Λ^*. The possibility of eliminating the constraint $g(x)=0$ from the explicit constraints in (P^β) relies on the following augmentability condition [H], [IK1].

(6.1)
$$\begin{cases} \text{There exist constants } \sigma > 0, r > 0, \text{ and } \bar{c} \geq 0 \text{ such that} \\ \mathcal{L}_c(x,\Lambda^*) - J(x^*) \geq \sigma |x - x^*|_X^2 \\ \text{for all } c \geq \bar{c}, \text{ and } x \in B(x^*,r) = \{x=(a,u) : |x-x^*|_X \leq r, \\ 0 \leq a(x) \leq 1\}. \end{cases}$$

Let us note that (6.1) implies that

$$J(x) \geq J(x^*) + \sigma |x - x^*|_X^2$$

for all $x \in B(x^*,r)$, which also satisfies $g(x)=0$. Thus x^* is a strict local minimum. Condition (6.1) will be analyzed in §7. If Λ^* was knwon, then based on (6.1) it would be natural to minimize $\mathcal{L}_c(x,\Lambda^*)$, subject to $0 \leq a \leq 1$. But since Λ^* is unknown it needs to be approximated as part of the numerical procedure to solve (P^β). This leads to the augmented Lagrangian algorithm, which we describe next.

Augmented Lagrangian algorithm.

(i) CHOOSE $\Lambda_0, \beta, \{c_n\}_{n=1}^\infty$ with $\bar{c} < c_1 \leq c_2, \ldots$

(ii) SET $n=1$

REPEAT

(iii) (P_n^β) minimize $\mathcal{L}_{c_n}(a,u,\Lambda_{n-1})$ subject to $(a,u) \in X$ and $0 \leq a \leq 1$ to obtain a solution (a_n,u_n).

(iv) UPDATE

$\Lambda_n = \Lambda_{n-1} + (c_n - \bar{c})g(a_n,u_n)$

(v) SET $n = n+1$.

Remark 6.1. (i) In our calculations we took $\Lambda_0 = (\lambda_0, \mu_{1,0}, \ldots, \mu_{4,0}) = 0$. The choice of $\lambda_0 = 0$ is suggested by (5.11), which implies that λ^* should be small if the data are almost attainable and if β is small.

(ii) For the computations in this paper we took $\{c_n\}$ to be a constant sequence and we chose β heuristically; compare, however, [IK3].

(iii) It is not difficult to argue that the problems (P_n^β) in step (iii) of the algorithm have a solution. To solve these problems numerically, a conjugate gradient algorithm

was used. In view of Lemma 5.1 it is simple to calculate the analytical gradient of $\mathcal{L}_{c_n}(a, u, \Lambda_{n-1})$ with respect to (a, u). Numerically, an alternate direction method was used. First the variable a was fixed and the quadratic problem was solved for u, then u was fixed at the value obtained and the nonlinear function was minimized with respect to a. We can repeat these steps several times before updating Λ.

(iv) As a startup value for the numerical solution of (P_n^β) the solution (a_{n-1}, u_{n-1}), $n = 2, 3, \ldots$ was taken. For $n = 1$ a very good choice for a startup value a_1^0 is given as a solution to the compatibility conditions $G_i(a) = A_i$.

(v) For some calculations the regularization term $\beta|a_x|^2_{L^2(0,2)}$ in (P_n^β) was replaced by $\beta|a_x - \tilde{a}_x|^2_{L^2(0,2)}$, where $\tilde{a}$ represents some a priori guess to a^*. A good choice for $\tilde{a}$ is again given by a function which satisfies the compatibility conditions.

(vi) The update rule in (iv) of the augmented Lagrangian algorithm realizes a steepest ascent rule for the dual optimization problem; compare [B], [IK1].

The convergence analysis is based on (6.1). A priori it is not guaranteed that the iterates x_n are contained in $B(x^*, r)$, which is the ball in which (6.1) is applicable. Therefore the auxilary constrained problems

$$(P_{n,c}^\beta) \text{ minimize } \mathcal{L}_{c_n}(a, u, \Lambda_{n-1}) \quad \text{subject to } (a, u) \in B(x^*, r)$$

are introduced. The following result guarantees that if in the augmented Lagrangian algorithm (P_n^β) is replaced by $(P_{n,c}^\beta)$, then the solutions $x_n \in \text{ int } B(x^*, r)$ either for all sufficiently large n, or for all n if $|\Lambda_0 - \Lambda^*|_Y$ is sufficiently small or c_1 is sufficiently large, i.e., the constraint $|x - x^*|_X \leq r$ is not active in these cases. Moreover it asserts convergence of $\{x_n\}$ and boundedness of the Lagrange multipliers $\{\lambda_n\}$.

THEOREM 6.2. *Let* (6.1) *hold. Then for every* $n = 1, 2, \ldots$, *there exists a solution* x_n *to* $(P_{n,c}^\beta)$. *Moreover, there exists* n_0 *such that* $x_n \in \text{ int } B(x^*, r)$ *for all* $n \geq n_0$. *Alternatively, if* $|\Lambda_0 - \Lambda^*|_Y$ *is sufficiently small or* c_n *are taken sufficiently large, then* x_n *is a solution to* (P_n^β) *in* int $B(x^*, r)$ *for every* $n = 1, 2, \ldots$. *Assuming that* x_n *are solutions of* (P_n^β) *in* $B(x^*, r)$ *we find for* $n = 1, 2, \ldots$,

$$\sigma|x_n - x^*|_X^2 + \frac{1}{2\sigma_n}|\Lambda_n - \Lambda^*|_Y^2 \leq \frac{1}{2\sigma_n}|\Lambda_{n-1} - \Lambda^*|_Y^2 \tag{6.2}$$

and

$$\sum_{n=1}^{\infty} \sigma_n |x_n - x^*|_X^2 \leq \frac{1}{2\sigma}|\Lambda^0 - \Lambda^*|_Y^2, \tag{6.3}$$

where $\sigma_n = c_n - \bar{c}$.

For the proof we refer to [IK1].

7. The augmentability condition. This section is devoted to the study of the augmentability condition (6.1) by means of a second-order analysis of $\mathcal{L}_c(x, \Lambda)$ at (x^*, Λ^*). The section is organized as follows. First a smoothing of the mapping $a \to f_a$ is introduced which will allow us to argue that a modified augmented Lagrangian functional which for convenience we again denote by $\mathcal{L}_c$, is twice continuously Fréchet differentiable with respect to x. Then Taylor's formula is used to verify the augmentability condition (6.1) under the assumption that the second Fréchet derivative $\mathcal{L}_c''(x^*, \Lambda^*)$ with respect to x at (x^*, Λ^*) is positive definite on the kernel of the linearized constraints. The latter condition is referred to as the second-order sufficient optimality condition, and is analyzed at the end of this section. At present

we cannot verify the augmentability condition for the original problem without the smoothing of $a \to f_a$.

The smoothing of $a \to f_a$ is defined next. The domain Ω and the right-hand side f_a are modified as follows:

$$\Omega = \left\{(x,y) : x \in [0,2], -\frac{\alpha}{2} \le y \le \frac{\alpha}{2} + 1\right\}, \tag{7.1}$$

where $\alpha > 0$ and

$$\tilde{f}_a(x,y) = \begin{cases} \rho_1 & \text{if } y > a(x) + \dfrac{\alpha}{2}, \\ \dfrac{\rho_1 - \rho_2}{\alpha}\left(y - a(x) + \dfrac{\alpha}{2}\right) + \rho_2 & \text{if } a(x) - \dfrac{\alpha}{2} \le y \le a(x) + \dfrac{\alpha}{2}, \\ \rho_2 & \text{if } y < a(x) - \dfrac{\alpha}{2}. \end{cases} \tag{7.2}$$

It is obvious that the results of the previous sections remain correct with f_a replaced by $\tilde{f}_a$.

Let us recall that

$$\mathcal{L}_c(x,\Lambda) = J(x) + \langle \Lambda, g(x)\rangle_Y + \frac{c}{2}|g(x)|_Y^2,$$

where g is understood with f_a replaced by $\tilde{f}_a$. Since $x \to J(x)$ is clearly twice continuously Fréchet differentiable we turn to the differentiability properties of

$$x \to \langle \Lambda, g(x)\rangle_Y + \frac{c}{2}|g(x)|_Y^2.$$

We note that $a \to \mu_i(G_i(a) - A_i) + (c/2)|G_i(a) - A_i|^2$ with $\mu_i \in \mathbf{R}$ is twice continuously Fréchet differentiable and therefore it suffices to consider

$$\tilde{G} : (a,u) = x \to \langle \lambda, N(\Delta u + \tilde{f}_a)\rangle_{H^1(\Omega)} + \frac{c}{2}|N(\Delta u + \tilde{f}_a)|_{H^1(\Omega)}^2, \tag{7.3}$$

where $\lambda \in H^1(\Omega)$.

Since $\tilde{G}$ is quadratic with respect to u, and since there are no mixed terms so that $\tilde{G}_{au} = 0$, we only specify the second Fréchet derivative of $\tilde{G}$ with respect to a. The first Fréchet derivative of $a \to \tilde{f}_a$ from $H^1(0,2)$ to $L^2(0,2)$ is given by

$$\tilde{f}'(a)h = \frac{\rho_2 - \rho_1}{\alpha}\tilde{\chi}_a h,$$

where $\tilde{\chi}_a$ is the characteristic function of the subdomain $\{(x,y) : 0 < x < 2, a(x) - \alpha/2 < y < a(x) + \alpha/2\}$. In view of Lemma 5.1 we put $\tilde{\tau}_a = (1/\alpha)\tilde{\chi}_a$ and we note that $(h\tau_a)$ is the limit of $h\tilde{\tau}_a$ as $\alpha \to 0$. In view of the proof of Lemma 5.1 we further find

$$\begin{aligned} \tilde{G}_{aa}(x)(h,\tilde{h}) &= \frac{\rho_2 - \rho_1}{\alpha}\langle h\tilde{h}(\tau_{a+\frac{\alpha}{2}} - \tau_{a-\frac{\alpha}{2}}), \lambda\rangle_{H^1(\Omega)^*, H^1\Omega)} \\ &\quad + c(\rho_2 - \rho_1)^2\langle N(h\tilde{\tau}_a), N(\tilde{h}\tilde{\tau}_a)\rangle_{H^1(\Omega)} \\ &= \frac{\rho_2 - \rho_1}{\alpha}\int_0^2 \left(\lambda\left(x, a(x) + \frac{\alpha}{2}\right) - \lambda\left(x, a(x) - \frac{\alpha}{2}\right)\right) h(x)\tilde{h}(x)dx \\ &\quad + c(\rho_2 - \rho_1)^2\langle N(h\tilde{\tau}_a), N(\tilde{h}\tilde{\tau}_a)\rangle_{H^1(\Omega)}, \end{aligned}$$

where $h, \tilde{h} \in H^1(0,2)$. We have thus shown that $\mathcal{L}_c(x,\lambda)$ is twice continuously Fréchet differentiable with respect to x if f_a is replaced by $\tilde{f}_a$.

Let us recall a lemma on positive definite bilinear forms. For a proof, see [H], for example.

LEMMA 7.1. *Let A be a bounded self-adjoint linear operator from a Hilbert space H into itself and let*

$$\langle Ax, x\rangle_H \geq m|x|_H^2 \quad \textit{for some } m > 0 \quad \textit{and all } x \in \ker C,$$

where C is a bounded linear operator from H onto a Hilbert space H_1. Then for every $\gamma \in (0, m)$ there exists $R > 0$ such that

$$\langle (A + KC^*C)x, x\rangle_H = \gamma|x|_H^2 \quad \textit{for all } x \in H \quad \textit{and } K \geq R.$$

Throughout this section $x^* = (a^*, u^*)$ denotes a solution of (P^β) with Lagrange multiplier Λ^*. It will be useful to observe that

$$\begin{aligned} \mathcal{L}_c''(x^*, \Lambda^*)(y, y) = J''(x^*)(y, y) + \langle \Lambda^*, g''(x^*)(y, y)\rangle_Y \\ + c|g'(x^*)y|_Y^2, \end{aligned} \tag{7.4}$$

where $y \in X$. The following second-order sufficient optimality condition will be used.

There exists a constant $K > 0$ such that

$$J''(x^*)(y, y) + \langle \Lambda^*, g''(x^*)(y, y)\rangle_Y \geq K|y|_X^2 \quad \text{for all } y \in \ker g'(x^*). \tag{7.5}$$

THEOREM 7.2. *Let* (7.5) *hold and assume that a^* is not a linear function. Then the augmentability condition* (6.1) *is satisfied.*

Proof. Due to Proposition 7.1, Taylor's theorem is applicable to $x \to \mathcal{L}_c(x, \Lambda^*)$. We find

$$\begin{aligned} \mathcal{L}_c(x, \Lambda^*) &= \mathcal{L}_c(x^*, \Lambda^*) + \mathcal{L}_c'(x^*, \Lambda^*)(x - x^*) + \mathcal{L}_c''(x^*, \Lambda^*)(x - x^*)^2 \\ &\quad + o(|x - x^*|_X^2) \\ &= J(x^*) + J'(x^*)(x - x^*) + < \Lambda^*, g'(x^*)(x - x^*) >_Y + \mathcal{L}_c''(x^*, \Lambda^*)(x - x^*)^2 \\ &\quad + o(|x - x^*|_X^2) \\ &\geq J(x^*) + \mathcal{L}_c''(x^*, \Lambda^*)(x - x^*)^2 + o(|x - x^*|_X^2), \end{aligned} \tag{7.6}$$

for all $x = (a, u)$ with $0 \leq a \leq 1$. In the last estimate we used (5.4). Lemma 7.1 will be employed next to argue positivity of $\mathcal{L}_c''(x^*, \Lambda^*)$. The operator $g'(x^*) : X \to Y \times \mathbf{R}^4$ is given by

$$g'(x^*)(h, v) = \begin{pmatrix} N(\Delta v + (\rho_2 - \rho_1)(h\tilde{\tau}_{a^*})) \\ \langle l_1, h\rangle_{L^2(0,2)} \\ \vdots \\ \langle l_4, h\rangle_{L^2(0,2)} \end{pmatrix}.$$

It is surjective since $\{l_i\}_{i=1}^4$ are linearely independent as a consequence of the assumption that a^* is not a linear function. In view of (7.4), $\mathcal{L}_c''(x^*, \Lambda^*)$ can be expressed as

$$\mathcal{L}_c''(x^*, \Lambda^*)(y, y) = \langle Ay, y\rangle_X + c|Cy|_Y^2, \qquad y \in X,$$

where $A \in \mathcal{L}(X), C \in \mathcal{L}(X,Y)$, are given by

$$\langle Ay, y\rangle_X = J''(x^*)(y,y) + \langle \Lambda^*, g''(x^*)(y,y)\rangle_Y$$

and $Cy = g'(x^*)y$ for $y \in X$. Let $\tilde{\sigma} \in (0,K)$, with K from (7.5). Then by Lemma 7.1 there exists $\bar{c}$ such that

$$\mathcal{L}''_{\bar{c}}(x^*,\Lambda^*)(y,y) \geq \tilde{\sigma}|y|_X^2 \quad \text{for all } y \in X.$$

Now (7.6) can be used to assert that for every $\sigma \in (0,\tilde{\sigma})$ there exists $r > 0$ such that

$$\mathcal{L}_c(x,\Lambda^*) \geq \mathcal{L}_{\bar{c}}(x,\Lambda^*) \geq J(x^*) + \sigma|x - x^*|_X^2$$

for all $c \geq \bar{c}$ and all $x \in B(x^*,r)$. This is the desired augmentability condition (6.1). □

Remark 7.3. The assumption that a^* is not a linear function is only required if the compatibility conditions $G_i(a) = A_i, i = 2,4$ are present. It can be replaced by the requirement that $\{l_i\}_{i\in I}$ are linearly independent, where I denotes the set of compatibility conditions which are considered.

Remark 7.4. The smoothing of f_a by $\tilde{f}_a$ guarantees the existence of the second Fréchet derivative of $\mathcal{L}_c(x,\Lambda)$ with respect to x. Let us briefly comment on the second derivative of $\mathcal{L}_c(x,\Lambda^*)$ with respect to x if f_a were not replaced by $\tilde{f}_a$. The only difference arises in the term $\tilde{G}$, which is then given by

$$\tilde{G} : x \to \langle \lambda^*, N(\Delta u + f_a)\rangle_{H^1} + \frac{c}{2}|N(\Delta u + f_a)|_{H^1}^2.$$

In this case, $\tilde{G}$ is not twice continuously differentiable in general, however, we can show that $\tilde{G}$ has a second Fréchet derivative at (x^*,Λ^*) provided that λ^* satisfies additional regularity properties, for instance $\lambda^*_{yy} \in L^p(\Omega)$ for some $p > 1$. In this case

$$\tilde{G}_{aa}(a^*)(h,\tilde{h}) = (\rho_2 - \rho_1)\int_0^2 \lambda^*_y(x,a^*(x))h(x)\tilde{h}(x)dx + c(\rho_2 - \rho_1)^2\langle N(h\tau_{a^*}), N(\tilde{h}\tau_{a^*})\rangle_{H^1(\Omega)}.$$

We now turn to the second order sufficient optimality condition (7.5), which will be verified to hold for various combinations of the compatibility conditions $G_i(a) = A_i$. The estimate will depend on the regularization parameter β, and to emphazise this fact we change the notation from (x^*,Λ^*,μ^*) to $(x^\beta,\Lambda^\beta,\mu^\beta)$. So let x^β denote a solution of $(P^\beta), \beta > 0$ and let $\Lambda^\beta = (\lambda^\beta,\mu_1^\beta,\dots,\mu_4^\beta)$ denote an associated Lagrange multiplier. If only a subset of all four compatibilities is considered, then the corresponding coordinates μ_i^β are deleted from Λ^β; see also the comment before Proposition 5.3 in this respect. We give some preparatory lemmas.

LEMMA 7.5. *There exists a constant σ_4 independent of β such that*

$$|\lambda^\beta|_{H^1} \leq \sigma_4(|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*}). \tag{7.7}$$

If, moreover,

$$\sigma_0 \leq a^\beta(x) \leq 1 - \sigma_0 \quad \text{for some constant } \sigma_0 > 0 \tag{7.8}$$

and only $G_2(a) = A_2$ *is present in* $g(a,u) = 0$, *then*

$$|\mu_2^\beta| \le \sigma_4(|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*}). \tag{7.9}$$

The same result holds with $G_2(a) = A_2$ *replaced by* $G_4(a) = A_4$.

Proof. We employ (5.11) of Theorem 5.5 and Lemma A.6. We find

$$|\nabla\lambda^\beta|^2_{L^2(\Omega)} \le \sigma_1\sigma_3^2|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)}|\lambda^\beta|_{H^1(\Omega)}$$

and

$$\begin{aligned}\left|\int_\Gamma \tau_0\lambda^\beta d\Gamma\right| &\le \text{ const}|\tau_0\mathcal{N}(z_2 - \tau_1 u^\beta)|_{H^{1/2}(\Gamma)} \\ &\le \text{ const } \sigma_1\sigma_3|\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*},\end{aligned}$$

where const is independent of β. From Lemma A.6 we find

$$\begin{aligned}|\lambda^\beta|^2_{H^1(\Omega)} &\le \sigma_2^2\left[|\nabla\lambda^\beta|^2_{L^2(\Omega)} + \left|\int_\Gamma \tau_0\lambda^\beta d\Gamma\right|^2\right] \\ &\le \sigma_1\sigma_2^2\sigma_3^2(|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)}|\lambda^\beta|_{H^1(\Omega)} \\ &\qquad + \text{ const}^2\cdot\sigma_1|\tau_1 u^\beta - z_2|^2_{H^{1/2}(\Gamma)^*}) \\ &\le \sigma_1\sigma_2^2\sigma_3^2\left[\frac{1}{2}\left(\tilde\alpha^2|\tau_0 u^\beta - z_1|^2_{H^{1/2}(\Gamma)} + \frac{1}{\tilde\alpha^2}|\lambda^\beta|^2_{H^1(\Omega)}\right)\right. \\ &\qquad \left. + \text{ const}^2\cdot\sigma_1|\tau_1 u^\beta - z_2|^2_{H^{1/2}(\Gamma)}\right]\end{aligned}$$

for every $\tilde\alpha > 0$. Take $\tilde\alpha^2 = \sigma_1\sigma_2^2\sigma_3^2$, then

$$|\lambda^\beta|^2_{H^1(\Omega)} \le \sigma_1^2\sigma_2^2\sigma_3^2[\sigma_2^2\sigma_3^2|\tau_0 u^\beta - z_1|^2_{H^{1/2}(\Gamma)} + 2\text{ const}^2\cdot|\tau_1 u^\beta - z_2|^2_{H^{1/2}(\Gamma)^*}]$$

and we obtain

$$|\lambda^\beta|_{H^1(\Omega)} \le \sigma_1\sigma_2\sigma_3(\sigma_2\sigma_3|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + \sqrt{2}\text{ const}^2\cdot|\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*}).$$

To verify the second part of the lemma, let (7.8) hold and let $G_2(a) = A_2$ be the only compatibility condition that is considered. In this case $\eta^\beta = 0$ and from (5.12) it follows that

$$2\mu_2^\beta\langle a^\beta, h\rangle_{L^2(0,2)} + (\rho_2 - \rho_1)\langle h\tilde\tau_{a^\beta}, \lambda^\beta\rangle_{L^2(\Omega)} + \beta\langle a_x^\beta, h_x\rangle_{L^2(0,2)} = 0$$

for all $h \in H^1(0,2)$. Coosing $h = 1$ we obtain

$$\mu_2^\beta = \frac{\rho_1 - \rho_2}{2\langle a^\beta, 1\rangle_{L^2(0,2)}}\langle\tilde\tau_{a^\beta}, \lambda^\beta\rangle_{L^2(\Omega)},$$

and therefore

$$|\mu_2^\beta| \le \frac{|\rho_1 - \rho_2|^2}{4\sigma_0\alpha}|\lambda^\beta|_{L^1(\Omega)}.$$

In view of (7.7), σ_4 can now be modified such that (7.7) and (7.9) hold simultanously. The proof for $G_4(a) = A_4$ is analogous. □

LEMMA 7.6. *There exists* $\sigma_5 > 0$ *such that*

$$|h|_{H^1(0,2)} \leq \sigma_5 |h_x|_{L^2(0,2)}$$

for all $h \in \ker G_i'(a^\beta)$, *where* $i \in \{1,3\}$. *If* $a^\beta(x) \geq \sigma_0 > 0$, *then* σ_5 *can be chosen independently of* β *such that*

$$|h|_{H^1(0,2)} \leq \sigma_5 |h_x|_{L^2(0,2)} \tag{7.10}$$

for all $h \in \ker G_i'(a^\beta)$, *where* $i \in \{2,4\}$.

The proof is left to the reader.

Remark 7.7. The condition $a^\beta(x) \geq \sigma_0 > 0$ is sufficient, but not necessary, for (7.10) to hold. If, for example, the data (z_1, z_2) are attainable by a unique pair $(\hat{a}, \hat{u})$, then $a^\beta \to \hat{a}$ in $H^1(0,2)$ (see [CK]) and there exists $\hat{\beta} > 0$ such that (7.10) holds for all $\beta \in [0, \hat{\beta}]$.

LEMMA 7.8. *An element* $(h, v) \in \ker(g'(x^\beta))$ *with* $x^\beta = (a^\beta, u^\beta)$ *is characterized by*

$$\begin{aligned} \langle \nabla v, \nabla \varphi \rangle_{L^2(\Omega)} &= \langle \tau_1 v, \tau_0 \varphi \rangle_{H^{1/2}(\Gamma)^*, H^{1/2}(\Gamma)} \\ &\quad + (\rho_2 - \rho_1) \langle h \tilde{\tau}_{a^\beta}, \varphi \rangle \quad \textit{for all } \varphi \in H^1(\Omega) \end{aligned} \tag{7.11}$$

and

$$\langle l_i(a^\beta), h \rangle_{L^2(0,2)} = 0 \quad \textit{for } i = 1, \ldots, 4.$$

There exists σ_6 *independent of* β *such that*

$$\begin{aligned} |v|^2_{H^1(\Omega)} &\leq \sigma_6 (|\mathcal{D}(\tau_0 v)|^2_{H^1(\Omega)} + |\mathcal{N}(\tau_1 v)|^2_{H^1(\Omega)} + |h|^2_{H^1(0,2)}) \\ &\qquad \textit{for every } (h, v) \in \ker(g'(x^\beta)). \end{aligned} \tag{7.12}$$

Proof. The first part of the lemma follows from a simple calculation and we therefore turn directly to (7.12). Using Lemma A.6(v) we can show that

$$|\tau_1 v|_{H^{1/2}(\Gamma)^*} \leq \sqrt{2} \sigma_3 |\mathcal{N}(\tau_1 v)|_{H^1(\Omega)}, \tag{7.13}$$

and with a similar proof as for Lemma A.6(iv) there exists $\hat{\sigma}_5$ such that

$$|\varphi|^2_{H^1(\Omega)} \leq \hat{\sigma}_5 (|\nabla \varphi|^2_{H^1(\Omega)} + |\mathcal{D}(\tau_0 \varphi)|^2_{H^1(\Omega)}) \tag{7.14}$$

for all $\varphi \in H^1(\Omega)$. From (7.11) we find, by (7.13) and Lemma A.6(i),

$$\begin{aligned} |\nabla v|^2_{L^2(\Omega)} &\leq |\tau_1 v|_{H^{1/2}(\Gamma)^*} |\tau_0 v|_{H^{1/2}(\Gamma)} + \frac{|\rho_2 - \rho_1|}{\alpha} \int_\Omega |\tilde{\chi}_{a^\beta} h v| d\Omega \\ &\leq \sigma_1 \sigma_3 \sqrt{2} |v|_{H^1(\Omega)} |\mathcal{N}(\tau_1 v)|_{H^1(\Omega)} + \tilde{\sigma}_6 |h|_{H^1(0,2)} |v|_{H^1(\Omega)}, \end{aligned}$$

where $\tilde{\sigma}_6$ is independent of β. This estimate together with (7.14) implies the desired estimate (7.12). □

THEOREM 7.9. *Assume that*

$$\rho(\beta) = \frac{\beta}{2\sigma_5^2} - \tilde{\sigma}_4 (|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*}) > 0, \tag{7.15}$$

where $\tilde{\sigma}_4$ and σ_5 are constants independent of β specified in the proof. If (i) *only $G_2(a) = A_2$ or $G_4(a) = A_4$ with $0 < \sigma_0 \le a^\beta(x) \le 1 - \sigma_0$ are considered, or* (ii) *$G_2(a) = A_2$ and $G_4(a) = A_4$ are not present but at least one of the other compatibilities are considered, then*

$$J''(x^\beta)(y,y) + \langle \Lambda^\beta, g''(x^\beta)(y,y)\rangle_Y \ge \eta \min\Big(\frac{1}{4\sigma_3}, \frac{1}{2\sigma_6}\Big)|v|^2_{H(\Gamma,\Omega)} + \rho(\beta)|h|^2_{H^1(0,2)} \quad \text{for all } y = (h,v) \in \ker g'(x^*),$$

where $\eta = \min\Big\{\beta/\sigma_5^2, 1\Big\}$, i.e. (7.5) *holds.*

Proof. We consider (i) with the compatibility condition $G_4(a) = A_4$. For $y = (h,v) \in \ker g'(x^*)$,

$$\begin{aligned} Q(y) &= J''(x^\beta)(y,y) + \langle \Lambda^\beta, g''(x^\beta)(y,y)\rangle_Y \\ &= |\mathcal{D}(\tau_0 v)|^2_{H^1(\Omega)} + |\mathcal{N}(\tau_1 v)|^2_{H^1(\Omega)} + \beta|h_x|^2_{L(0,2)} \\ &\quad + \frac{\rho_2 - \rho_1}{\alpha}\int_0^2 \int_{a^\beta - \alpha/2}^{a^\beta + \alpha/2} \lambda^\beta_y(x,s)h^2(x)dsdx + 2\mu_4^\beta \int_0^2 xh(x)^2dx. \end{aligned}$$

By (7.13), (7.10), and (7.12) we find for every $\eta \in [0,1]$,

$$\begin{aligned} Q(y) \ge{}& \frac{\eta}{4\sigma_3}|\tau_1 v|^2_{H^{1/2}(\Gamma)^*} + \frac{\eta}{2\sigma_6}|v|^2_{H^1(\Omega)} + \Big(\frac{\beta}{\sigma_5^2} - \frac{\eta}{2}\Big)|h|^2_{H^1(0,2)} \\ &- \frac{|\rho_1 - \rho_2|}{\alpha}|\lambda^\beta|_{H^1(\Omega)} \cdot |h|^2_{L^4(\Omega)} - 4|\mu_4^\beta||h|^2_{L^2(0,2)}. \end{aligned} \tag{7.16}$$

By Lemma 7.5 there exists $\tilde{\sigma}_4$ independent of β such that

$$\begin{aligned} Q(y) \ge{}& \eta \min\Big(\frac{1}{4\sigma_3}, \frac{1}{2\sigma_6}\Big)|v|^2_{H(\Gamma,\Omega)} \\ &+ \Big[\frac{\beta}{\sigma_5^2} - \frac{\eta}{2} - \tilde{\sigma}_4(|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*})\Big]|h|^2_{H^1(0,2)}. \end{aligned}$$

Choosing $\eta = \min\{\beta/\sigma_5^2, 1\}$ so that $\eta \in (0,1]$ we find

$$\begin{aligned} Q(y) &\ge \eta \min\Big(\frac{1}{4\sigma_3}, \frac{1}{2\sigma_6}\Big)|v|^2_{H(\Gamma,\Omega)} \\ &\quad + \Big[\frac{\beta}{2\sigma_5^2} - \tilde{\sigma}_4(|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^* - z_2|_{H^{1/2}(\Gamma)^*})\Big]|h|^2_{H^1(0,2)} \\ &= \eta \min\Big(\frac{1}{4\sigma_3}, \frac{1}{2\sigma_6}\Big)|v|^2_{H(\Gamma,\Omega)} + \rho(\beta)|h|^2_{H^1(0,2)}, \end{aligned}$$

which is the desired estimate. For the compatibility condition $G_2(a) = A_2$, the proof is almost indentical. In case of (ii) where any nontrivial combination of the linear compatibilities $G_i(a) = A_i, i \in \{1,3\}$ is considered, the second derivative of the Lagrangian term involving μ_i with $i \in \{1,3\}$ vanishes and the estimate becomes simpler. □

Remark 7.10. Assumption (7.15) is the weakest condition that we could find so far to obtain the conclusion of Theorem 7.9. It is unsatisfactory from the point of

view that in general we can only show that $|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*} = o(\sqrt{\beta})$ if (z_1, z_2) are attainable. Additional regularity conditions are required to improve the rate of convergence $|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*}$ to zero as $\beta \to 0^+$ [EKN], [N]. We do not investigate these aspects in this paper. Of course, (7.15) holds for values of β that are sufficiently large, since the set $\{|\tau_0 u^\beta - z_1|_{H^{1/2}(\Gamma)} + |\tau_1 u^\beta - z_2|_{H^{1/2}(\Gamma)^*} : \beta > 0\}$ is bounded.

8. Numerical examples. For the construction of a test example we proceed as follows.

Let

$$f_a(x,y) = \begin{cases} \rho_1 & \text{for } y > a(x), \\ \rho_2 & \text{for } y < a(x), (x,y) \in \mathbf{R}^2, \end{cases}$$

where $a \in H^2(\mathbf{R})$ and $a(x) \in [0,1]$ for $x \in [0,2]$. Let us denote

$$\Omega_a = \{(x,y) : -1 < x < 3, a(x) - 2 < y < a(x) + 2\},$$

and note that $\Omega_a \supset \Omega = (0,2) \times (0,1)$. It is well known that the two-dimensional fundamental solution of $-\Delta$ is given by $E(x-\tilde{x}, y-\tilde{y}) = -(1/4\pi)\log[(x-\tilde{x})^2+(y-\tilde{y})^2]$. The function $\tilde{u}$ defined by

$$\tilde{u}(x,y) = \int_{\Omega_a} E(x-\tilde{x}, y-\tilde{y}) f_a(\tilde{x},\tilde{y}) d\tilde{x} d\tilde{y}$$

satisfies

$$-\Delta \tilde{u} = f_a \quad \text{in } \tilde{\Omega}.$$

A calculation shows that

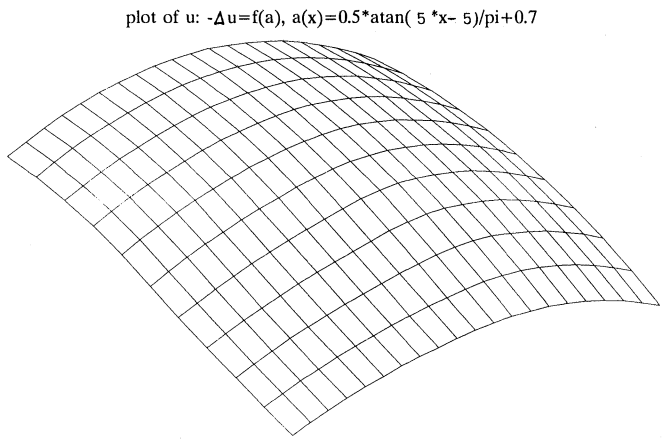
$$(8.1)\quad \begin{cases} \tilde{u}(x,y) &= -\dfrac{1}{4\pi}\displaystyle\int_{-1}^{3}\Big\{\rho_2\int_{-2}^{0}\log[(x-\tilde{x})^2+(y-\tilde{y}-a(x))^2]d\tilde{y} \\ &\qquad + \rho_1\displaystyle\int_{0}^{2}\log[(x-\tilde{x})^2+(y-\tilde{y}-a(x))^2]d\tilde{y}\Big\}d\tilde{x}, \\ \tilde{u}_x(x,y) &= -\dfrac{1}{2\pi}\displaystyle\int_{-1}^{3}\Big\{\rho_2\int_{-2}^{0}\frac{x-\tilde{x}}{(x-\tilde{x})^2+(y-\tilde{y}-a(x))^2}d\tilde{y} \\ &\qquad + \rho_1\displaystyle\int_{0}^{2}\frac{x-\tilde{x}}{(x-\tilde{x})^2+(y-\tilde{y}-a(x))^2}d\tilde{y}\Big\}d\tilde{x}, \\ \tilde{u}_y(x,y) &= -\dfrac{1}{2\pi}\displaystyle\int_{-1}^{3}\Big\{\rho_2\int_{-2}^{0}\frac{y-\tilde{y}-a(x)}{(x-\tilde{x})^2+(y-\tilde{y}-a(x))^2}d\tilde{y} \\ &\qquad + \rho_1\displaystyle\int_{0}^{2}\frac{y-\tilde{y}-a(x)}{(x-\tilde{x})^2+(y-\tilde{y}-a(x))^2}d\tilde{y}\Big\}d\tilde{x}. \end{cases}$$

Defining

$$z_1 = \tau_0 \tilde{u} \qquad \text{and} \qquad z_2 = \tau_1 \tilde{u}$$

provides exact data, which are attainable by the function a. For the test example we choose

$$a(x) = \frac{1}{2\pi}\arctan(5x-5) + 0.7.$$



N=10, meshdom: 20x10

FIG. 2.

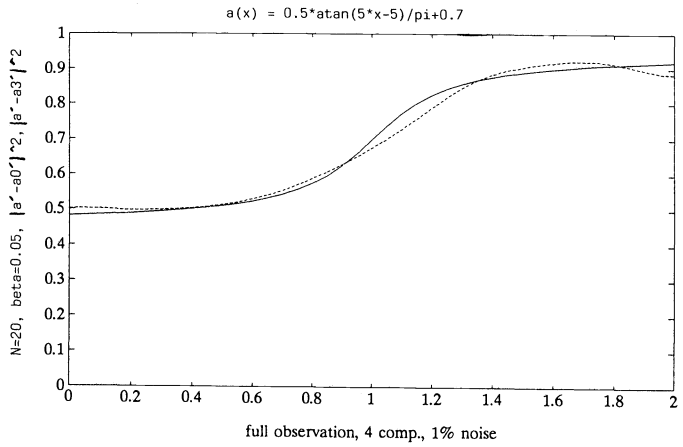


FIG. 3.

For the discretization of the variable $u \in H^1(\Omega)$ we use bilinear finite elements with respect to the grid $\{(i/N, j/N) : i = 0, \ldots, 2N; j = 0, \ldots N\}$, and the function $a \in H^1(0,2)$ was discretized by piecewise linear elements with respect to the grid $\{i/N, i = 0, \ldots, 2N\}$. A ten-point Gaussian-quadrature formula was used to calculate the exact data from (8.1). Further specifications for the numerical result of Fig. 3 to be presented below are as follows:

(i) $\rho_1 = 1, \rho_2 = 20$;

(ii) $N = 20$;

(iii) all four compatibilites are used;

(iv) observations are taken on all four sides of the rectangle;

(v) the regularization term has the form $\beta|a_x - a_x^0|^2$, with $\beta = 0.05$ and a^0 chosen such that the compatibility conditions hold;

(vi) the data are perturbed by adding 1% uniformly distributed relative noise at the nodal points of the observations (z_1, z_2).

Numerically, a^0 is calculated by solving (3.6) in the least-squares sense with start-up by a constant function with value .05. To illustrate the difficulty of the inverse

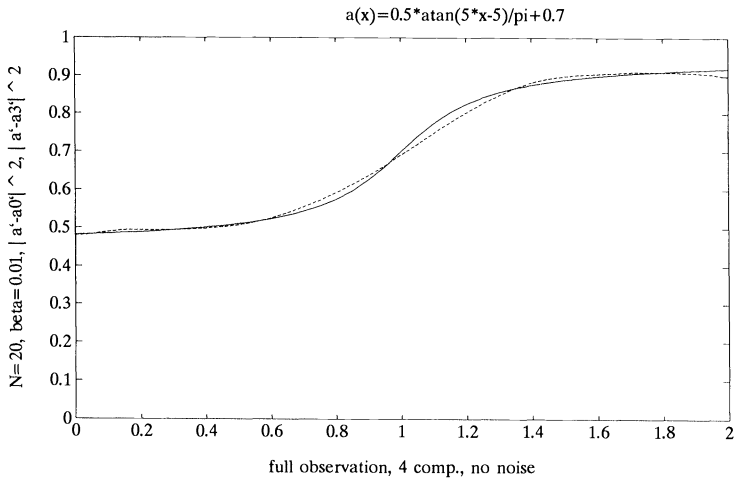


FIG. 4.

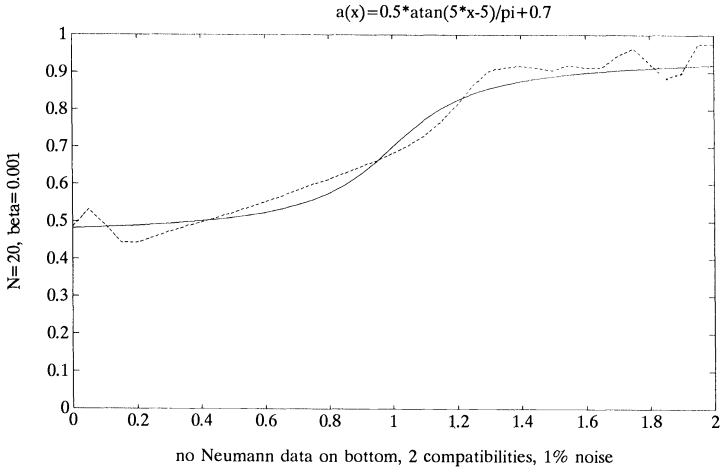


FIG. 5.

problem, a graph of $\tilde{u}$ on Ω is given in Fig. 2. The unperturbed data are obtained by evaluation of $\tilde{u}$ on the boundary Γ. Clearly, these data are extremely smooth thus making the inverse problem a difficult one. In Fig. 3 the solid line shows the exact function a and the dotted line gives the numerical result after three iterations. Observe that the scaling of the axis in the plot is not 1:1. It magnifies the error to our disadvantage.

For Fig. 4 all specifications are identical to those for Fig. 3, except for taking noise-free data and simultaneously decreasing the regularization parameter to $\beta = .01$. For the results of Fig. 5 only two compatibility conditions (the second and the fourth in (3.6)) and no Neumann data on the bottom are used. Moreover, the regularization term is changed to $\beta|a_x|^2$, which requires us to choose a smaller value for β. Here we take $\beta = .001$. As expected, the numerical reconstruction for a is now worse than in the two previous figures. Analyzing the effects of the changes between the specifications of the algorithm between Figs. 3 and 5 one at a time, it is found that

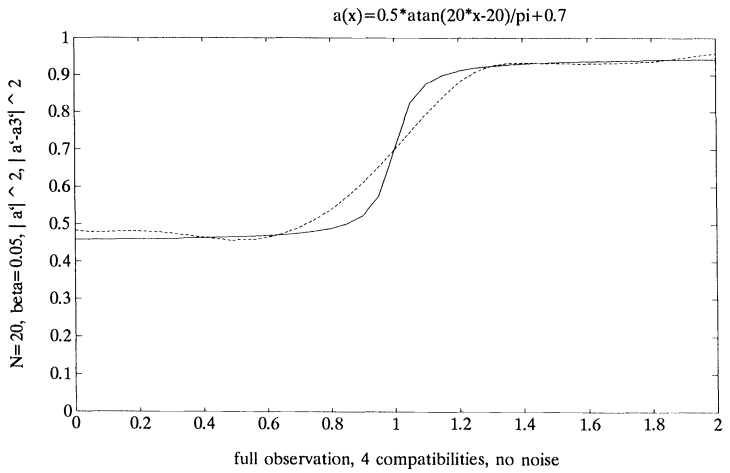


FIG. 6.

both changing the regularization term and using only two instead of four compatibility conditions contribute to the loss of accuracy. However, the former has less effect than the latter.

Returning once again to the result of Fig. 3, we recall that the compatibility conditions are used in three ways. First, they determine a very good start-up value for the first minimization with respect to a, as explained in Remark 6.1 (iii)–(iv).

Secondly, this start-up value is also used in the regularization term, and finally, the compatibility conditions are used as explicit constraints. We also tested the algorithm without the use of any compatibility condition, and using $a^\circ \equiv .5$ as a start-up value. The results (depending on various algorithmic parameters) vary from divergence to giving a qualitatively correct reconstruction of the unknown function, at best.

Figure 6 gives the numerical result in the case that the "true" inface function is given by

$$a(x) = \frac{1}{2\pi} \arctan(20x - 20) + 0.7,$$

which behaves almost like a discontinuity. The remaining specifications are those of Fig. 3 except that no noise is added.

Appendix. In this appendix we summarize facts on elliptic equations that are needed for this paper. For details we refer to [Ad], [G], [W]. Unless stated otherwise, Ω denotes an open subset of $\mathbf{R}^2$ whose boundary is a curvilinear polygon consisting of arcs $\Gamma_j, j = 1, \ldots, N$, [G], with outward normals w_j. We define $H(\Delta, \Omega) = \{u \in H^1(\Omega) : \Delta u \in L^2(\Omega)\}$.

LEMMA A.1 [G]. *Let Ω be a bounded open subset of $\mathbf{R}^2$ whose boundary is a curvilinear polygon of class $C^{1,1}$. Then for each j, the mapping*

$$u \to \left\{ u|\Gamma_j, \frac{\partial u}{\partial n}|\Gamma_j \right\},$$

which is well defined for $u \in \mathcal{D}(\bar{\Omega})$, has a unique continuous extension as an operator from $H^2(\Omega)$ into $\Pi_{i=0}^1 H^{3/2-i}(\Gamma_j)$. If, moreover, the boundary Γ of Ω is Lipschitzian, then $u \to u|\Gamma$ has an extension as a unique continuous operator from $H^2(\Omega)$ onto $H^{1/2}(\Gamma)$.

LEMMA A.2 [G]. *Let $u \in H^2(\Omega)$ and let Ω be divided into two regions Ω_1 and Ω_2 by a piecewise $C^{1,1}$ Jordan curve $\hat{\Gamma} = \bigcup_{j=1}^{p} \bar{\Gamma}_j$. Then for any smooth part of Γ the following transmissivity conditions hold:*

$$u^1|\Gamma_j = u^2|\Gamma_j \quad \text{and} \quad \frac{\partial u^1}{\partial n_j}\bigg|\Gamma_j = \frac{\partial u^2}{\partial n_j}\bigg|\Gamma_j, \quad j = 1, \dots, p,$$

where $u^1 = u|\Omega_1, u^2 = u|\Omega_2$, and n_j is normal to Γ_j and pointing into Ω.

LEMMA A.3 (Green's formula [DL], [G], [GR]). *For every $u \in H^2(\Omega)$ with Ω a polygonal domain we have*

$$\int_\Omega (\Delta u) v \, dx = -\int_\Omega \nabla u \cdot \nabla v \, dx + \sum_{j=1}^{N} \int_{\Gamma_j} \gamma_j \frac{\partial u}{\partial n_j} \gamma_j v \, dv,$$

where γ_j is the trace operator onto Γ_j. If the boundary Γ of Ω is Lipschitz continuous, then the Neumann boundary operator τ_1 has a unique continuous extension, again denoted by τ_1 from $H(\Delta, \Omega)$ to $H^{1/2}(\Gamma)^*$, and the generalized Green's formula

$$\langle \tau_1 u, v \rangle_{H^{1/2}(\Gamma)^*, H^{1/2}(\Gamma)} = \int_\Omega \Delta u \cdot v \, dx + \int_\Omega \nabla u \cdot \nabla v \, dx$$

holds for all $v \in H^1(\Omega)$.

LEMMA A.4. *The space $H^2(\Omega)$ is continuously embedded into $C^{0,1}(\bar{\Omega})$.*

LEMMA A.5. *Let Ω be a convex polygon $f \in L^2(\Omega)$ and $g \in C(\partial\Omega)$ with $g|\Gamma_j \in H^{3/2}(\Gamma_j)$. Then there exists a unique solution $u \in H^2(\Omega)$ of*

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u|\Gamma_j &= g|\Gamma_j, \qquad j = 1, \dots, N. \end{aligned} \tag{A.1}$$

Proof. By [G, p. 50] there exists an extension $\tilde{g} \in H^2(\Omega)$ of g such that $\tilde{g}|\Gamma_j = g|\Gamma_j$ for all j. Consider the equation

$$-\Delta w = f + \Delta \tilde{g} \qquad w|\partial\Omega = 0.$$

It has a unique solution $w \in H^2(\Omega)$ [G, p. 147], and $u = w + \tilde{g}$ is the desired solution of (A.1). □

LEMMA A.6. *Let $\Omega = \{(x, y) \in \mathbf{R}^2 : 0 < x < 2, 0 < y < 1\}, \Gamma = \partial\Omega$, and $\Gamma_a = \{(x, y) \in \Omega : y = a(x)\}$. Then there exist constants σ_i such that*

(i) $|\varphi|_{H^{1/2}(\Gamma)} \le \sigma_1 |\varphi|_{H^1(\Omega)}$ *for all* $\varphi \in H^1(\Omega)$;

(ii) $|\partial\varphi/\partial n|_{H^{1/2}(\Gamma)^*} \le \sigma_1 \Big(|\varphi|_{H^1(\Omega)} + |\Delta\varphi|^2_{L^2} \Big)^{1/2}$ *for all* $\varphi \in H(\Delta, \Omega)$;

(iii) *the embedding of $H^1(0, 2)$ into $C^0([0, 2])$ is compact*;

(iv) $|\varphi|^2_{H^1(\Omega)} \le \sigma_2^2 (|\nabla\varphi|^2_{L^2(\Omega)} + |\int_\Gamma \varphi \, d\Gamma|^2)$ *for all* $\varphi \in H^1(\Omega)$;

(v) $|\mathcal{D}\varphi|_{H^1(\Omega)} \le \sigma_3 |\varphi|_{H^{1/2}(\Omega)}$ *for all* $\varphi \in H^{1/2}(\Gamma)$, $|\mathcal{N}\varphi|_{H^1(\Omega)} \le \sigma_1 |\varphi|_{H^{-1/2}(\Gamma)}$ *for all* $\varphi \in H^{-1/2}(\Gamma)$.

Proof. (i)–(iii) are well-known trace and embedding properties (see [Ad], [DL, p. 380]). (iv) is a Poincaré type inequality (we refer to [DL, p. 127–133]). (v) can be verified with standard coercivity-type estimates. □

Acknowledgement. The authors thank Dr. R. Mayer who proposed the problem, and both him and Prof. H. Sünkel for discussions on the geophysical relevance of

the problem studied in this paper. Part of the paper was written while the first author visited Institut National de Recherche en Informatique et en Automatique Rocquencourt. He expresses his gratitude to the group of "Project Ident" and especially to Prof. G. Chavent.

REFERENCES

[Ad] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[Az] A. K. AZIZ, *The Mathematical Foundations of the Finite Element Methods with Applications to Partial Differential Equations*, Academic Press, New York, 1972, pp. 41–81.

[B] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[CK] F. COLONIUS AND K. KUNISCH, *Stability for parameter estimation in two point boundary value problems*, J. Reine Angew. Math., 370 (1986), pp. 1–29.

[DL] R. DAUTRAY AND J.LIONS, *Mathematical Analysis and Numerical Method for Science and Technology*, Vol. 2, Functional and Variational Methods, Springer-Verlag, Berlin, Heidelberg, 1988.

[EKN] H. W. ENGL, K. KUNISCH, AND A. NEUBAUER, *Tikhonov regularization for the solution of nonlinear ill-posed problems* I, Inverse Problems, 5 (1989), pp. 523–540.

[G] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, MA, 1985.

[GR] V. GIRAULT AND P. RAVIART, *Finite Element Methods for Navier–Stokes Equation. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[H] M. R. HESTENES, *Optimization Theory, The Finite Dimensional Case*, John Wiley, New York, 1975.

[IJ] K. ITO AND X. JIANG, *Augmented Lagrangian method for impedance computed tomography*, preprint.

[IK1] K. ITO AND K. KUNISCH, *The augmented Lagrangian method for equality and inequality constraints in Hilbert spaces*, Math. Programming, 46 (1990), pp. 341–360.

[IK2] ———, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113–136.

[IK3] ———, *On the choice of the regularization parameter in nonlinear inverse problems*, SIAM J. Optim., 2 (1992), pp. 1–29.

[IKK] K. ITO, M. KROLLER, AND K. KUNISCH, *A numerical study of the augmented Lagrangian method for the estimation of parameters in elliptic systems*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 884–910.

[KS] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed systems by regularization*, SIAM J. Control Optim., 23 (1985), pp. 217–241.

[L] K. LAMBECK, *Geophysical Geodesy, The Slow Deformation of the Earth*, Oxford Science Publication, Clarendon Press, Oxford, UK, 1988.

[M] H. MORITZ, *Advanced Physical Geodesy*, Herbert Wichman, Karlsruhe, 1989.

[Mi] C. MIRANDA, *Partial Differential Equations of Elliptic Type*, Springer-Verlag, Berlin, 1970.

[N] A. NEUBAUER, *Tikhonov regularization for nonlinear ill-posed problems: Optimal convergence rates and finite-dimensional approximation*, Inverse Problems, 5 (1989), pp. 541–558.

[SZ] J. SOKOLOWSKI AND J.-P. ZOLESIO, *Introduction to Shape Optimization, Shape Sensitivity Analysis*, Springer, Berlin, 1991.

[Tr] F. TREVES, *Basic Linear Partial Differential Equations*, Academic Press, New York, 1975.

[V] M. VOGELIUS, *A computational algorithm to determine cracks from electrostatic boundary measurements*, J. Eng. Sci., 28 (1991), pp. 917–938.

[W] J. WLOKA, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.

[ZK] J. ZOWE AND S. KURCYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1675–1695, November 1994

010

FEEDBACK STABILIZATION OVER COMMUTATIVE RINGS: THE MATRIX CASE*

V. R. SULE†

Abstract. This paper provides a solution of the feedback stabilization problem over commutative rings for *matrix transfer functions*. Stabilizability of a transfer matrix is realised as local stabilizability over the entire spectrum of the ring. For stabilizable plants, certain modules generated by its fractions and that of the stabilizing controller are shown to be projective compliments of each other. In the case of rings with irreducible spectrum, this geometric relationship shows that the plant is stabilizable if and only if the above modules of the plant are projective of ranks equal to the number of inputs and the outputs. If the maxspectrum of the ring is Noetherian and of zero (Krull) dimension, then this result shows that the stabilizable plants have doubly coprime fractions. Over unique factorization domains the above stabilizability condition is interpreted in terms of matrices formed by the fractions of the plant. Certain invariants of these matrices known as *elementary factors*, are defined and it is shown that the plant is stabilizable if and only if these elementary factors generate the whole ring. This condition thus provides a generalization of the coprime factorizability as a condition for stabilizabilty. A formula for the class of all stabilizing controllers is then developed that generalizes the previous well-known formula in factorization theory. For multidimensional transfer functions these results provide concrete conditions for stabilizabilty. Finally, it is shown that the class of polynomial rings over principal ideal domains is an additional class of rings over which stabilizable plants always have doubly coprime fractions.

Key words. feedback stabilization, coprime factorization, multidimensional systems

AMS subject classifications. 93D15, 93D25

1. Introduction.

The factorization approach to control systems over the past decade has provided important insights into the synthesis problems of linear control systems. The approach essentially emerged from the ring theoretic formulations of the feedback stabilization problem in [4] and [15] to obtain the algebraic characterization of the stabilizing controller on the lines of [16] and that of the achievable feedback system maps. However, this entire theory is founded on the coprime factorizability of transfer functions, a property always satisfied by transfer matrices over the fields of fractions of rings such as the principal ideal domain (PID) or the Bezout domain. This is the case, for example, in the stabilization of linear time-invariant transfer functions. This property fails for transfer functions having fractions over more general integral domains. In fact, in the well-known cases of n-dimensional systems and spatially distributed systems, the transfer functions belong to the fraction fields of unique factorization domains (UFDs), which do not always admit coprime fractions. Furthermore, some of the practically important problems, such as parametric scheduling of feedback controllers, involve transfer functions over several indeterminates for which coprime factorizability is not assured. For these reasons it seems worthwhile to develop a more general factorization theory of feedback systems which will encompass such examples. Thus, questions such as 1) *what are the necessary and sufficient conditions for stabilizability over a general commutative ring*? and 2) *if a plant is stabilizable, then what is the characterization of its stabilizing controllers*? need to be answered.

Recall that the coprime factorization theory (see [14]) which we refer to as the *standard factorization theory*, shows that the existence of doubly coprime fractions is

* Received by the editors October 7, 1991; accepted for publication (in revised form) April 30, 1993.

† Department of Electrical Engineering, Indian Institute of Technology, Kanpur 208 016, India.

sufficient for stabilization and characterization of the stabilizing controller. Moreover, one of the problems considered in this theory over general integral domains is to show existence of the doubly coprime fractions when any one, either left or right coprime fraction, exists. Hence for polynomial rings, an application of the celebrated Quillen–Suslin theorem shows that this is precisely true. See [14, Chap. 8] for details.

Our goal in this paper is, however, quite distinct from that of the above problems of the standard factorization theory. Primarily, our aim is to determine under what conditions a plant transfer matrix is stabilizable under the most natural fractional representation, so that these conditions also work even when coprime fractions do not exist. Next, we want to find a class of rings for which stabilizability will imply an existence of doubly coprime fractions. Thus for this class of rings, the parametrization of the controller and the achievable feedback maps will be identical to that of the standard factorization theory. Finally, we should also be able to determine the characterization of all stabilizing controllers in some concrete form in special cases of interest.

Although most rings of transfer functions needed thus far in feedback theory are integral domains, for the sake of generality, main results in this paper are developed for more general commutative rings with certain weak restrictions on the spectrum.

1.1. Previous background. The two-dimensional stabilization problem is considered in [2] for both the scalar as well as the matrix case. In [2] it is shown that the two-dimensional transfer matrices are stabilizable with the unit bidisc $\bar{U}^2$ as the domain of instability if and only if they have doubly coprime fractions. The standard coprime factorization theory of stabilization is also developed in [14, Chap. 8] for integral domains. Apart from [14, Chap. 8], the application of the Quillen–Suslin theorem is also considered in [5]. Formulation of the stabilization problem in [2] and [14] is purely in the input-output sense, where coprime factorizability of the transfer matrix plays a major role. In [8] relations between coprime factorizability and properties of split realizations are developed for rational transfer matrices over commutative rings.

For the case of scalar (i.e., single input, single output) plants, the stabilization problem over a general integral domain is solved in [13] and the conditions for stabilizability are developed in terms of coprimeness of ideals rather than in terms of fractions. Further, it is shown in [13] that the geometric interpretations for the n-dimensional case can be obtained for domains of instability that are *polynomially convex*. The purpose of this paper is to solve the corresponding matrix (i.e., multi input, multi output) case of the problem. This extension of the results of [13], which are primarily based on commutativity, to the matrix case should provide some insight into the noncommutative version of this problem.

1.2. Preview of results and organization of the paper. This section is devoted to the definition of the stabilization problem solved in the later sections as well as the relevant mathematical background and notations.

In §2 an algebraic formulation of the problem is developed. A linear system of equations is determined in Propositon 1 whose solution completely determines the solution of the stabilzation problem. Local solvability of this system of equations is then defined as local stabilizability, and it is shown in Proposition 2 that local stabilizability over the entire spectrum of the ring is equivalent to stabilizability. Analysis of these equations also reveals equivalence classes of plant transfer matrices whose stabilizability is determined by that of a given plant. Finally, an ideal theoretic interpretation of stabilizability of the plant is developed in Proposition 4, which generalizes the scalar result of [13].

Some of the main results are collected in §3. Lemma 2 shows that modules $\mathcal{T}$ and $\mathcal{W}$ determined by the fractions of the plant transfer matrix are projective complements of the similar modules of its stabilizing controller. Then it is shown that the plant has coprime fractions if and only if these modules of the plant are free. This observation, along with the local stabilizability, is then employed to show in Theorem 1 that if the spectrum of the ring is irreducible, then the plant is stabilizable if and only if the above modules are projective, of ranks equal to the number of inputs and outputs, respectively. This result is thus a geometrically necessary and sufficient condition for stabilizability over such rings. Finally, the problem of determining a class of rings for which these projective modules become free is addressd. A well-known theorem of Forster and Swan is used to show in Theorem 3 that these modules are free when the maxspectrum of the ring is Noetherian and has zero dimension. Thus, for this class of rings, stabilizability of a transfer matrix implies its coprime factorizability.

In §3.3 these results are developed further over unique factorization domains. A local-global characterization of projective modules is employed to determine a concrete interpretation of stabilizability in terms of certain invariants of the plant, called the *elementary factors*. Theorem 4 shows that the plant is stabilizable if and only if these elementary factors are coprime. A formula parametrizing all stabilizing controllers is then developed for this case.

Section 4 is devoted to the application of the above results for the problem of n-dimensional stabilization as well as for other classes of rings. Numerical examples are provided in which the stabilizability is checked by computing the elementary factors.

Next, an application of the well-known result of Quillen and Suslin is made to show that for polynomial rings over PIDs, stabilizable plants always have doubly coprime fractions. This result thus enlarges the class of rings for which the above modules are free, even when they may not have Noetherian and zero-dimensional maxspectrum. Finally, it is shown that the above theory specializes to the well-known standard factorization theory for PIDs and Bezout domains.

1.3. Stabilization problem. We begin by defining the stabilization problem for the standard feedback system considered in [14, Chap. 1, Fig. 1.2]. The reader is referred to [14] for all other details as these are well known. Let $\mathcal{A}$ be a commutative ring (with identity) denoting the ring of *stable causal* transfer functions and let $\mathcal{F}$ be its *total ring of fractions* (see notations below) denoting the class of all transfer functions. Let P, called the *plant transfer matrix*, belong to $\mathcal{F}^{n\times m}$, the set of $n \times m$ matrices over $\mathcal{F}$. Observe that P can always be represented in the form of a fraction $P = Nd^{-1}$, where N is an $n \times m$ matrix over $\mathcal{A}$ and d is a nonzero divisor. Let $\hat{F}$ denote the set $\mathcal{F}^{n\times m} \times \mathcal{F}^{m\times n}$ and consider $\hat{F}_{\text{ad}} \subset \hat{F}$ defined as

$$\hat{F}_{\text{ad}} = \{(X, Y) \in \hat{F} \mid \det(I + XY) \text{ is a unit of } \mathcal{F}\}.$$

Consider $(P, C) \in \hat{F}_{\text{ad}}$ and the map $H : \hat{F}_{\text{ad}} \longrightarrow \mathcal{F}^{(m+n)\times(m+n)}$ defined by

$$H(P,C) \triangleq \begin{pmatrix} (I+PC)^{-1} & -P(I+CP)^{-1} \\ C(I+PC)^{-1} & (I+CP)^{-1} \end{pmatrix}.$$

The stabilization problem. If the plant transfer matrix P belongs to $\mathcal{F}^{n\times m}$, when does there exist a *controller* C belonging to $\mathcal{F}^{m\times n}$ such that

1. $(P, C) \in \hat{F}_{\text{ad}}$, and
2. $H(P, C) \in \mathcal{A}^{(m+n)\times(m+n)}$?

Further, if such a C exists, what is the characterization of all such controllers? Call a plant *stabilizable* if, for its transfer function P, there exists such a C. Call C its *stabilizing controller.*

The significance of the above problem in the algebraic theory of linear feedback systems is by now quite well known and can be found in [2], [4], [14], and [15], as well as references therein. The above ring theoretic formulation of the problem of stabilizing the dynamics of a linear control system by means of a dynamic feedback was developed in [4] and [15], following its actual origin in [16]. When a controller C is a solution of the problem, the resulting feedback system is also called *externally stable* and produces bounded outputs from all the transfer functions for bounded external inputs. It is now well known that the external stability of such a feedback system also implies the stability of the dynamics, called *internal stability*, for a large class of linear systems. However, the ring theoretic stabilization problem in the above form is mathematically of a different variety and is more concerned with the structural aspects of the feedback system map $H(P, C)$, above. Its relevance to an internal state space realization (as dynamic system) has been shown in [8] for rational transfer matrices over rings. In this paper we restrict ourselves to the above form of the stabilization problem and do not attempt to relate it to state-space theoretic results.

Finally, for convenience, the problem defined above does not refer to the causality of the feedback controller C. However, following [2], [4], [15], causal transfer functions can be considerd in a ring of fractions that forms a subring of $\mathcal{F}$. In fact, since our primary goal in this paper is to study the stabilizability of a transfer matrix, causality of this matrix is not of much significance here, since it does not affect its stabilizability. We consider this condition in the following sections to show that all the solutions C of the above problem, when they exist, are causal if the plant P is strictly causal.

1.4. Notations and background results. The following notations and results are used throughout the paper. These are mentioned only briefly, as all are readily available in the standard texts indicated.

Total rings of fractions [1]. Let $\mathcal{A}$ be a commutative ring and $D \subset \mathcal{A}$ the set of all zero divisors of $\mathcal{A}$. The set $S = A - D$ is a *saturated multiplicatively closed* (MC) subset, and the *ring of fractions* $S^{-1}\mathcal{A}$ is called the *total ring of fractions* of $\mathcal{A}$. Total rings of fractions for our purpose are best, next to integral domains, for if $\mathcal{A}$ is an integral domain, then $\mathcal{F} = S^{-1}\mathcal{A}$ is its field of fractions. In the total ring of fractions $S^{-1}\mathcal{A}$, elements ab^{-1}, cd^{-1} are equal if and only if $ad - bc = 0$. The spectrum of $\mathcal{A}$ is denoted by spec $\mathcal{A}$ and its closed sets by $V(a)$ for an ideal $a \subset \mathcal{A}$. For $\mathrm{p} \in$ spec $\mathcal{A}$, the local ring at p is denoted by $\mathcal{A}_{\mathrm{p}}$. It is the ring of fractions $S^{-1}\mathcal{A}$ when the set $S = \mathcal{A} - \mathrm{p}$. Similarly, the localization of an $\mathcal{A}$-module M is denoted by M_{p}. If f in $\mathcal{A}$ is not nilpotent, then the ring of fractions with respect to the multiplicative subset $\{1, f, f^2, \ldots\}$ is denoted by $\mathcal{A}_f$. Its spectrum is homeomorphic to the open set $D(f) =$ spec $\mathcal{A} - V(f)$. Also, the maxspectrum of $\mathcal{A}$ is denoted by max $\mathcal{A}$. The *Jacobson's radical* of a commutative ring is the intersection of all its maximal ideals.

Matrix theory [12]. For a $p \times q$ matrix M with entries over a commutative ring $\mathcal{A}$, the ideal generated by the $k \times k$ minors of M is denoted by $I_k(M)$. By definition, $I_0 = \mathcal{A}$ and $I_t = 0$ for $t > \min(p, q)$. These ideals satisfy the chain $\mathcal{A} \supseteq I_1(M) \supseteq I_2(M) \supseteq \cdots \supseteq I_t(M) \supseteq 0$. A square matrix M is called *singular* if its determinant is a zero divisor of $\mathcal{A}$. For square M, the linear system of equations $Mx = 0$ over $\mathcal{A}$ has a nontrivial solution if and only if M is singular. We frequently refer to the Binet–Cauchy formula for the determinant of a product of matrices in terms of its maximal order minors. This formula is referred from [12, p. 21].

Matrix rings [11]. The ring of $n \times n$ matrices over a commutative ring $\mathcal{A}$ is denoted by $(\mathcal{A})_n$, while a_{R} and a_{L} denote its left and right ideals, respectively. Two-sided ideals are denoted without suffixes. For any right ideal $a_{\mathrm{R}} \subset (\mathcal{A})_n$, the *idealizer ring* of a_{R} is

$$\mathcal{I}(a_{\mathrm{R}}) = \{X \in (\mathcal{A})_n \mid Xa_{\mathrm{R}} \subseteq a_{\mathrm{R}}\}.$$

The center Z of $(\mathcal{A})_n$ belongs to $\mathcal{I}(a_{\mathrm{R}})$. Moreover $\mathcal{I}(a_{\mathrm{R}})$ is the largest subring of $(\mathcal{A})_n$ containing a_{R} as a two-sided ideal. Idealizer rings of left ideals are defined similarly.

Projective modules [9]. An $\mathcal{A}$-module M is called *projective* if it is a direct summand of a *free* $\mathcal{A}$-module, i.e., there is a module N such that $M \oplus N$ is free. The module N is then called its *projective complement*. For a *finitely generated* (or *finite*) projective module M, the projective complement N is also finite and $M \oplus N \simeq \mathcal{A}^n$ for some finite n. For a finite projective module M, the number of generators in a shortest system of generators (also called a *minimal generating system*) is denoted by $\mu(M)$. For a finite module M, $\mu_{\mathrm{p}}(M)$, for $\mathrm{p} \in \operatorname{spec} \mathcal{A}$, denotes the number of generators in a minimal generating system of the A_{p}-module M_{p} and is also called the *rank* of M at p. M is said to have a (constant) *rank* r if $\mu_{\mathrm{p}}(M) = r$ for all $\mathrm{p} \in \operatorname{spec} \mathcal{A}$. One of the standard results on projective modules used in the sequel is as follows (see [9, Chap. 4, §3] for details).

LEMMA 1. *If M is a finitely generated projective $\mathcal{A}$-module, then M_{p} is a free A_{p}-module, for all* p *in* spec $\mathcal{A}$.

2. Linear system of equations. In this section we develop a linear system of equations over the ring $\mathcal{A}$ whose solution determines the solution of the stabilization problem. However, observe that the stabilization problem defined above does not refer to the *causality* of the plant P or the controller C, which is an important physical constraint. In this respect we follow [2] and [15] and consider the set of causal transfer functions to be a ring of fractions contained in the ring $\mathcal{F}$ of all transfer functions. Thus consider a multiplicatively closed subset $R \subset S$ and let $R^{-1}\mathcal{A} \subset \mathcal{F}$ to be the ring of *causal* transfer functions. Observe that since $R \subset S$, it does not contain any zero divisors of $\mathcal{A}$. Next, we follow again [2], [15] and consider the set of strictly causal transfer functions as the Jacobson's radical of this ring of causal transfer functions. Hence let $\mathcal{J}$, the Jacobson's radical of $R^{-1}\mathcal{A}$, denote the set of *strictly causal* transfer functions. It is well known that this formulation of causal (strictly causal) transfer functions specializes to the familiar notion of proper (strictly proper) transfer functions of linear time-invariant systems. We extend these ideas for matrix transfer functions as follows.

DEFINITION 1. *A matrix M in $\mathcal{F}^{n\times m}$ is called causal if M has all entries in $R^{-1}\mathcal{A}$. A causal matrix M is called strictly causal if $I_t(M) \subseteq \mathcal{J}$, where $t = \min(n, m)$.*

Thus if $P = Nd^{-1}$ is strictly causal, then, since $d \in R$, the ideal $I_t(N)$ is contained in all maximal ideals of $\mathcal{A}$ which do not intersect R. Hence N treated as a matrix over $R^{-1}\mathcal{A}$ has $I_t(N) \subseteq \mathcal{J}$. Clearly, for 1×1 matrices we obtain a specialization of the above to familiar notions of causal and strictly causal transfer functions.

2.1. Linear system of equations and stabilizability. Consider an arbitrary fractional representation $P = Nd^{-1}$ for P. Any other fraction $P = Ab^{-1}$ satisfies $Ad - Nb = 0$. Fixing one fraction Nd^{-1} we have the following proposition.

PROPOSITION 1. *A strictly causal plant $P = Nd^{-1}$ is stabilizable if and only if there exists a solution $X \in (\mathcal{A})_n$, $U \in \mathcal{A}^{n\times m}$, $Y \in \mathcal{A}^{m\times n}$, and $W \in (\mathcal{A})_m$ for the*

equations

$$\begin{aligned} XN &= Ud, \\ YN &= Wd, \\ NY &= (I - X)d. \end{aligned} \tag{1}$$

Moreover, 1) *if* P *is stabilizable, then any stabilizing controller has the form* $C = YX^{-1}$, *where* Y, X *satisfy* (1), *and* 2) *conversely, if* (1) *has a solution* X, Y, *then* $\det X \in R$ *and* $C = YX^{-1}$ *is a stabilizing controller.*

Proof. First, the proof is by necessity and 1). Let an $m \times n$ matrix C over $\mathcal{F}$ be a stabilizing controller of P. Then $H(P, C)$ has all entries in $\mathcal{A}$. Call $(I + PC)^{-1} = X$, $C(I + PC)^{-1} = Y$, $P(I + CP)^{-1} = U$, and $(I + CP)^{-1} = I - W$. Clearly, since P is strictly causal, it follows that $\det (I + PC)^{-1}$ is a unit of $R^{-1}\mathcal{A}$. Hence $\det X \in R \subset S$. Also $C = Y(I + PC) = YX^{-1}$.

Now, $U = (I + PC)^{-1}P = XNd^{-1} \Leftrightarrow XN = Ud$.

Next, $I - W = I - C(I + PC)^{-1}P = I - YNd^{-1} \Leftrightarrow YN = Wd$.

Finally, $X = I - PC(I + PC)^{-1} \Leftrightarrow NY = (I - X)d$.

Second, the proof is by sufficiency and 2). Let (1) have a solution X. Consider the equation $NY = (I - X)d$ as an equation over $R^{-1}\mathcal{A}$ taking the denominator fractions as the identity. Since $I_t(N) \subseteq \mathcal{J}$, using the Binet–Cauchy formula it follows that $I_n(NY) = (\det (NY)) \subseteq \mathcal{J}$. Hence $\det (I - X)d$ is also in $\mathcal{J}$, which implies that $\det X \in R$ as d is a unit of $R^{-1}\mathcal{A}$. Hence $C = YX^{-1}$ is causal. For any such C, we have $(I + PC)^{-1} = Xd(dX + NY)^{-1} = dX(dX + (I - X)d)^{-1} = X$, i.e., $(I + PC)^{-1}$ belongs to $(\mathcal{A})_n$. Similarly it can be easily verified that all other entries of $H(P, C)$ belong to $\mathcal{A}$, which shows that C above stabilizes P. □

Thus the stabilizability of a strictly causal P by the above is equivalent to the solvability of (1). Moreover, every stabilizing controller is also causal. The result holds only for strictly causal P in this general case. Later, for integral domains, we relax this condition (see §4). At this stage it may be worthwhile to observe that if any other fractions $P = Ab^{-1}$ are used instead of Nd^{-1}, then $XN = Yd \Leftrightarrow XNb = Ydb \Leftrightarrow XA = Yb$ as both d and b are nonzero divisors if X, Y satisfy $XN = Yd$, since $Nb = Ad$ (identical arguments for $NX = Yd$). Thus solutions of (1) are invariant with respect to the choice of fractions of P. The above proposition is essentially just a reformulation of the well-known Q-parametrization in feedback stabilization.

2.1.1. Local stabilizability. The definition of stabilizability can be generalised as follows. Since $H(P, C)$ in the stable system has all entries in $\mathcal{A}$, we call the above notion $\mathcal{A}$-stabilizability. Hence for certain ring homomorphisms $f : \mathcal{A} \longrightarrow \mathcal{B}$, we can obtain $\mathcal{B}$-stabilizability. Using this idea, the following notion of stabilizability is obtained. Let $\mathcal{F}(\mathcal{B})$ denote the total ring of fractions of a ring $\mathcal{B}$. For the local ring $\mathcal{A}_{\mathrm{m}}$ at any maximal ideal m, it follows that a matrix P over $\mathcal{F}$ also belongs to $\mathcal{F}(\mathcal{A}_{\mathrm{m}})$.

DEFINITION 2 (Local stabilizability). *A matrix* P *belonging to* $\mathcal{F}^{n\times m}$ *is locally stabilizable at a maximal ideal* m *of* $\mathcal{A}$ *if there is a* C *in* $\mathcal{F}^{m\times n}$ *such that*

1. $\det H(P, C)$ *is a unit of* $\mathcal{F}(\mathcal{A}_{\mathrm{m}})$, *and*
2. *the matrix* $H(P, C)$ *has all entries in* $\mathcal{A}_{\mathrm{m}}$.

PROPOSITION 2. *A strictly causal* P *is stabilizable if and only if* P *is locally stabilizable at all maximal ideals of* $\mathcal{A}$.

Proof. First, the proof is by necessity. Let C be a stabilizing controller. Then $\det (I + PC)^{-1}$ is a nonzero divisor of $\mathcal{A}$ hence, treated as an element of $\mathcal{A}_{\mathrm{m}}$ also a nonzero divisor of $\mathcal{A}_{\mathrm{m}}$ and consequently a unit of $\mathcal{F}(\mathcal{A}_{\mathrm{m}})$ for any maximal ideal m.

Since all entries of $H(P,C)$ belong to $\mathcal{A}$ they also belong to $\mathcal{A}_{\mathrm{m}}$. Thus P is locally stabilizable at any m.

Second, the proof is by sufficiency. Imitating the steps of the necessity part of the proof of Proposition 1 with $\mathcal{A}$ replaced by $\mathcal{A}_{\mathrm{m}}$ and $\mathcal{F}$ replaced by $\mathcal{F}(\mathcal{A}_{\mathrm{m}})$, it follows that the linear system of equations is described over $\mathcal{A}_{\mathrm{m}}$ and has solutions of X, Y, U, W, with all entries in $\mathcal{A}_{\mathrm{m}}$.

Now we use a well-known result [12, p. 92] on linear systems of equations, which shows that a linear system of equations $Ax = b$ over a commutative ring $\mathcal{A}$ has a solution if and only if it has a solution over $\mathcal{A}_{\mathrm{m}}$ for every maximal ideal m.

Clearly, as (1) can be written in the form $Ax = b$ by choosing A and b from the entries of the fractions N and d, it follows from this lemma that (1) in fact has a solution over $\mathcal{A}$. Hence using the sufficiency part of the proof of Proposition 1, it follows that P is stabilizable. □

The notion of local stabilizability has also been useful in [6] and [7], though in the context of state feedback stabilization of systems over rings. In [7] the equivalence of local and global stabilizability is proved using analytic approaches. In comparison to these the approach developed above is purely algebraic.

2.1.2. Stabilizability and equivalence classes. We now show that stabilizability is in fact a property of much bigger equivalence classes of $\mathcal{A}^{n\times m} \times S$ than just those obtained by different choices of fractions of P. Two $n \times m$ matrices N, M over $\mathcal{A}$ are called *equivalent* (denoted $N \sim M$) if there exist *unimodular* (or invertible) matrices $A \in (\mathcal{A})_n$ and $B \in (\mathcal{A})_m$ such that $N = AMB$. Observe that two matrices P, Q in $\mathcal{F}^{n\times m}$ can always be written in terms of fractions having a common denominator. For if $P = Nd^{-1}$ and $Q = Ab^{-1}$, then we have $P = Nb(db)^{-1}$ and $Q = Ad(db)^{-1}$. Hence, call P, Q S-*equivalent* if $Nb \sim Ad$. It is easy to verify that this is an equivalent relation on $\mathcal{F}^{n\times m}$ and also on $\mathcal{A}^{n\times m} \times S$, where it defines bigger equivalence classes than those equivalent to elements in $\mathcal{F}^{n\times m}$.

PROPOSITION 3. *Let P be strictly causal. Then P is stabilizable if and only if Q is stabilizable for all Q that are S-equivalent to P.*

Proof. First, the proof is by sufficiency. P is S-equivalent to itself.

Second, the proof is by necessity. Let $P = Nd^{-1}$ and $Q = Mt^{-1}$ be S-equivalent to P. Thus there exist unimodular matrices A, B such that $ANtB = Md$. Since P is stabilizable, (1) has a solution X, Y, U, W for fractions Nt and dt. Define $\bar{X} = AXA^{-1}$, $\bar{Y} = B^{-1}YA^{-1}$, $\bar{U} = AUB$, and $\bar{W} = B^{-1}WB$. Then (1) in terms of these new matrices is as follows:

$$\begin{aligned}\bar{X}(ANBt) &= \bar{U}(dt),\\ \bar{Y}(ANBt) &= \bar{W}(dt),\\ (ANBt)\bar{Y} &= (I - \bar{X})(dt).\end{aligned}$$

Thus (1) is satisfied for Q as well. Since Q is also strictly causal, being S-equivalent to P, Q is stabilizable. □

The above proof shows that the stabilizing controllers of S-equivalent plants have fractions that are appropriately left and right associates.

2.1.3. Block diagonal plant. As another application of the linear equation formulation, consider a block decoupled transfer matrix

$$P = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix},$$

where P_1 and P_2 are $n_1 \times m_1$ and $n_2 \times m_2$ matrices over $\mathcal{F}$, respectively. Clearly, P is strictly causal if and only if both P_1 and P_2 are strictly causal. For strictly causal P of this kind we have the following result. The proof follows from Proposition 1.

COROLLARY 1. *The block diagonal P is stabilizable if and only if both P_1 and P_2 are stabilizable. If C_1 and C_2 are stabilizing controllers of P_1 and P_2, then*

$$C = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}$$

is also a stabilizing controller of P.

2.2. Coprime ideals and stabilizability. In [13] stabilizability of a scalar plant $p = nd^{-1}$ is obtained as coprimeness of ideals $\hat{a} = (n : d)$ and $\hat{b} = (d : n)$. We now determine the corresponding ideals for stabilizability of a matrix P over $\mathcal{F}$. We treat the case of only the square plant matrix with $n = m$. Nonsquare cases can be developed on similar lines but require more complex notations. For $P = Nd^{-1}$ in $(\mathcal{F})_n$, consider the set

$$b_{\mathrm{L}} = \{X \in (\mathcal{A})_n \mid XN \in (d)_n\},$$

where $(d)_n \subset (\mathcal{A})_n$ is the principal ideal generated by dI. Clearly, b_{L} is a left ideal of $(\mathcal{A})_n$. Consider the set Nb_{L}, which consists of all matrices of the form NX where X belongs to b_{L}. Let $\mathcal{I}(b_{\mathrm{L}})$ be the idealizer ring of b_{L}. Recall that

$$\mathcal{I}(b_{\mathrm{L}}) = \{X \in (\mathcal{A})_n \mid b_{\mathrm{L}}\, X \subseteq b_{\mathrm{L}}\}.$$

Observe that $\mathcal{I}(b_{\mathrm{L}})$ contains the center Z of $(\mathcal{A})_n$ as well as the ring $C(N)$ of matrices that commute with N. Since b_{L} is a two-sided ideal of $\mathcal{I}(b_{\mathrm{L}})$, the set Nb_{L} is a right ideal of $\mathcal{I}(b_{\mathrm{L}})$. Consider

$$a_{\mathrm{R}} = \{Y \in \mathcal{I}(b_{\mathrm{L}}) \mid Yd \in Nb_{\mathrm{L}}\}.$$

Clearly, a_{R} is a right ideal of $\mathcal{I}(b_{\mathrm{L}})$. Thus the sum of ideals $b_{\mathrm{L}} + a_{\mathrm{R}}$ is also a right ideal of $\mathcal{I}(b_{\mathrm{L}})$. Finally, consider the following ideals of the center Z of $(\mathcal{A})_n$:

$$b = b_{\mathrm{L}} \cap Z, \qquad a = a_{\mathrm{R}} \cap Z.$$

Thus a and b are ideals of the commutative ring $\mathcal{A}$.

PROPOSITION 4. *A strictly causal $P = Nd^{-1}$ in $(\mathcal{F})_n$ is stabilizable if and only if $a + b = \mathcal{A}$.*

Proof. First, the proof is by necessity. Since (1) has a solution, there is X in b_{L} and $X' = I - X$ in a_{R}. As b_L is a left ideal of $(\mathcal{A})_n$, $\det X = (\mathrm{adj}\ X)X$ belongs to $b_{\mathrm{L}} \cap Z = b$. Further, $X + X' = I$ implies that $(\mathrm{adj}\ X')X + \det X' = \mathrm{adj}\ X'$. Hence as $(\mathrm{adj}\ X')X \in b_{\mathrm{L}} \subset \mathcal{I}(b_{\mathrm{L}})$ and $\det X' \in Z \subset \mathcal{I}(b_{\mathrm{L}})$, it follows that adj X' belongs to $\mathcal{I}(b_{\mathrm{L}})$.

Now, as X' belongs to the right ideal a_{R} of $\mathcal{I}(b_{\mathrm{L}})$, $\det X' = X'\mathrm{adj}\ X'$ also belongs to a_{R}, but then $\det X' \in a$. As $\det X + \det X' = 1$, $a + b = \mathcal{A}$.

Second, the proof is by sufficiency. Let $x \in a$ and $y \in b$ such that $x + y = 1$. Then $xI \in b_{\mathrm{L}}$, $yI \in a_{\mathrm{R}}$, and $I = xI + yI \in b_{\mathrm{L}} + a_{\mathrm{R}}$. Hence it follows from the definitions of ideals b_{L} and a_{R} that there is a solution to (1). Hence P is stabilizable. □

The above algebraic condition of coprimeness of ideals a and b is now immediately amenable to geometrical interpretation in terms of the closed sets of the spectrum of

the ring $\mathcal{A}$. Since two ideals are coprime if and only if there is no prime ideal containing both of them, it follows that P is stabilizable if and only if

$$V(a) \cap V(b) = \emptyset,$$

where $V(a)$ and $V(b)$ are closed sets in spec $\mathcal{A}$. Following [13] we call $V(a)$ the *generalized zero set* of P and $V(b)$ the *generalized pole set* of P.

For the scalar plant $p = nd^{-1}$, the above ideals are $b = (d : n)$ and $a = (nb : d)$. However, the ideals obtained in [13] are $\hat{b} = (d : n)$ and $\hat{a} = (n : d)$. Thus $\hat{b} = b$. We now show that $\hat{a}$ is also equal to a. For $a = (nb : d) \subseteq (n \cap b : d) = (n : d)$ as $(d) \subseteq b$. Thus $a \subseteq \hat{a}$. Conversely let $x \in \hat{a}$. Then for some y in $\mathcal{A}$,

$$\begin{array}{rrcl} & xd & = & yn, \\ \Rightarrow & y & \in & (d : n) = b, \\ \Rightarrow & yn & \in & (n)b, \\ \Rightarrow & x(d) & \subseteq & (n)b, \\ \Rightarrow & x & \in & (nb : d) = a. \end{array}$$

Thus the result above specializes to the scalar case of [13]. In fact the above result is valid for any commutative ring, not just for integral domains.

Although the results of this section resolve the stabilization problem algebraically, they do not reveal the geometric relationship between the plant and its stabilizing controller. This relationship is an important aspect of the feedback stabilization problem to which we now turn.

3. Geometric conditions for stabilizability. In this section we first investigate the geometric relationship between the plant and its stabilizing controller whenever one exists, and then develop the geometric necessary and sufficient conditon for stabilizability of the plant. In the standard factorization theory, this relationship is exhibited by the doubly coprime factorization. For convenience we reproduce the definition of this factorization.

DEFINITION 3 (Doubly coprime factorization). *Let P be an $n \times m$ matrix over $\mathcal{F}$. Then P is said to have a doubly coprime factorization* (DCF) *if there exist the following matrices*:

1. N, $\tilde{N}$ *in* $\mathcal{A}^{n\times m}$, $\tilde{D}$ *in* $(\mathcal{A})_n$, D *in* $(\mathcal{A})_m$;
2. X, $\tilde{X}$ *in* $\mathcal{A}^{m\times n}$, Y *in* $(\mathcal{A})_n$, $\tilde{Y}$ *in* $(\mathcal{A})_m$;
3. D, $\tilde{D}$, Y, $\tilde{Y}$ *are all nonsingular and satisfy*;
4. $P = ND^{-1} = \tilde{D}^{-1}\tilde{N}$ *and the identity*

$$\begin{bmatrix} \tilde{X} & \tilde{Y} \\ \tilde{N} & -\tilde{D} \end{bmatrix} \begin{bmatrix} D & Y \\ N & -X \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

In the standard factorization theory (i.e., over PIDs or Bezout domains) such a factorization exists for every plant. Moreover each such factorization gives rise to a stabilizing controller. Now consider some additional notations. For $P = Nd^{-1}$ and $C = N_c d_c^{-1}$, denote various matrices as follows:

$$T = \begin{bmatrix} N \\ dI \end{bmatrix}, \qquad T_c = \begin{bmatrix} N_c \\ d_c I \end{bmatrix},$$

$$W = \begin{bmatrix} N & dI \end{bmatrix}, \qquad W_c = \begin{bmatrix} N_c & d_c I \end{bmatrix},$$

$$Q_t = \begin{bmatrix} d_c I & N \\ -N_c & dI \end{bmatrix}, \qquad Q_w = \begin{bmatrix} dI & N \\ -N_c & d_c I \end{bmatrix},$$

$$\triangle_t = \begin{bmatrix} d_c I & 0 \\ 0 & dI \end{bmatrix}, \qquad \triangle_w = \begin{bmatrix} dI & 0 \\ 0 & d_c I \end{bmatrix}.$$

In the above matrices the dimension of the identity matrix I is made clear from the context. For example, in T the identity matrix is $m \times m$, while in W it is $n \times n$. Further, let $\mathcal{T}$ denote the $\mathcal{A}$-module generated by the rows of the matrix T, and let $\mathcal{T}_c$ denote the $\mathcal{A}$-module generated by the rows of the matrix T_c. Similarly let $\mathcal{W}$ denote the $\mathcal{A}$-module generated by the columns of the matrix W, while $\mathcal{W}_c$ denotes the module generated by the columns of the matrix W_c. In general, for a matrix X let $M_r(X)$ denote the module generated by the rows of X. Finally, let H denote the matrix $H(P, C)$ of the feedback system. The following lemma gives the relationship between P and C.

LEMMA 2. *Let $P = Nd^{-1}$ be stabilizable and $C = N_c d_c^{-1}$ be its stabilizing controller. Then*

$$\mathcal{T} \oplus \mathcal{T}_c \simeq \mathcal{A}^{n+m},$$
$$\mathcal{W} \oplus \mathcal{W}_c \simeq \mathcal{A}^{n+m}.$$

Proof. Recall that in the feedback system,

$$H = \begin{bmatrix} I & P \\ -C & I \end{bmatrix}^{-1}.$$

Hence $\det H = \det (I + PC)^{-1}$, which implies that both $\det H$ and $\det(\text{adj } H)$ are not zero divisors. As both d and d_c are also not zero divisors, from the above we obtain the following identities:

$$(\text{adj } H)\triangle_t = Q_t \det H, \tag{2}$$

$$\triangle_w(\text{adj } H) = Q_w \det H. \tag{3}$$

For (2) we obtain, upon padding the matrices on both sides by $\det H \triangle_t$,

$$\begin{bmatrix} (\text{adj } H)\triangle_t \\ \triangle_t \det H \end{bmatrix} = \begin{bmatrix} Q_t \\ \triangle_t \end{bmatrix} \det H,$$

which is equivalent to

$$\begin{bmatrix} I \\ H \end{bmatrix} (\text{adj } H)\triangle_t = \begin{bmatrix} Q_t \\ \triangle_t \end{bmatrix} \det H.$$

Now observe that the module generated by the rows of $[I\ H^T]^T$ is free of rank $m + n$. Hence as $\det((\text{adj } H)\triangle_t)$ is not a zero divisor, we obtain

$$M_r\left(\begin{bmatrix} I \\ H \end{bmatrix} (\text{adj } H)\triangle_t\right) \simeq M_r\left(\begin{bmatrix} I \\ H \end{bmatrix}\right) \simeq \mathcal{A}^{m+n}.$$

Using this fact and the fact that $\det H$ is not a zero divisor, it follows from the above equation that

$$M_r\left(\begin{bmatrix} Q_t \\ \triangle_t \end{bmatrix}\right) \simeq \mathcal{A}^{m+n}.$$

Now, a simple observation shows the equivalence of following matrices:

$$\begin{bmatrix} Q_t \\ \triangle_t \end{bmatrix} \sim \begin{bmatrix} T & 0 \\ 0 & T_c \end{bmatrix}.$$

Thus clearly,

$$M_r\left(\begin{bmatrix} T & 0 \\ 0 & T_c \end{bmatrix}\right) = \mathcal{T} \oplus \mathcal{T}_c \simeq \mathcal{A}^{m+n}.$$

An exactly similar computation starting from (3) shows that

$$\mathcal{W} \oplus \mathcal{W}_c \simeq \mathcal{A}^{m+n}$$

is also satisfied. □

Thus in a stable feedback system the modules $\mathcal{T}$ and $\mathcal{T}_c$ generated by the fractions of the plant and the controller, respectively, are projective complements of each other in $\mathcal{A}^{m+n}$. In the standard factorization theory, an algebraic characterization of the stabilizing controller is obtained once a DCF of P is computed. In contrast to this, the above result reveals a geometrical characteristic of all stabilizing controllers of P. However, this does not give a geometrical characterization of the controller since this is only a necessary condition that a controller must satisfy in order to be a stabilizing controller.

3.1. Doubly coprime factorization. We now give a necessary and sufficient condition under which a DCF exists for the plant P.

LEMMA 3. *$P = Nd^{-1}$ has a doubly coprime factorization if and only if both $\mathcal{T}$ and $\mathcal{W}$ are free $\mathcal{A}$-modules of ranks m and n, respectively.*

Proof. First, the proof is by sufficiency. Let $v_1, v_2, \ldots, v_m$, $v_i \in \mathcal{A}^m$ be a basis of $\mathcal{T}$ and $u_1, u_2, \ldots, u_n$, $u_j \in \mathcal{A}^n$ be a basis of $\mathcal{W}$. Cosider the matrix V in $(\mathcal{A})_m$ whose rows are v_i, and the matrix U in $(\mathcal{A})_n$ whose columns are u_j. Since $\{v_i\}$ and $\{u_j\}$ are both linearly independent sets, $\det V$ and $\det U$ are both nonzero divisors in $\mathcal{A}$. Thus we can write the matrices T and W in the form

$$T = \begin{bmatrix} A \\ B \end{bmatrix} V, \qquad W = U \begin{bmatrix} \tilde{A} & \tilde{B} \end{bmatrix},$$

by uniquely choosing the matrices A, $\tilde{A}$ in $\mathcal{A}^{n\times m}$, B in $(\mathcal{A})_m$, and $\tilde{B}$ in $(\mathcal{A})_n$. Clearly, as $d^m = \det(BV)$ and $d^n = \det(UB)$, both $\det B$ and $\det \tilde{B}$ are also nonzero divisors. Hence we also have $Nd^{-1} = AB^{-1} = \tilde{B}^{-1}\tilde{A}$. Now, since $v_i, i = 1, \ldots, m$ belong to $\mathcal{T}$ and $u_j, j = 1, \ldots, n$ belong to $\mathcal{W}$, we can obtain matrices $\tilde{Y}$, Y' in $\mathcal{A}^{m\times n}$, $\tilde{X}$ in $(\mathcal{A})_m$, and X' in $(\mathcal{A})_n$ such that

$$V = \begin{bmatrix} \tilde{Y}\tilde{X} \end{bmatrix} T = (\tilde{Y}A + \tilde{X}B)V,$$

$$U = W \begin{bmatrix} Y \\ X \end{bmatrix} = (\tilde{A}Y' + \tilde{B}X')U.$$

Hence, as $\det V$, $\det U$ are nonzero divisors, we obtain

$$\tilde{Y}A + \tilde{X}B = I, \qquad \tilde{A}Y' + \tilde{B}X' = I.$$

Now, by a well-known procedure [14], define $\tilde{Y}X' - \tilde{X}Y' = M$ and let $Y = BM + Y'$ and $X = X' - AM$. The doubly coprime factorization is then obtained as

$$\begin{bmatrix} \tilde{X} & \tilde{Y} \\ \tilde{A} & -\tilde{B} \end{bmatrix} \begin{bmatrix} B & Y \\ A & -X \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Let this equation be denoted as $LR = I$, where L and R denote the unimodular matrices of the above DCF.

Second, the proof is by necessity. If P has a DCF $LR = I$ as above, then a small computation shows that there exist unimodular matrices E and F such that

$$ET = \begin{bmatrix} V \\ 0 \end{bmatrix} \quad \text{and} \quad WF = \begin{bmatrix} U & 0 \end{bmatrix},$$

where U and V are nonsingular. Hence, as E and F are unimodular, $\mathcal{T}$ is isomorphic to $M_r(V)$ which is free of rank m. Similarly, it follows that $\mathcal{W}$ is free of rank n. □

Remark. The matrix interpretation of the lemma, which is clear from the the proof above, is as follows. The modules $\mathcal{T}$ and $\mathcal{W}$ are free of ranks m and n, respectively, if and only if there exist unimodular matrices L and R such that

$$LT = \begin{bmatrix} V \\ 0 \end{bmatrix} \quad \text{and} \quad WR = \begin{bmatrix} U & 0 \end{bmatrix},$$

where V is $m \times m$, U is $n \times n$, and both are nonsingular.

It is well known that once the DCF is given as above then the plant P is stabilizable by the controller $C = \tilde{X}^{-1}\tilde{Y} = YX^{-1}$. We use this next to prove one of the central results of this paper.

3.2. Geometric necessary and sufficient condition. We now develop a necessary and sufficient condition for stabilizability of P purely in terms of the modules $\mathcal{T}$ and $\mathcal{W}$. Hence it is first necessary to establish the uniqueness of $\mathcal{T}$ and $\mathcal{W}$ with respect to fractions Nd^{-1} of P. Let $P = Ab^{-1}$ be another choice of fractions with $T' = [A^T\ bI]^T$. Then as $Nb = Ad$ and d and b are both nonzero divisors, we have $\mathcal{T} \simeq M_r(TbI) \simeq M_r(T'dI) \simeq M_r(T')$. Thus the module $\mathcal{T} = M_r(T)$ is obtained uniquely up to an isomorphism for any choice of fractions of P. Similarly, it follows that the module $\mathcal{W}$ is also obtained uniquely. In fact, it is easy to observe that if P and Q are S-equivalent, then the modules $\mathcal{T}$ of P and Q (similarly, $\mathcal{W}$) are isomorphic. Thus a stabilizability condition in terms of the modules $\mathcal{T}$ and $\mathcal{W}$ will be invariant, not only with respect to the choice of fractions, but also with respect to the S-equivalent class of fractions. We thus fix an arbitrary fraction $P = Nd^{-1}$.

Consider the class of rings $\mathcal{A}$ whose spectrum is an irreducible topological space. In such a space every nonempty open subset is dense. Hence any two nonempty open subsets have a nonempty intersection. A ring $\mathcal{A}$ has spec $\mathcal{A}$ irreducible if and only if its nilradical (the set of nilpotent elements) is a prime ideal. For this class of rings, which is sufficiently broad to cover applications of our interest, we obtain the necessary and sufficient conditon for stabilizability in the theorem below. We first need the following well-known lemma, which shows the semicontinuity of ranks of projective modules.

LEMMA 4. *If M is a finitely generated projective $\mathcal{A}$-module and r is a natural number, then the set of all prime ideals* p *in* spec $\mathcal{A}$, *for which* $\mu_{\mathrm{p}}(M) < r$, *is open.*

THEOREM 1. *Let* spec $\mathcal{A}$ *be irreducible. Then a strictly causal plant P is stabilizable if and only if $\mathcal{T}$ and $\mathcal{W}$ are both projective of ranks m and n, respectively.*

Proof. First, the proof is by necessity. From Lemma 2 there is a $C = N_c d_c^{-1}$ with $\mathcal{T} \oplus \mathcal{T}_c \simeq \mathcal{A}^{n+m}$. Hence, as projective modules are locally free (see Lemma 1) the localizations $\mathcal{T}_{\mathrm{p}}$ and $\mathcal{T}_{c\mathrm{p}}$ are free $\mathcal{A}_{\mathrm{p}}$-modules for all p $\in$ spec $\mathcal{A}$, and moreover, we have

$$\mathcal{T}_{\mathrm{p}} \oplus \mathcal{T}_{c\mathrm{p}} \simeq \mathcal{A}_{\mathrm{p}}^{n+m},$$

which implies that

$$\mu_{\mathrm{p}}(\mathcal{T}) + \mu_{\mathrm{p}}(\mathcal{T}_c) = n + m \quad \forall \mathrm{p} \in \text{spec } \mathcal{A},$$

where $\mu_{\mathrm{p}}(\mathcal{T})$ is the rank of $\mathcal{T}$ at p. Now, if d is not contained in a prime ideal p, then clearly $\mathcal{T}$ contains m linearly independent rows. Hence $\mu_{\mathrm{p}}(\mathcal{T}) = m$ for all p in the open set $D(d) =$ spec $\mathcal{A} - V(d)$. So let p be a prime ideal containing d. Thus p belongs to the closed set $V(d)$ in spec $\mathcal{A}$. Now if $\mu_{\mathrm{p}}(\mathcal{T}) < m$ at this p, then due to the semicontinuity property of the rank as given in the above lemma, there is an open set D' in spec $\mathcal{A}$ such that $\mu_{\mathrm{p}}(\mathcal{T}) < m$ for all p in D'. However, since spec $\mathcal{A}$ is irreducible, $D(d)$ and D' have a nonempty intersection which gives a contradiction. Hence rank of $\mathcal{T}_{\mathrm{p}}$ is equal to m for all p. On similar lines it can be proved that the rank of $\mathcal{W}_{\mathrm{p}}$ is equal to n over the entire spectrum.

Second, the proof is by sufficiency. By hypothesis, $\mathcal{T}_{\mathrm{m}}$ and $\mathcal{W}_{\mathrm{m}}$ are free of rank m and n, respectively, for all maximal ideals m. Hence from Lemma 3 it follows that P has a doubly coprime factorization over $\mathcal{A}_{\mathrm{m}}$ (i.e., all matrices in the DCF are obtained over $\mathcal{A}_{\mathrm{m}}$). Thus there is a $C = YX^{-1}$, with X, Y having all entries in $\mathcal{A}_{\mathrm{m}}$ and for which $\det(I + PC)$ is a unit of $\mathcal{F}(\mathcal{A}_{\mathrm{m}})$ as well as $H(P, C)$, that has all entries in $\mathcal{A}_{\mathrm{m}}$. Thus P is locally stabilizable for all maximal ideals m. Hence, by Proposition 2, P is stabilizable. □

In the proof of the above theorem we have used the results of Proposition 1 and the subsequent local-to-global implication of stabilizability of Proposition 2. This is because the only other sufficient condition for stabilizability available so far is via existence of DCF for which the modules $\mathcal{T}$ and $\mathcal{W}$ are required to be free, which is in general not the case. For this reason it will be useful to know when these modules are free because then, by Lemma 3, there will actually be a DCF available. We thus investigate the class of rings for which stabilizability will imply that the modules $\mathcal{T}$ and $\mathcal{W}$ are free of ranks m and n, respectively.

3.2.1. Number of generators and coprimeness. For convenience we reproduce below some standard definitions and results, see [9] for details. A topological space X is called *Noetherian* if every descending chain $V_1 \supset V_2 \supset \ldots$ of closed sets $V_i \subset X$ is stationary. The space X is called *irreducible* if, for any decomposition $X = V_1 \cup V_2$ with closed subsets $V_i \subset X$, we have either $X = V_1$ or $X = V_2$. A subset $Y \in X$ is called irreducible if Y is an irreducible space with the induced topology. The *Krull dimension* of X denoted as $\dim X$ is the supremum of the lengths n of all chains $V_0 \subset V_2 \subset \ldots V_n$, $V_{i+1} \neq V_i$, of nonempty closed irreducible subsets V_i of X.

Our above problem of determining when the modules $\mathcal{T}$ and $\mathcal{W}$ are free can be answered by calculating $\mu(\mathcal{T})$ and $\mu(\mathcal{W})$ the number of generators in any minimal generating systems of $\mathcal{T}$ and $\mathcal{W}$. In this respect the theorem of Forster and Swan (see

[9]) turns out to be of great importance. We reproduce below only the directly useful consequence of this theorem and refer the reader to [9] for other details.

THEOREM 2. *Let* $X = \max \mathcal{A}$ *be Noetherian and of finite Krull dimension, and let* M *be any finitely generated* $\mathcal{A}$*-module. Then*

$$\mu(M) \leq \dim X + \operatorname{Max} \{\mu_{\mathrm{m}}(M) \mid \mathrm{m} \in X \cap \operatorname{supp} M\}.$$

Hence, if M *is generated by* r *elements for each* m *in* $\max \mathcal{A}$*, then* M *is generated globally by* $r + \dim X$ *elements.*

The application of the above theorem to our problem is now quite obvious and is given in the following.

THEOREM 3. *Let* $\max \mathcal{A}$ *be Noetherian and* $\dim \max \mathcal{A} = 0$*. Then* P *is stabilizable if and only if* P *has a doubly coprime factorization.*

Proof. Only necessity need be proved. By Theorem 1, $\mathcal{T}$ and $\mathcal{W}$ are free of ranks m and n, respectively, at every maximal ideal m (hence also generated by m and n generators at each m). Hence by the theorem of Forster and Swan above, $\mathcal{T}$ and $\mathcal{W}$ are generated globally by m and n generators. But then all of these generators of $\mathcal{T}$, as well as $\mathcal{W}$, must be linearly independent, since $\mathcal{T}$ and $\mathcal{W}$ contain, respectively, m and n linearly independent elements, namely those of the m and n scalar matrices dI with d a nonzero divisor. Thus the modules $\mathcal{T}$ and $\mathcal{W}$ are free of ranks m and n. Consequently, we have DCF by Lemma 3. □

Remark. The above theorem shows that for the class of rings $\mathcal{A}$ that have a Noetherian and zero-dimensional maxspectrum, the stabilization theory is identical to that of the standard factorization theory, since it is only those plants that are stabilizable that have DCF. In the next section, we show that the class of polynomial rings over PIDs is also in another class for which this holds.

3.3. Stabilizability condition over UFDs. We now develop further geometrical interpretation of the condition of Theorem 1. This interpretation can in fact be developed for Noetherian rings with irreducible spectrum; however, to keep matters simple we restrict ourselves only to unique factorization domains (UFDs). Moreover, results over UFDs are directly relevant to our main application of multidimensional stabilization treated in the next section. To begin, consider the following local-global characterization of a projective module. This result is in fact valid for general rings. We refer the reader to [3, Thm. 1(d), p. 110] for the proof.

PROPOSITION 5. *The following statements are equivalent.*

1. *An* $\mathcal{A}$ *module* M *is finitely generated and projective.*

2. *There exists a finite family* $F = \{f_1 \dots f_r\}$ *of elements of* $\mathcal{A}$ *generating* $\mathcal{A}$*, i.e., the ideal* $(f_1 \dots f_r) = \mathcal{A}$ *and* M_f *is a free* A_f*-module of finite rank for all* f *in the finite set* F*.*

Observe that in our problem the modules $\mathcal{T}$ and $\mathcal{W}$ are finitely generated over $\mathcal{A}$ by the rows and columns of matrices T and W, respectively. Hence there already exist m (respectively, n) linearly independent generators of $\mathcal{T}$ (respectively, $\mathcal{W}$). It thus follows that the localizations $\mathcal{T}_{\mathrm{p}}$, respectively, $\mathcal{W}_{\mathrm{p}}$, over the zero prime ideal $\mathrm{p} = 0$ are free $\mathcal{A}_{\mathrm{p}}$-modules of rank m (respectively, n). Thus to prove that $\mathcal{T}$ and $\mathcal{W}$ are projective of ranks m and n it suffices to just prove that they are projective, since the rank is constant over the entire spectrum of an $\mathcal{A}$ a UFD.

From the above characterization of a projective module it follows that the problem of checking whether a plant is stabilizable can be solved, if we can identify the finite families corresponding to our modules $\mathcal{T}$ and $\mathcal{W}$ from the matrices T and W obtained from the fractions of the plant. A step in this direction is provided by the following.

PROPOSITION 6. *There exists a nonzero element h in $\mathcal{A}$ such that $\mathcal{T}_h$ and $\mathcal{W}_h$ are both free $\mathcal{A}_h$-modules of ranks m and n, respectively.*

Proof. Let T_1 be an $m \times m$ nonsingular submatrix of T. Recall that there is at least one. Consider the matrix $T_0 = TT_1^{-1}$ over the field of fractions $\mathcal{F}$ of $\mathcal{A}$. Let the entries of T_0 be considered in the reduced form, with the numerator and the denominator of each entry considered relatively prime after the greatest common denominator between the modulo units of $\mathcal{A}$ is cancelled. Let f be the radical of the least common multiple of all the denominators of T_0. Then for a sufficiently large integer k the matrix $K = f^k T_0$ has all entries in $\mathcal{A}$ and there is also an $m \times m$ submatrix of $f^{-k}K$ with row indices same as that of T_1 such that the determinant of this submatrix is a unit of $\mathcal{A}_f$. This implies a factorization of the form

$$T = f^{-k} K T_1,$$

where entries of $f^{-k}K$ are in $\mathcal{A}_f$ and the $m \times m$ minors of $f^{-k}K$ generate $\mathcal{A}_f$. Moreover, entries of T_1 are also in $\mathcal{A}_f$ and $\det T_1 \neq 0$. Thus it follows that $\mathcal{T}_f$ is a free $\mathcal{A}_f$ module. On similar lines we can construct a nonzero element g in $\mathcal{A}$ from the transposed matrix W^T, such that $\mathcal{W}_g$ is a free $\mathcal{A}_g$ module. Now set $h = \operatorname{rad}(fg)$. □

From the above proposition it follows that it would be useful to collect such elements f and g for the matrices T and W. Let $\{T_1, T_2 \dots T_r\}$ be the family of all nonsingular $m \times m$ submatrices of the matrix T, and for each index j, let f_j be the element corresponding to T_j (as constructed in the proof of the above proposition) with minimum number of irreducible factors such that $\mathcal{T}_{f_j}$ is a free $\mathcal{A}_{f_j}$ module. We call the family $F = \{f_1, f_2 \dots f_r\}$ the family of *elementary factors* of the matrix T. Similarly, let $G = \{g_1, g_2 \dots g_l\}$ denote the family of elementary factors of the matrix W^T. Now let the finite family $H = \{f_i g_j, i = 1 \dots r, j = 1 \dots l\}$ of rl elements be denoted by FG. We call H the *family of elementary factors of the transfer matrix* P. In fact, it can be shown that the closed sets $V(F)$, $V(G)$, and hence also $V(H)$, of the ideals (F), (G), and (H) generated by each of the above respective families of elementary factors remain invariant with respect to the choices of the generators of the modules $\mathcal{T}$ and $\mathcal{W}$. Moreover, two S-equivalent plants have the same $V(H)$. We omit these details as these are not needed further. However this observation shows that $V(H)$ is an invariant of the transfer matrix P.

Our main result now is as follows. We drop the strict causality condition on P as in Theorem 1 in view of the proof of Proposition 1 for integral domains given in §4.4.

THEOREM 4. *The plant transfer matrix P is stabilizable if and only if the family H of elementary factors of P satisfies*

$$\bigcap_{h \in H} V(h) = \emptyset$$

(i.e., the elementary factors of P are coprime).

Proof. We first establish the proof by necessity. By hypothesis, the modules $\mathcal{T}$ and $\mathcal{W}$ are projective of rank m and n, respectively. Hence by Proposition 5, there exist finite families F' and G' both generating $\mathcal{A}$ such that $\mathcal{T}_{f'}$ is a free $\mathcal{A}_{f'}$ module for all f' in F', and $\mathcal{W}_{g'}$ is a free $\mathcal{A}_{g'}$ module for all g' in G'. Let H' be the product family $F'G'$. Now consider the module $\mathcal{T}$ and a member f' of F'. It follows that there exist m linearly independet vectors of $\mathcal{T}$ forming a nonsingular $m \times m$ matrix $V_{f'}$, which extend to provide a basis of $\mathcal{T}_{f'}$ over $\mathcal{A}_{f'}$. Hence for a sufficiently large

integer ν there is a factorization

$$T = f'^{-\nu} K_{f'} V_{f'},$$

where the mth determinantal ideal $I_m(K_{f'})$ contains a unit of $\mathcal{A}_{f'}$. Now, let f in F be an elementary factor of T corresponding to a submatrix T_1. Also, let $K_{f'1}$ denote the corresponding submatrix of $K_{f'}$. Then, in the field of fractions $\mathcal{F}$ of $\mathcal{A}$ we obtain

$$TT_1^{-1} = K_{f'} K_{f'1}^{-1},$$

which shows that the elementary factor f is a factor of $\det K_{f'1}$. Hence the ideal $I_m(K_{f'})$ is contained in the ideal (F) generated by the elemetary factors of T. Since the family F' generates $\mathcal{A}$ and this inclusion holds for every f' in F' we have,

$$V(F) \subset V(F') = \emptyset.$$

On similar lines it can be shown that $V(G) = \emptyset$. Thus, it follows that

$$V(H) = V(F) \cup V(G) = \emptyset.$$

Now we establish the proof by sufficiency. Since $H = FG$, the ideal (H) generated by the family H is equal to the product of ideals $(F)(G)$. Hence $V(H) = V(F) \cup V(G)$ is empty implies that the families F and G both generate $\mathcal{A}$. Since F is the family of elementary factors of T, the module $\mathcal{T}_f$ is free $\mathcal{A}_f$-module for all f in F. Hence by Proposition 5 $\mathcal{T}$ is a projective $\mathcal{A}$-module. Its rank is thus clearly m. Similarly it can be shown that $\mathcal{W}$ is a projective $\mathcal{A}$-module of rank n. P is thus stabilizable by Theorem 1. □

The discussion preceding the above theorem clearly shows that when the family of elementary factors of P contains a unit, then both the modules $\mathcal{T}$ and $\mathcal{W}$ are free and the condition of the theorem is satisfied, which is a fact in confirmation of the standard factorization theory, i.e., the plant is stabilizable when it has a DCF. In general, stabilizabilty may not imply existence of DCF. It can be shown by similar reasoning as employed in the above proof that if P has DCF then the elementary factors of P generate $\mathcal{A}$. It has however not been possible to establish that the DCF may fail to exist for a stabilizable plant.

We now consider a necessary condition for stabilizability as a corollary of the above theorem, which should at times be a simpler check for stabilizabilty of P before we embark on the more exhaustive computation of the elementary factors of P. This is developed via some notation and a simplifying lemma below.

For the matrix T, let $t_i = a_i d_t$ for $i = 1 \ldots r$ denote the nonzero maximal order minors of T, where d_t is the greatest common divisor of t_i. Call $\{a_1 \ldots a_r\}$ the family of *reduced minors* of T. Similarly, let $\{b_1, b_2 \ldots b_l\}$ denote the family of reduced minors of the matrix W. The following lemma is stated without proof, which follows from a straightforward application of the determinant formula but is lengthy.

LEMMA 5. *For any transfer matrix P, the families of reduced minors of T and W are identical modulo units, i.e., $r = l$ and $a_i = b_i$ up to unit factors for $i = 1 \ldots r$.*

We now consider the corollary of Theorem 4.

COROLLARY 2. *If P is stabilizable, then the family of reduced minors of T (and also of W) generates $\mathcal{A}$, i.e.,*

$$(a_1, a_2, \ldots, a_r) = \mathcal{A}.$$

Proof. Stabilizabilty of P by Theorem 4 implies that the family H of elementary factors of P generates $\mathcal{A}$. Since $\mathcal{T}_h$ is a free $\mathcal{A}_h$-module for every h in H, we have a factorization $T = T_h D_h$, where $\det D_h$ is nonzero and the $m \times m$ minors of T_h generate $\mathcal{A}_h$. Thus $\det D_h$ is the greatest common divisor over $\mathcal{A}_h$ of the nonzero $m \times m$ minors of T. Hence these minors of T_h are of the form $u_i a_i$ for $i = 1 \ldots r$, where a_i are the reduced minors of T and u_i are units of $\mathcal{A}_h$. This implies that $\cap_{i=1}^{r} V(a_i) \subset V(h)$ for every h in H. Hence we have

$$\bigcap_{i=1}^{r} V(a_i) \subset \bigcap_{h \in H} V(h) = \emptyset.$$

From Lemma 5 above, a_i are also the reduced minors of W. □

Clearly, the above condition also becomes sufficient, and in that case also implies DCF of P if the reduced minors are also the minors of both the matrices T and W. Thus in this case the maximal minors of T and W are both coprime. Stabilization theory in such cases is thus trivial.

3.4. Formula for all stabilizing controllers. An important landmark of the standard factorization theory is its formula that characterizes all stabilizing controllers of the plant. The development above shows that the plant P always has DCFs over $\mathcal{A}_h$ for all its elementary factors h and becomes stabilizable if and only if these generate the whole ring $\mathcal{A}$. Thus we have the formula of stabilizing controllers for each elementary factor h and the ring $\mathcal{A}_h$. When P is stabilizable, we should thus be able to glue together these formulae to obtain all the stabilizing controllers of P. We carry out this construction in this section.

As before, let $H = \{h_1, h_2, \ldots, h_k\}$ denote the family of elementary factors of the plant P, which we now assume is stabilizable. Since P has a DCF over $\mathcal{A}_h$ for each h in H, it is stabilizable over $\mathcal{A}_h$ by standard factorization theory. Hence using Proposition 1, all its stabilizing controllers are of the form

$$C_h = Y_h X_h^{-1},$$

where the matrices Y_h and X_h have all entries in $\mathcal{A}_h$ and satisfy (1) for some other matrices U_h and W_h also over $\mathcal{A}_h$. Recall that all these matrices can be obtained from the controller formula of the standard factorization theory. Now, for a sufficiently large integer ν, (1), after multiplying by h^ν, becomes

$$\begin{aligned}(h^\nu X_h)N &= (h^\nu U_h)d,\\ (h^\nu Y_h)N &= (h^\nu W_h)d,\\ N(h^\nu Y_h) &= (h^\nu I - (h^\nu X_h))d,\end{aligned}$$

where all the matrices in the parentheses (.) have their entries in $\mathcal{A}$. For index $i = 1, \ldots, k$, let X_i, Y_i, U_i, and W_i be the quadruple of matrices over $\mathcal{A}_i = \mathcal{A}_{h_i}$ satisfying (1). Then for each of this index i there exist integers n_i such that the quadruple of matrices $h_i^{n_i} X_i$, $h_i^{n_i} Y_i$, $h_i^{n_i} U_i$, $h_i^{n_i} W_i$, all have entries in $\mathcal{A}$ and satisfy the above equations.

Now, since P is stabilizable, the elementary factors h_i generate $\mathcal{A}$. Hence there exist elements α_i such that

$$\sum_{i=1}^{k} \alpha_i h_i^{n_i} = 1.$$

Consider the quadruple

$$\begin{aligned} X &= \sum_{i=1}^{k} \alpha_i h_i^{n_i} X_i, \\ Y &= \sum_{i=1}^{k} \alpha_i h_i^{n_i} Y_i, \\ U &= \sum_{i=1}^{k} \alpha_i h_i^{n_i} U_i, \\ W &= \sum_{i=1}^{k} \alpha_i h_i^{n_i} W_i. \end{aligned}$$

Then these matrices satisfy (1) and hence YX^{-1} is a stabilizing controller. Moreover, by Proposition 1, every stabilizing controller can be obtained in this form. This formula thus clearly shows the gluing of the coprime fractions of the stabilizing controllers obtained over each ring $\mathcal{A}_h$ as h varies over the entire set of elementary factors H.

4. Applications. The purpose of this section is to apply the results of the previous sections to examples of integral domains. In particular, the problem of multidimensional stabilization is discussed and numerical examples are provided to illustrate the stabilizability conditions.

4.1. Multidimensional stabilization. One of the problems that motivates the present work is the problem of multidimensional or n-dimensional stabilization. In [2] the matrix problem is solved for the two-dimensional case, while in [13] the scalar problem is solved for the general n-dimensional case. Following [13], we now define the k-dimensional stabilization problem as follows. (Most of the definitions are well known and are given in [9]. Certain specific results are from [13]).

Let $\mathcal{B}$ denote the ring $\mathbf{R}[X_1 \dots X_k]$ of polynomials in k indeterminates over the real field $\mathbf{R}$, and let $V(f)$ (respectively, $V(I)$) denote the algebraic variety of an $f \in \mathcal{B}$ (respectively, an ideal $I \subset \mathcal{B}$) in the affine k-dimensional complex space $A^k(\mathbf{C})$. Let $\Gamma \subset \mathbf{C}^k$ be a *compact symmetric polynomially convex domain.* The domain Γ defines the saturated MC subset

$$S = \{f \in \mathcal{B} \mid V(f) \cap \Gamma = \emptyset\}.$$

The domain Γ is called the *domain of instability.* The polydisc $\bar{U}^n$ considered in [13] is, in fact, an example of such a domain. The following property of such domains, which immediately follows from the developments in [13], is relevant here.

LEMMA 6. *If Γ is a compact symmetric polynomially convex domain, then there is a bijective correspondence between the conjugate pairs of points of Γ and the maximal ideals of $S^{-1}\mathcal{B}$.*

Now returning to the k-dimensional stabilization problem, we consider $\mathcal{A} = S^{-1}\mathcal{B}$ as the ring of stable causal transfer functions and the k-dimensional plant P is an $n \times m$ matrix with entries in $\mathcal{F}$ the field of fractions of $\mathcal{A}$. Since this ring $\mathcal{A}$ is a UFD, the necessary and sufficient condition for stabilizability of such a plant is given by Theorem 4 in the previous section. We thus only consider some numerical examples for applying the condition of Theorem 4 to determine stabilizabilty. From the above lemma, the domain Γ now plays the role of spec $\mathcal{A}$, and the closed sets $V(I)$ for an ideal I are the zeros of generators of I in Γ. We thus denote the zeros in Γ of an element f of $\mathcal{A}$ again by $V(f)$.

Example 1. We have that

$$P = \frac{1}{1 + X_1 X_2} \begin{bmatrix} X_1 & X_2 \\ X_3 & X_4 \\ 1 & X_5 \end{bmatrix}.$$

The reduced minors of matrices T and W are obtained (modulo units) as $X_1X_4 - X_2X_3$, $X_1X_5 - X_2$, $X_2(1 + X_1X_2)$, $X_1(1 + X_1X_2)$, $X_3X_5 - X_4$, $X_3(1 + X_1X_2)$, $X_4(1 + X_1X_2)$, $X_5(1 + X_1X_2)$, $(1 + X_1X_2)$, $(1 + X_1X_2)^2$. Clearly, for $X_1 = 1$, $X_3 = X_5 = -1$ we can find a solution to the equation $X_4 + X_3 = 0$ such that the reduced minors above have a common zero in the polydisc $\bar{U}^5$. Thus this plant does not satisfy the necessary condition for stabilizabity of Corollary 2. Hence P is not stabilizable.

Example 2. We have that

$$P = \frac{1}{X_1^2 + X_2^2 + X_3^2 - 1} \begin{bmatrix} X_1X_2 \\ 1 \end{bmatrix}.$$

In this example, the minors of the matrix T have no common zero in the polydisc $\bar{U}^3$, hence T is free of rank 1. This is clearly observed since one of the elementary factors of T is a unit. The matrix W is obtained as

$$W = \begin{bmatrix} X_1X_2 & d & 0 \\ 1 & 0 & d \end{bmatrix},$$

where $d = X_1^2 + X_2^2 + X_3^2 - 1$. The submatrix $W(1,2)$ formed by the first two columns of W is nonsingular. The elementary factor of W with respect to $W(1,2)$ can be easily observed to be a unit again. Thus P has a unit elementary factor that shows that P has a DCF and is thus stabilizable.

Example 3. Let f and d be polynomials in three indeterminates each having zeros in the polydisc $\bar{U}^3$ but having no common zeros in this polydisc. Thus f and d are coprime in $\mathcal{A}$. Let d have irreducible factors $d = d_1d_2 = (X_1 - 1)(X_1X_2 - X_3)$. Then we have

$$P = \frac{1}{d} \begin{bmatrix} fX_1 & X_3 \\ f & X_2 \end{bmatrix}.$$

After a laborious computation, the elementary factors of both T and W become

$$f \quad d_1X_3 \quad fd_1X_1 \quad d_1X_2 \quad fd_1 \quad d.$$

Clearly, this is also the set of elementary factors of P. Since f and d are coprime, this set also generates $\mathcal{A}$. Thus P is stabilizable.

Above examples show that the condition of Theorem 4 allows a constructive procedure for checking stabilizabilty of multidimensional plants. However, this condition does not indicate whether DCFs exist for stabilizable plants as shown in [2] for the two-dimensional case.

4.2. Stabilization over polynomial rings. We now consider the stabilization problem over the polynomial ring $K[X_1, \ldots, X_k]$, where K is a PID. Thus our ring $\mathcal{A}$ of stable transfer functions is now this polynomial ring $K[X_1, \ldots, X_k]$.

Recall that the celebrated Quillen–Suslin resolution of the well-known Serre's conjecture [9, Thm. 3.15] shows that projective modules over the above class of polynomial rings are free. Hence for this class of rings the stabilizability conditions of Theorem 1 reduce to the following simpler form.

COROLLARY 3. *A strictly causal transfer matrix P over the field of fractions of $K[X_1, \ldots, X_k]$ is stabilizable if and only if the modules $\mathcal{T}$ and $\mathcal{W}$ are free.*

Proof. The proof follows by the Quillen–Suslin theorem [10] and Theorem 1. □

Clearly such transfer matrices also have DCFs. Thus it follows from the above result that for stabilization problems over this class of polynomial rings, the theorem of Forster and Swan and the zero dimensionality condition of Theorem 3 are not needed to show existence of DCF. It follows that, *for this class of rings, the parametrization of the stabilizing controllers is identical to that in the standard factorization theory.* In other words, existence of DCF is necessary for stabilization. Observe that this fact is not available from the standard factorization theory over the class of polynomial rings developed in [14, Chap. 8]. Thus the polynomial rings above are an additional class of rings for which DCFs exist for stabilizable plants apart from the class of rings satisfying conditions of Theorem 3. This is a reason for our motivation in considering the problem over polynomial rings.

The Serre's conjecture referred to above has been investigated for many other classes of rings. In particular, rings K that are not PIDs, but for which the polynomial rings satisfy the conjecture, have been investigated. The well-known reference [10] gives an excellent survey of results and developments on this problem. However, any discussion on the use of further results on the Quillen–Suslin theorem as well as Serre's conjecture are beyond the scope of this paper.

4.3. PID and Bezout domain. These integral domains are the main cases considered in the standard factorization theory. Now clearly, if $\mathcal{A}$ is a PID or a Bezout domain, then the modules $\mathcal{T}$ and $\mathcal{W}$ are automatically free. Hence every matrix P over $\mathcal{F}$ has a DCF and consequently is stabilizable. Thus our results above specialize to the well-known results of standard factorization theory.

4.4. Relaxing strict causality. We now show that when the ring $\mathcal{A}$ is an integral domain, Proposition 1 can be made stronger by relaxing the strict causality of P. Recall that the strict causality of P is required in the necessity part of the proof of Proposition 1.

Thus let (1) be satisfied. Then in the idealizer ring $\mathcal{I}(b_{\mathrm{L}})$, we have $b_{\mathrm{L}}+a_{\mathrm{R}} = \mathcal{I}(b_{\mathrm{L}})$. Now two cases arise.

Case 1. $a_{\mathrm{R}} \neq \mathcal{I}(b_{\mathrm{L}})$. We claim that there is $X \in b_{\mathrm{L}}$ with $\det X \neq 0$ for which $I - X$ belongs to a_{R}. Let $X \in b_{\mathrm{L}}$ and $Y \in a_{\mathrm{R}}$ be such that $X + Y = I$. Now if $\det X = 0$ then since $\det X = 1 - \det Y$, it follows that $\det Y = 1$. Thus a_{R} contains a unimodular matrix. Let Y^{-1} be the inverse of Y in $(\mathcal{A})_n$. Since $Y^{-1} = Y^{-1}X + I$ and as b_{L} is a left ideal of $(\mathcal{A})_n$ it follows that $Y^{-1}X$ belongs to b_{L} hence also in $\mathcal{I}(b_{\mathrm{L}})$. Hence Y^{-1} belongs to $\mathcal{I}(b_{\mathrm{L}})$. Thus as a_{R} is a right ideal of $\mathcal{I}(b_{\mathrm{L}})$, $I = YY^{-1}$ belongs to a_{R} which is a contradiction. Thus $\det X \neq 0$.

Case 2. $a_{\mathrm{R}} = \mathcal{I}(b_{\mathrm{L}})$. Since $d \neq 0$, the ideal $(d : I_1(N)) \neq 0$, where $I_1(N)$ is the ideal generated by all the entries of N, i.e., equal to $\sum(n_{ij})$. Now $(d : \sum(n_{ij})) = \cap_{i,j}(d : n_{ij}) \neq 0$. Choose $x \neq 0$ in this this ideal. Then clearly $xI \in b_{\mathrm{L}}$ and $\det xI = x^n \neq 0$. Thus $(I - xI) \in \mathcal{I}(b_{\mathrm{L}}) = a_{\mathrm{R}}$.

Thus what we have proved is that if (1) has a solution, then every solution X has $\det X \neq 0$. Now the remaining steps verifying that YX^{-1} is a stabilizing controller are identical to those in Proposition 1.

Although as shown above, stabilizability can be shown without using strict causality, it has not been possible to relax this condition for showing the causality of the stabilizing controller for general integral domains.

5. Conclusions. In this paper geometric stabilizability criteria are developed that are invariant with respect to fractions chosen for the transfer matrix. Next, a class of rings of stable transfer functions is determined for which stabilizability is

equivalent to existence of doubly coprime factorization. Thus for all such rings, the stabilization theory is identical to the standard factorization theory. Moreover, it is shown that the class of polynomial rings over PIDs is also an additional class of rings over which stabilizable transfer matrices have doubly coprime fractions. For general unique factorization domains, however, stabilizabilty is determined by the coprimeness of the elementary factors of the plant transfer matrix. The formula parametrizing all stabilizing controllers of a plant in this case also generalizes the well-known formula of the standard factorization theory. The question of whether stabilizable multidimensional plants have doubly coprime fractions remains unanswered in this development.

Acknowledgments. The author is grateful to a reviewer whose suggestions led to important improvements in the paper. He is also thankful Shiva Shankar, Aparna Dar, and Anurag Singh for useful discussions.

REFERENCES

[1] M. F. Atiyah and I. G. Macdonald, *Introduction to Commutative Algebra*, Addison–Wesley, Reading, MA, 1969.

[2] N. K. Bose and J. P. Guiver, *2-D causal and weakly causal filters with applications to stabilization*, in Multidimensional system theory, N. K. Bose, ed., D. Reidel, Boston, MA, 1985.

[3] N. Bourbaki, *Commutative Algebra*, Addison–Wesley, Reading, MA, 1972.

[4] C. A. Desoer, R. W. Liu, J. Murray, and R. Saeks, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 399–412.

[5] R. Gattazzo, *Polynomial matrices with a given determinant*, Linear Algebra Appl., 144 (1991), pp. 107–120.

[6] E. W. Kamen, *Stabilization of linear spatially distributed continuous and discrete time linear systems*, in Multidimensional System Theory, N. K. Bose, ed., D. Reidel, Boston, MA, 1985.

[7] E. W. Kamen and P. P. Khargonekar, *On the control of linear systems whose co-efficients are functions of parameters*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 25–33.

[8] P. P. Khargonekar and E. D. Sontag, *On the relation between stable matrix fraction factorization and regulable realization of linear systems over rings*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 627–638.

[9] E. Kunz, *Introduction to commutative algebra and algebraic geometry*, Birkhauser, Boston, 1985.

[10] T. Y. Lam, *Serre's conjecture*, Lecture Notes in Math., No. 635, Springer-Verlag, Berlin, New York, 1978.

[11] J. C. McConnell and J. C. Robson, *Noncommutative Noetherian Rings*, John Wiley, New York, 1987.

[12] B. R. McDonald, *Linear algebra over commutative rings*, Marcel Dekker, New York, 1984.

[13] Shiva Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.

[14] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT press, Cambridge, MA, 1985.

[15] M. Vidyasagar, H. Schnider, and B. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880–894.

[16] D. C. Youla, J. Bongiorno, and H. A. Jabr, *Modern Wiener Hopf design of optimal controllers part* II, *The multivariable case*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 319–338.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1696–1725, November 1994

011

ROBUST INDIRECT ADAPTIVE CONTROL OF TIME-VARYING PLANTS WITH UNMODELED DYNAMICS AND DISTURBANCES*

SANJEEV M. NAIK† AND P. R. KUMAR†

Abstract. It is shown that indirect pole-zero placement adaptive controllers are robust for systems with time-varying parameters as well as unmodeled dynamics and disturbances. A parameter estimator with projection is used. No special signal normalization is employed to ensure robustness.

The nominal system parameters need only be bounded, and their variations need only be small in an average sense. This allows them to vary slowly with time, as well as to take large jumps occasionally. The adaptive controllers can also simultaneously tolerate small unmodeled dynamics, as well as bounded disturbances, with no restriction on the magnitude of the bound.

Key words. adaptive systems, disturbances, indirect adaptive control, robustness, robust performance, stability, time-varying plants, unmodeled dynamics

AMS subject classifications. 93B55, 93C10, 93C40, 93C41, 93C50, 93C55, 93D21, 93D25, 93D30, 93E12, 93E99

1. Introduction.

In [1] Åström and Wittenmark have proposed the use of indirect, certainty-equivalent self-tuning controllers based on pole-zero placement for the servo-problem. Thus, if the goal is to enforce a response

$$A^*(q^{-1})y_k = q^{-d}B^*(q^{-1})r_k \tag{1}$$

to a command signal $\{r_k\}$, then we first estimate system polynomials $\widehat{A}_{k-1}(q^{-1}) := 1 + \sum_{i=1}^{p} \hat{a}_{i,k-1}q^{-i}$ and $\widehat{B}_{k-1}(q^{-1}) := \sum_{i=0}^{\ell-1} \hat{b}_{i+1,k-1}q^{-i}$ at each time instant k by fitting a model

$$\widehat{A}_k(q^{-1})y_t = \widehat{B}_k(q^{-1})u_{t-d}$$

to the available data $\{y_t, u_{t-1} | 0 \le t \le k\}$ at time k. Then, the control input u_k is chosen such that

$$\widehat{R}_k(q^{-1})\widehat{B}_k(q^{-1})u_k = -\widehat{S}_k(q^{-1})y_k + B^*(q^{-1})r_k, \tag{2}$$

where the polynomials $\widehat{R}_k$ and $\widehat{S}_k$ are obtained by solving the Diophantine equation

$$\widehat{R}_k(q^{-1})\widehat{A}_k(q^{-1}) + q^{-d}\widehat{S}_k(q^{-1}) = A^*(q^{-1}) \tag{3}$$

at each time instant. To estimate the parameter vector $\widehat{\theta}_k := (-\hat{a}_{1,k}, \ldots, -\hat{a}_{p,k}, \hat{b}_{1,k}, \ldots, \hat{b}_{\ell,k})^T$, we may use a recursive algorithm of the form

$$\widehat{\theta}_k = \widehat{\theta}_{k-1} + \Gamma_{k-1}\phi_{k-1}(y_k - \phi_{k-1}^T\widehat{\theta}_{k-1}),$$

where

$$\phi_{k-1} = (y_{k-1}, \ldots, y_{k-p}, u_{k-d}, \ldots, u_{k-\ell-(d-1)})^T. \tag{4}$$

* Received by the editors June 22, 1992; accepted for publication (in revised form) May 4, 1993. This research has been supported in part by Army Research Office contract DAAL 03-91-G-0182 and Joint Services Electronics Program contract N00014-90-J-1270.

† University of Illinois, Coordinated Science Laboratory, 1308 West Main Avenue, Urbana, Illinois 61801.

The choice of the "gain" matrix $\Gamma_k = \gamma_k I$ yields a gradient update law. Control schemes of this nature have proven popular in practice, and many successful implementations have been reported; see [2], [3], [4], [23]. Also, the model reference adaptive control method, the backbone of continuous-time adaptive control, is a "direct" version of such a pole-zero placement scheme.

Often, such adaptive control algorithms are employed to control plants that are subject to time-variations, for which there is consequently an ever-present need to adapt to the changing system characteristics. It is therefore important to develop a theory of robustness for such adaptive control algorithms for their use in the face of such system variations, as well as in the presence of uncertainties such as unmodeled dynamics and disturbances.

In this paper we consider an indirect adaptive control law, as above, with a parameter estimation scheme employing "projection" to keep the parameter estimates confined to a compact convex set. We establish that this simple modification is powerful enough to provide robustness simultaneously with respect to system time-variations, small unmodeled dynamics, and bounded disturbances. In particular, no special signal normalization is used.

Some background on the robustness problem for adaptive systems is in order. Much attention has been devoted over the past decade to the robustness problem caused by unmodeled dynamics and bounded disturbances, and an excellent unification of the work up to 1988 is provided by Ioannou and Sun [5]. Recently, in an important paper, Ydstie [6] has shown that just the simple mechanism of "projection," which confines the parameter estimates to a compact convex set, is sufficient for robustness with respect to bounded disturbances and some unmodeled dynamics. This paper is notable also for the introduction of a new proof technique involving a "switched signal." This work has been extended to continuous-time, while also enlarging the class of unmodeled dynamics, by Naik, Kumar, and Ydstie [7]. The net effect of these investigations is to show that many of the modifications, proposed in the 1980's to prove robustness with respect to unmodeled dynamics, are not necessary. The original simple "projection" modification, established by Egardt [8] to be robust with respect to bounded disturbances, is also robust with respect to small unmodeled dynamics.

Comparatively less attention has been devoted to the problem of robustness with respect to time-variations. Solo [9] established the boundedness of signals for a direct one-step-ahead adaptive control scheme using an a posteriori estimate-based gradient update law with leakage, when the parameters are slowly varying, while Kreisselmeier [10] has done so for an indirect adaptive control scheme with projection for slow-in-the-mean parameter variations, which allows occasional large parameter jumps. Tsakalis and Ioannou [11] obtained similar results for continuous-time plants, while Guo [12] and Meyn and Guo [13] have done so for discrete-time stochastic systems. In these works, while the plant is allowed to vary with time, the effect of unmodeled dynamics is not considered. In [14] de Larminat and Raynaud considered such robustness for a fairly general indirect adaptive control law that uses two parallel estimators as well as a specially constructed "normalization" signal. Middleton and Goodwin [15] also incorporated a normalization signal, and additionally assumed knowledge of the constant factor by which it overbounds the unmodeled dynamics. This constant is then used to set up a normalized dead-zone, for which they established robustness to slow-in-the-mean parameter variations and small unmodeled dynamics. Giri et al. [16] proved the robustness of an adaptive *regulator* for a

plant with small-in-the-mean parameter variations and unmodelled dynamics, using only the knowledge of the order of a nominal plant model. However, this is done using a complicated control law involving an identification–stabilization time-splitting, using a least-squares-based adaptation law that also employs a special normalization signal. Furthermore, arbitrarily large bounded disturbances cannot be handled, and the regulation objective is *not* achieved in the "ideal" case unless the algorithm is appropriately modified.

In all these works, the signals entering the adaptation law are normalized by a specially constructed normalization signal first proposed by Praly [17]. The effect of such normalization is to ensure that the modeling error entering the adaptation law is bounded or small in the mean. Construction of such a normalization signal requires a priori system knowledge and involves extra computation. In addition to these practical considerations, it is of theoretical interest to see if normalization is necessary for robustness.

The key point of this paper is that such normalization is not required to ensure the robustness of adaptive controllers to small unmodeled dynamics, bounded disturbances with arbitrarily large bound, and slow-in-the-mean parameter variations. We show that we obtain robust adaptive control by merely utilizing projection, together with "extended regressor" normalization, without recourse to any other modifications.

An indirect adaptive pole-zero placement controller is considered. The resulting necessity to cancel process zeros requires the nominal time-varying plant to be minimum phase at every instant. On the other hand, the problem of potential loss of estimated model controllability/stabilizability faced in adaptive pole placement is not an issue here. This itself results in a simpler adaptive control scheme than those in [14], [16], and [18], for instance. A result similar to ours has recently been reported by Wen [21] using a different proof technique developed earlier by Wen and Hill [22]. They consider a unit delay plant with the time-variations restricted to be slow. The true nominal time-varying parameter vector is assumed to lie in a convex compact set, which is assumed to have the property that the nominal system polynomials induced by any parameter vector in the set are uniformly coprime. This restrictive assumption is the consequence of choosing an indirect pole-placement controller design. A final point worth noting is that unlike the signal bounds derived in the present paper, those in [21] are not uniform in that they depend on the system initial condition.

Our main results are the following.

(i) A certainty equivalent adaptive controller, using a gradient based parameter estimator with projection and employing normalization based on an "extended regressor" ensures that all closed-loop signals are bounded, when applied to a nominally minimum phase discrete-time plant with slow-in-the-mean parameter variations with bounded disturbances and small unmodelled dynamics (Theorem 1).

(ii) In the absence of unmodeled dynamics and disturbances, and in case the parameter variations asymptotically tend to zero, i.e., in the nominal case, the error in tracking a reference trajectory converges to zero (Theorem 2). When unmodeled dynamics as well as bounded disturbances are present, the mean-squared output prediction error is linear in the magnitude of the unmodelled dynamics, bounded disturbances, and average rate of parameter variations (Theorem 3). Thus the adaptive controller provides robust performance in addition to robust boundedness.

2. System description. Consider a plant

$$y_k + \sum_{i=1}^{p} a_{i,k-1} y_{k-i} = \sum_{j=1}^{\ell} b_{j,k-1} u_{k-j+1-d} + v_k, \tag{5}$$

where u and y denote the input and output, while v represents the cumulative effect of unmodeled dynamics and disturbances. The coefficients $\{a_{i,k}, b_{i,k}\}$ may vary with time, and so the system is allowed to be time-varying. We wish to investigate the robustness of indirect adaptive control schemes for such systems that simultaneously consist of time-variations, unmodeled dynamics, and disturbances.

Let us suppose that the goal of adaptive control is to generate a closed-loop response to a command signal r_k, which satisfies (1), where q^{-1} denotes the usual backward-shift operator, i.e., $q^{-1}z_k := z_{k-1}$. If the plant in (5) were time-invariant and "ideal" without any unmodeled dynamics or disturbances, i.e., given by $A(q^{-1})y_k = q^{-d}B(q^{-1})u_k$, then the control law

$$R(q^{-1})B(q^{-1})u_k = -S(q^{-1})y_k + B^*(q^{-1})r_k,$$

where $R(q^{-1})$ and $S(q^{-1})$ satisfy the Diophantine equation

$$A(q^{-1})R(q^{-1}) + q^{-d}S(q^{-1}) = A^*(q^{-1}), \tag{6}$$

results in the cancellation of all original process zeros (hence necessitating some sort of minimum-phase condition) and a closed-loop transfer-function $q^{-d}B^*(q^{-1})/A^*(q^{-1})$ from r_k to y_k, as desired.

We consider here a certainty-equivalent indirect adaptive control scheme, which first forms parameter estimates

$$\widehat{\theta}_k = (-\hat{a}_{1,k}, \ldots, -\hat{a}_{p,k}, \hat{b}_{1,k}, \ldots, \hat{b}_{\ell,k})^T \tag{7}$$

at each time-step k, and then uses the resulting estimated polynomials $\widehat{A}_{k-1}(q^{-1}) := 1 + \sum_{i=1}^{p} \hat{a}_{i,k-1}q^{-i}$ and $\widehat{B}_{k-1}(q^{-1}) := \sum_{i=0}^{\ell-1} \hat{b}_{i+1,k-1}q^{-i}$ to solve the Diophantine equation

$$\widehat{R}_k(q^{-1})\widehat{A}_k(q^{-1}) + q^{-d}\widehat{S}_k(q^{-1}) = A^*(q^{-1}) \tag{8}$$

for $\widehat{R}_k$ and $\widehat{S}_k$.[1] We consider the minimum degree solution for $\widehat{R}_{k-1}$, so that it is monic and of degree $(d-1)$. We note that such a solution can be found through a $(d-1)$-step long division of the polynomial A^* by $\hat{A}$. Then we apply the input u_k, given by

$$\widehat{R}_k(q^{-1})\widehat{B}_k(q^{-1})u_k = -\widehat{S}_k(q^{-1})y_k + B^*(q^{-1})r_k. \tag{9}$$

It is well known in adaptive control (see Egardt [8]) that if we simply use a pure gradient-based parameter estimator, then the resulting adaptive system is destabilized by even a small bounded disturbance. It is therefore necessary to somehow modify

[1] Since we are dealing with time-varying polynomials in a shift-operator, we distinguish between the notations $C_k(q^{-1})D_k(q^{-1}) := \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} c_{i,k}d_{j,k}q^{-(i+j)}$ and $C_k(q^{-1}) \circ D_k(q^{-1}) := \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} c_{i,k}d_{j,k-i}q^{-(i+j)}$ when multiplying $C_k(q^{-1}) = \sum_{i=1}^{\ell} c_{i,k}q^{-i}$ and $D_k(q^{-1}) = \sum_{j=1}^{\ell} d_{j,k}q^{-j}$. For a signal x_k, $C_k(q^{-1})x_k \equiv C_k(q^{-1}) \circ x_k$ denotes $\sum_{i=1}^{\ell} c_{i,k}x_{k-i}$, as is usual.

the parameter estimator to secure robustness with respect to even just bounded disturbances.

We consider a parameter update law using parameter projection. This modification, motivated by the seminal work of Egardt [8], simply projects the parameter estimate vector onto a compact convex set $\mathcal{C}$ at each time step k. The set $\mathcal{C}$ is chosen such that

$$((p+1)\text{th component of } \theta) \geq b_{\min} > 0 \text{ for every } \theta \in \mathcal{C}.$$

(By thus ensuring that the estimated leading coefficient of $\widehat{B}_k(q^{-1})$ is positive, we also ensure that the control law (9) is well defined, since it does not involve division by zero.) The parameter estimates are recursively specified by

$$\widehat{\theta}'_k = \widehat{\theta}_{k-1} + \frac{\mu \phi_{k-1} e_k}{1 + \|\psi_{k-1}\|^2}, \tag{10}$$

$$e_k := y_k - \phi_{k-1}^T \widehat{\theta}_{k-1}, \tag{11}$$

and

$$\widehat{\theta}_k = \text{ Proj }_{\mathcal{C}}[\widehat{\theta}'_k], \tag{12}$$

where the regressor ϕ is given by (4). Here Proj $_{\mathcal{C}}[\cdot]$ denotes "orthogonal" projection onto the compact convex set $\mathcal{C}$, defined uniquely by Proj $_{\mathcal{C}}[x] \in \mathcal{C}$, and $\|$ Proj $_{\mathcal{C}}[x] - x\| \leq \|y - x\|, \forall y \in \mathcal{C}$. The vector ψ_k is an "extension" of the regression vector ϕ_k defined in (4)

$$\psi_{k-1} := (y_{k-1}, \ldots, y_{k-p'-d-m'+1}, u_{k-1}, \ldots, u_{k-\ell-2d-n'+2})^T, \tag{13}$$

where $p' := p, m' := 0, n' := 0$, if $d = 1$; and $p' := \max\{p, \deg(A^*)\}$, $m' := \max\{0, p - d\}$, $n' := \max\{0, p' - d + 1\}$, if $d > 1$.[2] Above, the constant μ in the step-size is chosen such that $0 < \mu < 2$, and the algorithm is initialized with $\hat{\theta}_0 \in \mathcal{C}$. We refer to this as the *parameter estimator with projection* (PEP).

Remarks. All the results of this paper can be extended to least-squares type parameter estimators, and also to "direct" pole-zero placement schemes. In fact, though we do not show it here, if we use a direct one-step-ahead adaptive control law, then robustness can be established easily through a similar analysis.

3. The assumptions. Our goal is to analyze the behavior of these adaptive controllers when they are applied to plants of the form shown in (5), which consist simultaneously of time-variations, unmodelled dynamics, and disturbances. We make the following assumptions on the plant (5) and the reference model (1).

Assumption 1. $\theta_k \in \mathcal{C}$ for all k, where $\mathcal{C}$ is a compact convex set such that the $(p+1)$th component of every vector in $\mathcal{C}$ is larger than or equal to $b_{\min} > 0$. Let K_θ be a constant that bounds $\|\theta\|$ for all $\theta \in \mathcal{C}$.

Assumption 2. The zeros of $B_k(q^{-1}) = \sum_{i=1}^{\ell} b_{i,k} q^{-i+1} = 0$ lie in the open disk $|q|^2 < \sigma < 1$, for all k.

[2] When the delay is unity, i.e., $d = 1$, then $\psi = \phi$, i.e., the extended regressor is identical to the regressor.

Assumption 3. $\sum_{k=t+1}^{t+T} \|\theta_k - \theta_{k-1}\| \leq K_\delta + k_\delta T$ for all $t, T \geq 0$, for some constants $K_\delta, k_\delta > 0$. Here $\theta_k := (a_{1,k}, \ldots, a_{p,k}, b_{1,k}, \ldots, b_{\ell,k})^T$ is the time-varying parameter vector.

Assumption 4. $v_k^2 \leq K_v m_{k-1} + k_v$ for all k, for some $K_v, k_v > 0$, where m_k satisfies a recursion

$$m_k = \sigma m_{k-1} + K_y y_k^2 + K_u u_k^2 + K_3, \qquad m_0 > 0, \tag{14}$$

and σ is as in Assumption 2.

Assumption 5. $A^*(q^{-1}) = 1 + \sum_1^{p'} a_i^* q^{-i}$ has all its zeros in the open disk $|q|^2 < \sigma < 1$, with σ as in Assumption 2.[3]

Assumptions 1 and 2 require that the time-varying parameter vector θ_k be bounded, have leading b_1-coefficient uniformly bounded away from zero, and be strictly minimum-phase, at every time-instant k. This latter restriction is necessitated by the requirement of canceling all process zeros. Assumption 3 allows slow-in-the-mean parameter time-variations, and thus occasional jumps too. Assumption 4 allows small unmodelled dynamics, and bounded disturbances with arbitrarily large bound. Finally, Assumption 5 simply requires the reference model to be stable with a prescribed margin of stability.

Remark. It is worthwhile to note that Assumption 4 includes the case of an incorrect assumption on the delay, since m_{k-1} includes input terms up to time $k-1$. Furthermore, the class of unmodeled dynamics covered by Assumption 4 allows the true plant to be nonminimum phase [19], [20].

In order to implement the above adaptive control schemes, we thus need to know upper bounds p and ℓ on the orders of the *nominal* time-varying portion of the plant, the nominal delay d, and the sign (say, positive) and a lower bound $b_{\min} > 0$ on the leading term in the polynomial $B_k(q^{-1})$ modeling the numerator of the nominal portion of the plant, for every k. In addition, since we must project the parameter estimates onto a compact convex set guaranteed to contain the true parameter θ_k for all k, we must know the bound K_θ on the norms of the time-varying parameters $\|\theta_k\|$ for all k. Finally, we must know the reference model given by $A^*(q^{-1})$, $B^*(q^{-1})$, and r_k.

Our central result in this paper is that under the above assumptions, the adaptive control laws are stable for all K_v and k_δ small enough, i.e., whenever the unmodeled dynamics are small enough, the parameter variations are small enough on the average, and the disturbances are bounded, though without any restriction on the magnitude of the bound.

4. Properties of the parameter estimators. We now derive some important properties of the parameter estimator PEP, which are independent of the control laws used. Let us denote by $\widetilde{\theta}_k := \widehat{\theta}_k - \theta_k$ the parameter estimation error. Using the definitions of ϕ_k and θ_k in (4) and Assumption 3, we can express the plant (5) as

$$y_k = \phi_{k-1}^T \theta_{k-1} + v_k. \tag{15}$$

Further, substituting (15) in (11) gives

$$e_k = -\phi_{k-1}^T \widetilde{\theta}_{k-1} + v_k. \tag{16}$$

LEMMA 1. *The following properties hold for the parameter estimator PEP.*

[3] Without loss of generality, if $\deg(A^*) < p'$, we set $a_i^* = 0$ for $i = \deg(A^*) + 1, \ldots, p'$.

(i) $\hat{b}_{k,1} \geq b_{\min} > 0, \forall k \geq 0$.

(ii) *The parameter estimation errors are uniformly bounded, i.e.,* $\|\widetilde{\theta}_k\| \leq K_{\widetilde{\theta}}$ *for all* k, *where*

$$K_{\widetilde{\theta}} := 2K_\theta.$$

(iii) *The parameter estimates* $\{\widehat{\theta}_k\}$ *are uniformly bounded also, with* $\|\widehat{\theta}_k\| \leq K_\theta$.

(iv) *Let* ϵ *satisfy* $0 < \epsilon < 2 - \mu$. *Then*

$$\sum_{k=t+1}^{t+T} \frac{e_k^2}{\rho_{k-1}} \leq K_{ev} \sum_{k=t+1}^{t+T} \frac{v_k^2}{\rho_{k-1}} + k_e T + K_e. \tag{17}$$

The quantity ρ_k *is defined as*

$$\rho_{k-1} := 1 + \|\psi_{k-1}\|^2, \tag{18}$$

and the constants are given by,

$$K_{ev} := \frac{\left(1 + \frac{1}{\epsilon}\right)}{2 - \mu - \epsilon}, \qquad k_e := \frac{2(K_\theta + K_{\widetilde{\theta}})(1+\mu)}{\mu(2-\mu-\epsilon)} k_\delta,$$

and

$$K_e := \frac{K_{\widetilde{\theta}}^2 + 2(K_\theta + K_{\widetilde{\theta}})(1+\mu)K_\delta}{\mu(2-\mu-\epsilon)}.$$

Proof. Define $\widetilde{\theta}'_k := \widehat{\theta}'_k - \theta_k$, $\overline{\phi}_{k-1} := \phi_{k-1}/\sqrt{\rho_{k-1}}$, $\bar{e}_k := e_k/\sqrt{\rho_{k-1}}$, and $\bar{v}_k := v_k/\sqrt{\rho_{k-1}}$.

(i) This is immediate from (12), and the property Assumption 1 of $\mathcal{C}$.

(ii) This is immediate from Assumption 1 and (12).

(iii) This follows from (ii) by Assumption 1.

(iv) Let us define $\delta_k := \theta_k - \theta_{k-1}$. By (10),

$$\widetilde{\theta}'_k = \widetilde{\theta}_{k-1} - \delta_k + \mu \overline{\phi}_{k-1} \bar{e}_k. \tag{19}$$

Using $\|\overline{\phi}_{k-1}\|^2 \leq 1$ and (16), we obtain

$$2\mu \overline{\phi}_{k-1}^T \widetilde{\theta}_{k-1} \bar{e}_k = 2\mu(\bar{v}_k - \bar{e}_k)\bar{e}_k \leq \frac{\mu \bar{v}_k^2}{\epsilon} + \mu\epsilon \bar{e}_k^2 - 2\mu \bar{e}_k^2.$$

Using (19) and noting that $\|\widetilde{\theta}'_k\| \leq \|\widetilde{\theta}_k\|$ due to projection, we obtain

(20)

$$\|\widetilde{\theta}_k\|^2 \leq \|\widetilde{\theta}_{k-1}\|^2 + \|\delta_k\|^2 + \mu^2 \bar{e}_k^2 - 2\widetilde{\theta}_{k-1}^T \delta_k - 2\mu \overline{\phi}_{k-1}^T \delta_k \bar{e}_k + \frac{\mu \bar{v}_k^2}{\epsilon} + \mu\epsilon \bar{e}_k^2 - 2\mu \bar{e}_k^2.$$

Thus, using (16) we obtain,

$$\begin{aligned}
\|\delta_k\|^2 - 2\delta_k^T(\widetilde{\theta}_{k-1} + \mu \overline{\phi}_{k-1} \bar{e}_k) &\leq \|\delta_k\|^2 + 2\|\delta_k\| \left(\|\widetilde{\theta}_{k-1}\| + \mu |\bar{e}_k| \right) \\
&\leq \|\delta_k\|^2 + 2\|\delta_k\| \left(K_{\widetilde{\theta}} + \mu |\bar{v}_k - \overline{\phi}_{k-1}^T \widetilde{\theta}_{k-1}| \right) \\
&\leq \|\delta_k\|^2 + 2\|\delta_k\| \left(K_{\widetilde{\theta}} + \mu |\bar{v}_k| + \mu \|\widetilde{\theta}_{k-1}\| \right) \\
&\leq \|\delta_k\|^2 + 2\|\delta_k\|(1+\mu)K_{\widetilde{\theta}} + \mu(\|\delta_k\|^2 + \bar{v}_k^2) \\
&\leq \mu \bar{v}_k^2 + (1+\mu)\|\delta_k\|(\|\delta_k\| + 2K_{\widetilde{\theta}}) \\
&\leq \mu \bar{v}_k^2 + 2(1+\mu)\|\delta_k\|(K_\theta + K_{\widetilde{\theta}}).
\end{aligned}$$

Substituting this into (20) gives

$$\mu(2-\mu-\epsilon)\bar{e}_k^2 \le \|\widetilde{\theta}_{k-1}\|^2 - \|\widetilde{\theta}_k\|^2 + \|\delta_k\|^2 + \frac{\mu}{\epsilon}\bar{v}_k^2 - 2\delta_k^T(\widetilde{\theta}_{k-1} + \mu\overline{\phi}_{k-1}\bar{e}_k)$$
$$\le \|\widetilde{\theta}_{k-1}\|^2 - \|\widetilde{\theta}_k\|^2 + \mu\left(1+\frac{1}{\epsilon}\right)\bar{v}_k^2 + 2(1+\mu)(K_\theta + K_{\widetilde{\theta}})\|\delta_k\|.$$

Summing from $t+1$ to $t+T$, telescoping, and using Assumption 3 as well as $\|\widetilde{\theta}_k\|^2 \le K_{\widetilde{\theta}}^2$ gives the desired result. □

The boundedness of parameter estimates yields the following important result that the controller parameters are Lipschitz functions of the plant parameters.

LEMMA 2. *Let $\theta^c :=$ (coefficients of $R(q^{-1})$, coefficients of $S(q^{-1}))^T$ denote the "controller" parameter vector, where R and S are functions of $\theta =$ (coefficients of $A(q^{-1};\theta)$, coefficients of $B(q^{-1};\theta))^T$ through (6). Let $\theta^{c,1}$ and $\theta^{c,2}$ denote the controller parameter vectors corresponding to two different plant parameter vectors θ^1 and θ^2, respectively. Then*

$$\|\theta^{c,1} - \theta^{c,2}\| \le K(C)\|\theta^1 - \theta^2\|$$

for all θ^1, $\theta^2 \in L_C := \{\theta \in R^{p+\ell} : \|\theta\| \le C\}$, where $K(C)$ is a constant that only depends on C.

Proof. Relation (6) can be rewritten as a system of linear equations

$$M(\theta^j)\theta^{c,j} = a^*, \qquad j = 1, 2,$$

where $M(\theta^j)$ denotes the Sylvester matrix corresponding to the polynomials $A(q^{-1})$ and q^{-d}, and a^* is formed from the coefficients of $A^*(q^{-1})$. Then, A being monic implies that there exists a positive ϵ such that, for all θ in L_C, $|\det(M(\theta^j))| \ge \epsilon$. We therefore obtain

$$\begin{aligned}\|\theta^{c,1} - \theta^{c,2}\| &\le \|M^{-1}(\theta^1) - M^{-1}(\theta^2)\| \cdot \|a^*\| \\ &\le \|M^{-1}(\theta^1)\| \cdot \|M(\theta^1) - M(\theta^2)\| \cdot \|M^{-1}(\theta^2)\| \cdot \|a^*\| \\ &\le \frac{(\|M(\theta^1)\| \cdot \|M(\theta^2)\|)^{p+\ell-1}}{|\det(M(\theta^1))\det(M(\theta^2))|}\|M(\theta^1) - M(\theta^2)\| \cdot \|a^*\| \\ &\le K(C)\|M(\theta^1) - M(\theta^2)\| \\ &\le K(C)\|\theta^1 - \theta^2\|,\end{aligned}$$

thereby concluding the proof. □

5. The switched system. We now introduce, for purposes of analysis only, the "switched" system,

(21)
$$z_k = I_{k-1}(\sigma z_{k-1} + K_y e_k^2 + K_u u_{k-1}^2 + K_3) + (1 - I_{k-1})(g z_{k-1} + 2K_3), \quad z_0 > 0,$$

where $0 < \sigma < g < 1$, and the indicator function I_{k-1} is defined as

$$I_{k-1} = \begin{cases} 1 & \text{if } \sigma z_{k-1} + K_y e_k^2 + K_u u_{k-1}^2 + K_3 \ge g z_{k-1} + 2K_3, \\ 0 & \text{otherwise.} \end{cases}$$

LEMMA 3. *For all k,*

(i) $\|\phi_k\|^2 \leq Km_k$, $\|\psi_k\|^2 \leq Km_k$;[4]
(ii) $m_k \leq K_{mz} z_k + k_{mz}$;
(iii) $z_k \leq K_z z_{k-1} + k_z$;
(iv) $v_k^2 \leq K_{vz} z_{k-1} + k_{vz}$;
(v) $u_k^2 \leq K\rho_{k-1} + K_{uz} z_{k-1} + K$;
(vi) $\rho_k \leq K\rho_{k-1} + K_{\rho z} z_{k-1} + K$;
(vii) $\rho_k \leq K_{zk} + K$.

Above, the constants K_{vz}, K_{uz}, and $K_{\rho z}$ can all be made as small as desired by choosing K_v small enough.

Proof. (i) These are obvious from (4), (13), and (14), where we also note that by choosing m_0 large, the constant K can be chosen independently of the initial conditions.

(ii) From

$$m_k = \sigma^k m_0 + \sum_{j=0}^{k} \sigma^{k-j}[K_y y_j^2 + K_u u_j^2 + K_3],$$

$$z_k \geq \sigma^k z_0 + \sum_{j=0}^{k} \sigma^{k-j}[K_y e_j^2 + K_u u_{j-1}^2 + K_3],$$

it is clear that it suffices to show that

$$\sum_{j=0}^{k} \sigma^{k-j} y_j^2 \leq K \sum_{j=0}^{k} \sigma^{k-j} e_j^2 + K \sum_{j=0}^{k} \sigma^{k-j} u_{j-1}^2 + K + K'\sigma^k, \quad \text{and} \tag{22}$$

$$\sum_{j=0}^{k} \sigma^{k-j} u_j^2 \leq K \sum_{j=0}^{k} \sigma^{k-j} y_j^2 + K \sum_{j=0}^{k} \sigma^{k-j} u_{j-1}^2 + K + K'\sigma^k, \tag{23}$$

where K' denotes a constant that may depend on the initial conditions, since the effects of the initial conditions can then be accounted for by making z_0 appropriately large, thus ensuring that K_{mz} and k_{mz} do not depend on the initial conditions.

For simplicity, let us first consider the case $d = 1$. In that case, $\widehat{R}_{k-1}(q^{-1}) = 1$ and $\widehat{S}_{k-1}(q^{-1}) = q(A^*(q^{-1}) - \widehat{A}_{k-1}(q^{-1}))$. Hence, dropping q^{-1} for brevity, and recalling the notation for multiplying time-varying polynomials in the shift operator, we obtain

$$\begin{aligned} A^* y_k &= (\widehat{R}_{k-1}\widehat{A}_{k-1} + q^{-1}\widehat{S}_{k-1}) y_k \\ &= \widehat{A}_{k-1} y_k - \widehat{R}_{k-1}\widehat{B}_{k-1} u_{k-1} + B^* r_{k-1} \quad \text{(from (9))} \\ &= q^{-1} B^* r_k + y_k - \phi_{k-1}^T \widehat{\theta}_{k-1} \quad (\text{since } \widehat{A}_{k-1} y_k - \widehat{B}_{k-1} u_{k-1} = y_k - \phi_{k-1}^T \widehat{\theta}_{k-1}) \\ &= q^{-1} B^* r_k + e_k. \end{aligned} \tag{24}$$

From Assumption 5 it follows that there exist constants $0 < \gamma, \delta < 1$ such that $|q|^2 < \gamma^2 < \gamma^{1+\delta} < \sigma < 1$ for every root q of $A^*(q^{-1}) = 0$. Using the boundedness of

[4] In the remainder of the paper, we use the symbol K to generically denote any positive constant that does not depend on either K_v or k_δ, and whose exact value is unimportant for the proofs which follow.

$\{r_k\}$ it follows from (24) that

$$\sum_{j=0}^{k} \sigma^{k-j} y_j^2 \le K + K'\sigma^k + K \sum_{i=0}^{k} \sigma^{k-i} e_i^2, \tag{25}$$

thus proving (22). For (23), recall from (9) that

$$u_k = \frac{1}{\hat{b}_{1,k}} \left[-\sum_{j=2}^{\ell} \hat{b}_{j,k} u_{k-j+1} - \sum_{j=0}^{\deg(\widehat{S})} \hat{s}_{j,k} y_{k-j} + B^*(q^{-1}) r_k \right]. \tag{26}$$

Since $\hat{b}_{1,k} > b_{\min} > 0$, $\widehat{\theta}_k$ is bounded, and the coefficients of $\widehat{S}_k(q^{-1})$ depend continuously on $\widehat{\theta}_k$, (23) follows from (26).

Now let us turn to the case $d > 1$. In that case,

$$\begin{aligned}
\widehat{A}_{k-1} y_k &= q^{-d} \widehat{B}_{k-1} u_k + (\widehat{A}_{k-1} - A_{k-1}) y_k + q^{-d}(B_{k-1} - \widehat{B}_{k-1}) u_k \\
&\quad + A_{k-1} y_k - q^{-d} B_{k-1} u_k \\
&= q^{-d} \widehat{B}_{k-1} u_k - \phi_{k-1}^T \tilde{\theta}_{k-1} + v_k \\
&= \widehat{B}_{k-1} u_{k-d} + e_k.
\end{aligned}$$

Hence operating on both sides by $\widehat{R}_{k-1}$, and keeping in mind the consequences of multiplying time-varying polynomials in the shift-operator and the associated notation, we have

$$\begin{aligned}
\widehat{R}_{k-1} e_k &= \widehat{R}_{k-1} \circ e_k = \widehat{R}_{k-1} \circ [\widehat{A}_{k-1} y_k - \widehat{B}_{k-1} u_{k-d}] \\
&= \widehat{R}_{k-1} \widehat{A}_{k-1} y_k - \widehat{R}_{k-1} \widehat{B}_{k-1} u_{k-d} - (\widehat{R}_{k-1} \widehat{A}_{k-1} - \widehat{R}_{k-1} \circ \widehat{A}_{k-1}) y_k \\
&\quad + (\widehat{R}_{k-1} \widehat{B}_{k-1} - \widehat{R}_{k-1} \circ \widehat{B}_{k-1}) u_{k-d} \\
&= (A^* - q^{-d} \widehat{S}_{k-1}) y_k - \widehat{R}_{k-d} \widehat{B}_{k-d} u_{k-d} + (\widehat{R}_{k-d} \widehat{B}_{k-d} - \widehat{R}_{k-1} \widehat{B}_{k-1}) u_{k-d} \\
&\quad - (\widehat{R}_{k-1} \widehat{A}_{k-1} - \widehat{R}_{k-1} \circ \widehat{A}_{k-1}) y_k + (\widehat{R}_{k-1} \widehat{B}_{k-1} - \widehat{R}_{k-1} \circ \widehat{B}_{k-1}) u_{k-d} \\
&= A^* y_k + (\widehat{S}_{k-d} - \widehat{S}_{k-1}) y_{k-d} - B^* r_{k-d} \\
&\quad + (\widehat{R}_{k-d} \widehat{B}_{k-d} - \widehat{R}_{k-1} \widehat{B}_{k-1}) u_{k-d} - (\widehat{R}_{k-1} \widehat{A}_{k-1} - \widehat{R}_{k-1} \circ \widehat{A}_{k-1}) y_k \\
&\quad + (\widehat{R}_{k-1} \widehat{B}_{k-1} - \widehat{R}_{k-1} \circ \widehat{B}_{k-1}) u_{k-d} \quad \text{(using (9))}.
\end{aligned}$$

Hence,

$$\begin{aligned}
A^* y_k &= q^{-d} B^* r_k + \widehat{R}_{k-1} e_k + (\widehat{R}_{k-1} \widehat{A}_{k-1} - \widehat{R}_{k-1} \circ \widehat{A}_{k-1}) y_k \\
&\quad + (\widehat{S}_{k-1} - \widehat{S}_{k-d}) y_{k-d} - (\widehat{R}_{k-1} \widehat{B}_{k-1} - \widehat{R}_{k-1} \circ \widehat{B}_{k-1}) u_{k-d} \qquad (27) \\
&\quad + (\widehat{R}_{k-1} \widehat{B}_{k-1} - \widehat{R}_{k-d} \widehat{B}_{k-d}) u_{k-d}. \qquad (28)
\end{aligned}$$

Now note that

$$\begin{aligned}
&|(\widehat{R}_{k-1} \widehat{A}_{k-1} - \widehat{R}_{k-1} \circ \widehat{A}_{k-1}) y_k - (\widehat{R}_{k-1} \widehat{B}_{k-1} - \widehat{R}_{k-1} \circ \widehat{B}_{k-1}) u_{k-d} \\
&\quad + (\widehat{R}_{k-1} \widehat{B}_{k-1} - \widehat{R}_{k-d} \widehat{B}_{k-d}) u_{k-d} + (\widehat{S}_{k-1} - \widehat{S}_{k-d}) y_{k-d}| \\
&\le \left| \sum_{i=1}^{d-1} \sum_{j=1}^{p} \hat{r}_{i,k-1} (\hat{a}_{j,k-1} - \hat{a}_{j,k-i-1}) y_{k-i-j} \right|
\end{aligned}$$

$$
\begin{aligned}
&+\left|\sum_{i=1}^{d-1}\sum_{j=1}^{\ell}\hat{r}_{i,k-1}(\hat{b}_{j,k-1}-\hat{b}_{j,k-i-1})u_{k-i-j-d+1}\right| \\
&+\left|\sum_{i=1}^{d-1}(\widehat{R}_{k-i}\widehat{B}_{k-i}-\widehat{R}_{k-i-1}\widehat{B}_{k-i-1})u_{k-d}\right|+\left|\sum_{i=1}^{d-1}(\widehat{S}_{k-i}-\widehat{S}_{k-i-1})y_{k-d}\right| \\
&\leq K\sum_{i=1}^{d-1}\sum_{j=1}^{p}\sum_{n=1}^{i}|\hat{a}_{j,k-n}-\hat{a}_{j,k-n-1}||y_{k-i-j}| \\
&+K\sum_{i=1}^{d-1}\sum_{j=1}^{\ell}\sum_{n=1}^{i}|\hat{b}_{j,k-n}-\hat{b}_{j,k-n-1}||u_{k-i-j-d+1}| \\
&+\left|\sum_{i=1}^{d-1}(\widehat{R}_{k-i}-\widehat{R}_{k-i-1})\widehat{B}_{k-i}u_{k-d}+(\widehat{B}_{k-i}-\widehat{B}_{k-i-1})\widehat{R}_{k-i-1}u_{k-d}\right| \\
&+\left|\sum_{i=1}^{d-1}\sum_{j=0}^{\deg(\widehat{S})}(\hat{s}_{j,k-i}-\hat{s}_{j,k-i-1})y_{k-d-j}\right| \\
&\leq K\sum_{i=1}^{d-1}\sum_{j=1}^{p}\sum_{n=1}^{i}\|\widehat{\theta}_{k-n}-\widehat{\theta}_{k-n-1}\|\|\psi_{k-1-n}\| \\
&+K\sum_{i=1}^{d-1}\sum_{j=1}^{\ell}\sum_{n=1}^{i}\|\widehat{\theta}_{k-n}-\widehat{\theta}_{k-n-1}\|\|\psi_{k-1-n}\| \\
&+\left|\sum_{i=1}^{d-1}\sum_{j=1}^{d-1}\sum_{m=1}^{\ell}(\hat{r}_{j,k-i}-\hat{r}_{j,k-i-1})\hat{b}_{m,k-i}u_{k-d-j-m+1}\right. \\
&\qquad\left.+\sum_{i=1}^{d-1}\sum_{m=1}^{\ell}\sum_{j=1}^{d-1}(\hat{b}_{m,k-i}-\hat{b}_{m,k-i-1})\hat{r}_{j,k-i-1}u_{k-d-j-m+1}\right| \\
&+K\sum_{i=1}^{d-1}\sum_{j=0}^{\deg(\widehat{S})}\|\widehat{\theta}_{k-i}-\widehat{\theta}_{k-i-1}\||y_{k-d-j}| \\
&\leq K\sum_{i=1}^{d-1}\sum_{j=1}^{p}\sum_{n=1}^{i}\|\widehat{\theta}'_{k-n}-\widehat{\theta}_{k-n-1}\|\|\psi_{k-1-n}\| \\
&+K\sum_{i=1}^{d-1}\sum_{j=1}^{\ell}\sum_{n=1}^{i}\|\widehat{\theta}'_{k-n}-\widehat{\theta}_{k-n-1}\|\|\psi_{k-1-n}\| \\
&+K\sum_{i=1}^{d-1}\sum_{j=1}^{d-1}\sum_{m=1}^{\ell}\|\widehat{\theta}'_{k-i}-\widehat{\theta}_{k-i-1}\|\|\psi_{k-i-1}\| \\
&+K\sum_{i=1}^{d-1}\sum_{j=0}^{\deg(\widehat{S})}\|\widehat{\theta}'_{k-i}-\widehat{\theta}_{k-i-1}\|\|\psi_{k-i-1}\| \\
&\leq K\sum_{n=1}^{d-1}\|\widehat{\theta}'_{k-n}-\widehat{\theta}_{k-n-1}\|\|\psi_{k-1-n}\|\leq K\sum_{n=1}^{d-1}|e_{k-n}|,
\end{aligned}
$$

where we have used Lemma 2. Now employing this bound in (27), we can establish (22) just as in the case of $d = 1$. Result (23) again follows from (22) by (26).

(iii) From (21), $z_k \le g z_{k-1} + K_y e_k^2 + K_u u_{k-1}^2 + 2K_3$. Thus we only need to bound e_k^2 and u_{k-1}^2 in terms of z_{k-1}. For the first, by Assumption 4 and (i), it follows from (16) that

$$\begin{aligned} e_k^2 &\le 2K_{\widetilde{\theta}}^2 \|\phi_{k-1}\|^2 + 2K_v m_{k-1} + 2k_v \le 2(K_{\widetilde{\theta}}^2 K + K_v) m_{k-1} + 2k_v \\ &\le 2K_{mz}(K_{\widetilde{\theta}}^2 K + K_v) z_{k-1} + 2(K_{\widetilde{\theta}}^2 K + K_v) k_{mz} + 2k_v. \end{aligned}$$

Also, from (14), $u_{k-1}^2 \le m_{k-1}/K_u \le (K_{mz} z_{k-1} + k_{mz})/K_u$.

(iv) From Assumption 4 and (ii), $v_k^2 \le K_v(K_{mz} z_{k-1} + k_{mz}) + k_v$.

(v) From the control law (26), and the boundedness of $\{r_k\}$,

$$u_k^2 \le K[\sum_{i=1}^{\ell-1} u_{k-i}^2 + \sum_{i=0}^{\deg(\widehat{S})} y_{k-i}^2 + 1].$$

Hence, $u_k^2 \le K\rho_{k-1} + K y_k^2 + K$. Since $y_k = \phi_{k-1}^T \theta_{k-1} + v_k$, we obtain

$$\begin{aligned} y_k^2 &\le K\rho_{k-1} + 2v_k^2 \\ &\le K\rho_{k-1} + KK_v z_{k-1} + K, \end{aligned} \tag{29}$$

and the required bound follows.

(vi) Clearly, $\rho_k \le K\rho_{k-1} + K y_k^2 + K u_k^2$. Hence, from (29) and (v), $\rho_k \le K\rho_{k-1} + K(K_{uz} + K_{vz}) z_{k-1} + K$.

(vii) Using (i) and (ii) of Lemma 3 gives $\|\psi_k\|^2 \le K z_k + K$. Hence $\rho_k \le K z_k + K$ for PEP. □

Now we examine the implications of the switching mechanism in more detail.

LEMMA 4. (i) *There exists a constant* $\epsilon_{\rho z} > 0$ *such that* $\rho_k \ge \epsilon_{\rho z} z_k - k_{\rho z}$, *whenever* $I_k = 1$.

(ii) *For every positive integer* N, *there exist positive constants* $L(N), \bar{k}(N)$, *and* $K_{v\max}(N)$ *with the following property. If* $K_v \in [0, K_{v\max}]$, $I_{t_1} = 1$, *and* $z_k \ge L$ *for all* $k \in [t_1 - N, t_1]$, *then* $\rho_k \ge \bar{k}(N) z_k$ *for all* $k \in [t_1 - N, t_1]$.

Proof. (i) Suppose $I_{k-1} = 1$. Then, from the definition of I_{k-1}, $K_y e_k^2 + K_u u_{k-1}^2 \ge (g - \sigma) z_{k-1} + K_3$. Hence, we have

$$K_y e_k^2 + K_u \rho_{k-1} \ge (g - \sigma) z_{k-1} + K_3. \tag{30}$$

Now, if $\rho_{k-1} \ge ((g - \sigma) z_{k-1} + K_3)/2K_u$, then the claim is true with $\epsilon_{\rho z} > 0$ chosen smaller than $(g - \sigma)/2K_u$. So let us consider only the interesting case, $\rho_{k-1} \le ((g - \sigma) z_{k-1} + K_3)/2K_u$. From (30) we then have $e_k^2 \ge ((g - \sigma) z_{k-1} + K_3)/2K_y$, while from (16), $e_k^2 \le 2K_{\widetilde{\theta}}^2 \|\phi_{k-1}\|^2 + 2v_k^2$, and so

$$2K_{\widetilde{\theta}}^2 \|\phi_{k-1}\|^2 \ge e_k^2 - 2v_k^2 \ge \frac{(g - \sigma) z_{k-1} + K_3}{2K_y} - 2K_{vz} z_{k-1} - 2k_{vz}.$$

Since $\rho_{k-1} \ge \|\phi_{k-1}\|^2$, the required result follows by choosing K_{vz} small enough, and $\epsilon_{\rho z} < (K_\phi / 2K_{\widetilde{\theta}}^2)((g - \sigma)/2K_y - 2K_{vz})$.

(ii) We bound the growth rate of ρ_k / z_k, and then use this in a reversed time argument. Consider $k \in [t_1 - N, t_1]$. Then

$$\frac{\rho_{k+1}}{z_{k+1}} \le \frac{K\rho_k + K_{\rho z 1} z_{k+1} + K_{\rho z} z_k + K}{z_{k+1}},$$

by Lemma 3(vi), where we set $K_{\rho z1} := 0$ in the case of PEP. Hence,

$$\begin{aligned}\frac{\rho_{k+1}}{z_{k+1}} &\le \frac{K}{\sigma}\frac{\rho_k}{z_k} + \left(K_{\rho z1} + \frac{K_{\rho z}}{\sigma}\right) + \frac{K}{\sigma L} \\ &= K_a \frac{\rho_k}{z_k} + K_b \quad (\text{since } z_{k+1} \ge \sigma z_k \text{ and } z_k \ge L),\end{aligned}$$

where K_a and K_b are appropriately chosen. So,

$$\begin{aligned}\frac{\rho_{t_1}}{z_{t_1}} &\le K_a^{t_1 - t}\left[\frac{\rho_t}{z_t} + \frac{K_b}{K_a - 1}\right], \qquad \forall t \in [t_1 - N, t_1], \\ \frac{\rho_{t_1}}{z_{t_1}} &\le K_a^{N}\left[\frac{\rho_t}{z_t} + \frac{K_b}{K_a - 1}\right], \qquad \forall t \in [t_1 - N, t_1].\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\rho_t}{z_t} &\ge K_a^{-N}\frac{\rho_{t_1}}{z_{t_1}} - \frac{K_b}{K_a - 1} \\ &\ge K_a^{-N}\left(\epsilon_{\rho z} - \frac{k_{\rho z}}{L}\right) - \frac{K_b}{K_a - 1} \quad (\text{by (i) and since } z_{t_1} \ge L).\end{aligned}$$

Now we simply define

$$\bar{k}(N) := K_a^{-N}\left[\epsilon_{\rho z} - \frac{k_{\rho z}}{L}\right] - \frac{[K_{\rho z1} + \frac{K_{\rho z}}{\sigma} + \frac{K}{\sigma L}]}{K_a - 1},$$

noting that by choosing $L = L(N)$ large enough and $K_{\rho z1}, K_{\rho z}$ small enough (by choosing K_v small enough), we can ensure that $\bar{k}(N) > 0$. □

6. A representation of the closed-loop system. In this section, we obtain a nonminimal state space description of the closed-loop system consisting of a stable state-transition matrix, and driven by a composite input consisting of the output prediction error, the unmodelled dynamics and disturbances, the filtered reference input, and in the case of $d > 1$, also of "small" fractions of the input and output.

Consider first, for simplicity, the case $d = 1$. Then, using (24),

$$\begin{aligned}B_{k-1}(q^{-1})u_{k-1} &= A_{k-1}(q^{-1})y_k - v_k \\ &= A_{k-1}(q^{-1})y_k - v_k - [A^*(q^{-1})y_k - B^*(q^{-1})r_{k-1} - e_k] \\ &= [A_{k-1}(q^{-1}) - A^*(q^{-1})]y_k + e_k - v_k + B^*(q^{-1})r_{k-1}.\end{aligned}$$

Hence,

$$\begin{aligned}u_{k-1} &= \frac{1}{b_{1,k-1}}[A_{k-1}(q^{-1}) - A^*(q^{-1})]y_k + \frac{e_k - v_k}{b_{1,k-1}} \\ &\quad + \frac{B^*(q^{-1})r_{k-1}}{b_{1,k-1}} - \frac{(b_{2,k-1}u_{k-2} + \ldots + b_{\ell,k-1}u_{k-\ell})}{b_{1,k-1}} \\ &= -b'_{2,k-1}u_{k-2} - \cdots - b'_{\ell,k-1}u_{k-\ell} + a'_{1,k-1}y_{k-1} + \cdots + a'_{p',k-1}y_{k-p'} \\ &\quad + \frac{e_k - v_k}{b_{1,k-1}} + r'_{k-1},\end{aligned}$$

where b'_{k-1}, a'_{k-1} and r'_{k-1} are defined appropriately. Hence, defining the "state-vector"

$$x_k = [y_k, y_{k-1}, \ldots, y_{k-p'}, u_{k-1}, \ldots, u_{k-\ell+1}]^T,$$

we obtain the closed-loop system representation

$$x_k = F_k x_{k-1} + b_{e,k} e_k + b_{v,k} v_k + b_{r,k} r'_{k-1}, \tag{31}$$

where

$$F_k := \begin{bmatrix} J & 0 \\ G_k & H_k \end{bmatrix},$$

$$J = \begin{bmatrix} -a_1^* & \cdots & -a_{p'}^* \\ & & 0 \\ & \mathrm{I} & \vdots \\ & & 0 \end{bmatrix},$$

$$H_k = \begin{bmatrix} -b'_{2,k-1} & \cdots & -b'_{\ell,k-1} \\ & & 0 \\ & \mathrm{I} & \vdots \\ & & 0 \end{bmatrix},$$

$$G_k = \begin{bmatrix} a'_{1,k-1} & \cdots & a'_{p',k-1} \\ & & 0 \\ & 0 & \vdots \\ & & 0 \end{bmatrix},$$

$$b_{e,k} = [1\ 0\ \cdots\ 0\ 1/b_{1,k-1}\ 0\ \cdots\ 0]^T,$$
$$b_{v,k} = [0\ \cdots 0\ -1/b_{1,k-1}\ 0\ \cdots\ 0]^T,$$
$$b_{r,k} = [b_{1,k-1}\ 0\ \cdots\ 0\ 1\ 0\ \cdots\ 0]^T.$$

We note, for future use, that the eigenvalues of F_k, at each instant k, are the roots of $A^*(q^{-1})$ and $B_{k-1}(q^{-1})$. Hence, by Assumptions 2 and 5, F_k is a stable matrix with all eigenvalues λ_i lying in the disk $|\lambda_i|^2 < \sigma < 1$.

Now let us turn to the case of $d > 1$. First, recall that (27) gives

$$A^* y_k = q^{-d} B^* r_k + \widehat{R}_{k-1} e_k + \Delta_{1,k} u_{k-d} + \Delta_{2,k} u_{k-d} + \Delta_{3,k} y_k + \Delta_{4,k} y_{k-d}, \tag{32}$$

where $\Delta_{1,k} := -(\widehat{R}_{k-1}\widehat{B}_{k-1} - \widehat{R}_{k-1} \circ \widehat{B}_{k-1})$, $\Delta_{2,k} := \widehat{R}_{k-1}\widehat{B}_{k-1} - \widehat{R}_{k-d}\widehat{B}_{k-d}$, $\Delta_{3,k} := \widehat{R}_{k-1}\widehat{A}_{k-1} - \widehat{R}_{k-1} \circ \widehat{A}_{k-1}$, and $\Delta_{4,k} := \widehat{S}_{k-1} - \widehat{S}_{k-d}$.

Next, operating by $\widehat{A}_k$ on (9), adding $q^{-d}\widehat{S}_k\widehat{B}_k u_k$ to both sides, and then using (8) gives

$$A^* \widehat{B}_k u_k = (\widehat{A}_k \widehat{R}_k \widehat{B}_k - \widehat{A}_k \circ (\widehat{R}_k \widehat{B}_k)) u_k + q^{-d} \widehat{S}_k \widehat{B}_k u_k - \widehat{A}_k \circ \widehat{S}_k y_k + \widehat{A}_k B^* r_k.$$

This implies

$$\begin{aligned} A^* B_k u_k = {} & -A^* \widetilde{B}_{k-1+d} u_k + A^* (\widetilde{B}_{k-1+d} - \widetilde{B}_k) u_k + (\widehat{A}_k \widehat{R}_k \widehat{B}_k - \widehat{A}_k \circ (\widehat{R}_k \widehat{B}_k)) u_k \\ & + \widehat{A}_k B^* r_k + (\widehat{A}_k \widehat{S}_k - \widehat{A}_k \circ \widehat{S}_k) y_k + \widehat{S}_k (-\widehat{A}_{k-1} y_k + q^{-d} \widehat{B}_{k-1} u_k) \\ & + \widehat{S}_k (q^{-d} (\widehat{B}_k - \widehat{B}_{k-1}) u_k - (\widehat{A}_k - \widehat{A}_{k-1}) y_k), \end{aligned} \tag{33}$$

where $\widetilde{B}_k := \widehat{B}_k - B_k$. Next, note that by (4), (7), (15), and (16),

$$\begin{aligned} -\widehat{A}_{k-1} y_k + q^{-d} \widehat{B}_{k-1} u_k &= -y_k + \phi_{k-1}^T \widehat{\theta}_{k-1} \\ &= -e_k, \end{aligned} \tag{34}$$

and that by (34) and (5),

$$\begin{aligned} -\widetilde{B}_{k-1+d}u_k &= \widetilde{A}_{k-1+d}y_{k+d} - \widetilde{B}_{k-1+d}u_k - \widetilde{A}_{k-1+d}y_{k+d} \\ &= e_{k+d} - v_{k+d} - \widetilde{A}_{k-1+d}y_{k+d}, \end{aligned} \tag{35}$$

where $\widetilde{A}_k := \widehat{A}_k - A_k$. Using (34) and (35) in (33), and defining $\Delta_{5,k} := A^*(\widetilde{B}_{k-1} - \widetilde{B}_{k-d})$, $\Delta_{6,k} := (\widehat{A}_{k-d}\widehat{R}_{k-d}\widehat{B}_{k-d} - \widehat{A}_{k-d} \circ (\widehat{R}_{k-d}\widehat{B}_{k-d}))$, $\Delta_{7,k} := (\widehat{A}_{k-d}\widehat{S}_{k-d} - \widehat{A}_{k-d} \circ \widehat{S}_{k-d})$, $\Delta_{8,k} := \widehat{S}_{k-d}(\widehat{B}_{k-d} - \widehat{B}_{k-d-1})$, and $\Delta_{9,k} := \widehat{S}_{k-d}(\widehat{A}_{k-d} - \widehat{A}_{k-d-1})$, we thus obtain

$$\begin{aligned} A^*B_{k-d}u_{k-d} = &-A^*\widetilde{A}_{k-1}y_k + A^*(e_k - v_k) + \Delta_{5,k}u_{k-d} + \Delta_{6,k}u_{k-d} + \widehat{A}_{k-d}B^*r_{k-d} \\ &+\Delta_{7,k}y_{k-d} - \widehat{S}_{k-d}e_{k-d} + \Delta_{8,k}u_{k-2d} - \Delta_{9,k}y_{k-d}. \end{aligned} \tag{36}$$

We can now use (36) and (32) to obtain the required closed-loop system representation. Define

$$\begin{aligned} x_k &:= [y_k, \ldots, y_{k-p'-\max(p,d)+1}, u_{k-d}, \ldots, u_{k-p'-\ell-d+1}]^T, \\ l_{1,k} &:= q^{-d}B^*r_k + \widehat{R}_{k-1}e_k + \Delta_{1,k}u_{k-d} + \Delta_{2,k}u_{k-d} + \Delta_{3,k}y_k + \Delta_{4,k}y_{k-d}, \\ l_{2,k} &:= A^*(e_k - v_k) + \Delta_{5,k}u_{k-d} + \Delta_{6,k}u_{k-d} + \widehat{A}_{k-d}B^*r_{k-d} \\ &\quad + \Delta_{7,k}y_{k-d} - \widehat{S}_{k-d}e_{k-d} + \Delta_{8,k}u_{k-2d} - \Delta_{9,k}y_{k-d}, \quad \text{and} \\ C_k(q^{-1}) &:= A^*(q^{-1})B_k(q^{-1}) \\ &=: b_{1,k}(1 + c'_{2,k}q^{-1} + \cdots + c'_{p'+\ell+1,k}q^{-(p'+\ell)}). \end{aligned}$$

Note that by Assumptions 2 and 5, all roots of $C_k(q^{-1})$ lie in the open disk $|q|^2 < \sigma < 1$ for all $k \geq 1$.

Using the above definitions, we obtain the representation

$$x_k = F_k x_{k-1} + b_1 l_{1,k} + b_2 l_{2,k}, \tag{37}$$

where

$$F_k := \begin{bmatrix} J' & 0 \\ G'_k & H'_k \end{bmatrix},$$

$$J' = \begin{bmatrix} -a_1^* & \cdots & -a_{p'}^* & 0 & \cdots & 0 \\ & & & & & 0 \\ & & \mathrm{I} & & & \vdots \\ & & & & & 0 \end{bmatrix},$$

$$G'_k = \begin{bmatrix} \star & \cdots & \star \\ & & 0 \\ & 0 & \vdots \\ & & 0 \end{bmatrix},$$

$$H'_k = \begin{bmatrix} -c'_{2,k-1} & \cdots & -c'_{p'+\ell+1,k-1} \\ & & 0 \\ & \mathrm{I} & \vdots \\ & & 0 \end{bmatrix},$$

$$b_1 = [1, 0, \ldots, 0]^T,$$

$$b_2 = [0, \ldots, 0, 1, 0, \ldots, 0]^T,$$

and the $\star$'s represent nonzero scalars whose exact values are unimportant. Note that the eigenvalues of F_k are the zeros of $A^*(q^{-1})$ and the zeros of $B_{k-1}(q^{-1})$. One difference from (31) is that for the case $d > 1$, (37) is driven by terms involving the input and output. However, these occur only in products with the Δ_i terms, which are either parameter or parameter estimate differences or "swapping" terms.

7. The contraction property. We consider the "composite" Lyapunov function

$$W_k := k_F x_k^T P_k x_k + z_k,$$

where $k_F > 0$, and $P_k = P_k^T > 0$ satisfies the discrete-time Lyapunov equation

$$F_k^T P_k F_k - P_k = -I.$$

We note that since the F_k's lie in a compact set and each F_k has all its eigenvalues inside the disk of radius $\sigma^{1/2}$, we have $\|F_k^n\| \le \epsilon_F \gamma^n \le \epsilon_F \sigma^{n/2}$ for some $\epsilon_F > 0$, for all n, k. Hence, $P_k = \sum_{j=0}^{\infty} (F_k^T)^j (F_k)^j \le \lambda_m I$, where $\lambda_m := \epsilon_F^2/(1-\sigma)$.

In what follows, we first show that W_k has a bounded growth rate, and then that W_k has a certain "contraction" property, namely, $W_{k+T} < W_k$ for a certain T whenever W_k is large enough. These are then used to prove the boundedness of W_k, and hence of all signals in the closed-loop system, thus establishing "robust boundedness" of the adaptive system.

LEMMA 5 (Bounded Growth Rate of W). *There exist constants K_w and k_w such that*

$$W_k \le K_w W_{k-1} + k_w \quad \textit{for all } k.$$

Proof. First,

$$\begin{aligned} W_k &\le \frac{k_F \epsilon_F \|x_k\|^2}{1-\sigma} + z_k \\ &\le \frac{k_F \epsilon_F}{1-\sigma} \left(y_k^2 + \|\phi_{k-1}\|^2 + \sum_{j=p+1}^{p'+\max\{p,d\}-1} y_{k-j}^2 + \sum_{j=\ell+d}^{\ell+p'+d-1} u_{k-j}^2 \right) + z_k. \end{aligned}$$

Hence, for some K_{wz} and k_{wz},

$$W_k \le K_{wz} z_k + k_{wz} \le K_{wz}(K_z z_{k-1} + k_z) + k_{wz} \le K_{wz} K_z W_{k-1} + K_{wz} k_z + k_{wz}. \qquad \square$$

LEMMA 6 (the Key Lemma). *For every constant L large enough, whenever there is an interval $[a,b]$ with $W_k \ge 2K_{wz}L$ for all $k \in [a,b]$, the following properties hold.*

(i) $z_k \le W_k \le 2K_{wz} z_k$ *for all* $k \in [a,b]$.

(ii) $W_k \le K_{ww} W_{k-1}$ *for all* $k \in [a+1, b+1]$, *where* $K_{ww} := K_w + \frac{k_w}{2K_{wz}L}$.

(iii) *If* $I_{k-1} = 0$ *for all* $k \in [a,b]$, *then*

$$W_b \le 2K_{wz} \left(g^{b-a} + \frac{2K_3}{(1-g)L} \right) W_a.$$

(iv) *Let* $\overline{\gamma} \in (\max(1 - 1/\lambda_m, g), 1)$. *Let* $I_j = 1$ *for all* $j \in [a,b]$ *or, if* $I_j = 0$ *for any* $j \in [a,b]$, *then suppose there exists an* N *such that* $I_{j+n} = 1$ *for some* $n \in [0, N]$. *Then there exists* $0 < \lambda < 1$ *such that*

$$W_b \le K \exp[-(b-a)\lambda] \left[1 + \frac{K}{L\overline{\gamma}^{b-a}} \right] W_a.$$

Proof. (i) We already have $z_k \leq W_k$, by the definition of W_k. To show that $W_k \leq 2K_{wz}z_k$ for L large enough, we note that because $W_k \leq K_{wz}z_k + k_{wz}$ and $W_k \geq 2K_{wz}L$,

$$z_k \geq \frac{2K_{wz}L - k_{wz}}{K_{wz}} = 2L - \frac{k_{wz}}{K_{wz}}.$$

Hence if we choose L large enough that $k_{wz} \leq K_{wz}[2L - \frac{k_{wz}}{K_{wz}}]$, then $k_{wz} \leq K_{wz}z_k$, thus yielding $W_k \leq 2K_{wz}z_k$.

(ii) This follows easily from the lemma above.

(iii) Since $I_{k-1} = 0$, we have $z_k = gz_{k-1} + 2K_3$ for $a \leq k \leq b$. Hence,

$$\begin{aligned} z_b &= g^{b-a}z_a + 2K_3 \sum_{j=a+1}^{b} g^{b-j} \leq g^{b-a}W_a + \frac{2K_3}{(1-g)} \\ &\leq \left[g^{b-a} + \frac{2K_3}{L(1-g)}\right] W_a, \end{aligned} \tag{38}$$

since $2K_{wz} \geq 1$ implies $W_a \geq 2K_{wz}L \geq L$. Hence $W_b \leq 2K_{wz}z_b \leq 2K_{wz}[g^{b-a} + 2K_3/L(1-g)]W_a$.

(iv) First, let us consider the case $d = 1$ for simplicity. We note that

$$\begin{aligned} W_k &= k_F x_k^T P_k x_k + z_k \\ &\leq k_F(F_k x_{k-1} + b_{e,k}e_k + l_k)^T P_k(F_k x_{k-1} + b_{e,k}e_k + l_k) + gz_{k-1} \\ &\quad + K_y e_k^2 + K_u u_{k-1}^2 + 2K_3 \quad (\text{with } l_k := b_{v,k}v_k + b_{r,k}r'_{k-1}) \\ &\leq k_F x_{k-1}^T (P_k - I)x_{k-1} + 2k_F x_{k-1}^T F_k^T P_k b_{e,k}e_k + 2k_F x_{k-1}^T F_k^T P_k l_k \\ &\quad + 2k_F b_{e,k}^T P_k l_k e_k + k_F b_{e,k}^T P_k b_{e,k} e_k^2 + k_F l_k^T P_k l_k \\ &\quad + gz_{k-1} + K_y e_k^2 + K_u u_{k-1}^2 + 2K_3. \end{aligned} \tag{39}$$

Now defining $\gamma_1 := \sup_k \|F_k^T P_k b_{e,k}\|$, $\gamma_2 := \sup_k \|F_k^T P_k\|$, $\gamma_3 := \sup_k \|P_k b_{e,k}\|$, and since $P_k \leq \lambda_m I$, we have the following inequalities:

$$\left|2k_F x_{k-1}^T F_k^T P_k b_{e,k} e_k\right| \leq 2\gamma_1 k_F \|x_{k-1}\||e_k| \leq \gamma_1 k_F \left[\epsilon_5 \|x_{k-1}\|^2 + \frac{1}{\epsilon_5} e_k^2\right],$$

$$\left|2k_F x_{k-1}^T F_k^T P_k l_k\right| \leq \gamma_2 k_F \left(\epsilon_6 \|x_{k-1}\|^2 + \frac{1}{\epsilon_6}\|l_k\|^2\right),$$

$$\left|2k_F b_{e,k}^T P_k l_k e_k\right| \leq \gamma_3 k_F \left(\|l_k\|^2 + e_k^2\right),$$

and

$$u_{k-1}^2 \leq \frac{1}{b_{\min}^2}\left(K + K(K_\theta + K_{\widetilde{\theta}})^2 \|x_{k-1}\|^2\right) \quad (\text{from (26)}). \tag{40}$$

Hence,

$$\begin{aligned} W_k &\leq k_F x_{k-1}^T P_{k-1} x_{k-1} + gz_{k-1} + k_F\left[-1 + \epsilon_5\gamma_1 + \epsilon_6\gamma_2 + \frac{K_u K(K_\theta + K_{\widetilde{\theta}})}{b_{\min}^2 k_F}\right]\|x_{k-1}\|^2 \\ &\quad + k_F x_{k-1}^T (P_k - P_{k-1})x_{k-1} + \left[\frac{\gamma_1 k_F}{\epsilon_5} + \gamma_3 k_F + \left(1 + \frac{1}{b_{\min}^2}\right)k_F\lambda_m + K_y\right] e_k^2 \\ &\quad + \left[\frac{\gamma_2 k_F}{\epsilon_6} + \gamma_3 k_F + k_F\lambda_m\right]\|l_k\|^2 + 2K_3 + \frac{KK_u}{b_{\min}^u}. \end{aligned}$$

Now, choose $\overline{\gamma} \in (\max(1 - 1/\lambda_m, g), 1)$, k_F large enough, and $\epsilon_5, \epsilon_6 > 0$ small enough so that

$$-1 + \epsilon_5\gamma_1 + \epsilon_6\gamma_2 + \frac{K_u K(K_\theta + K_{\tilde{\theta}})^2}{b_{\min}^2 k_F} \leq -(1 - \overline{\gamma})\lambda_m, \tag{41}$$

and define $K_{ee} := \gamma_1 k_F/\epsilon_5 + \gamma_3 k_F + (1 + 1/b_{\min}^2)k_F\lambda_m + K_y$, $K_l := \gamma_2 k_F/\epsilon_6 + \gamma_3 k_F + k_F\lambda_m$ and $\overline{K}_3 := K_3 + KK_u/2b_{\min}^2$. This gives

$$\begin{aligned} W_k \leq\ & k_F x_{k-1}^T P_{k-1} x_{k-1} + g z_{k-1} - k_F(1 - \overline{\gamma})\lambda_m \|x_{k-1}\|^2 \\ & + k_F x_{k-1}^T (P_k - P_{k-1}) x_{k-1} + K_{ee} e_k^2 + K_l \|l_k\|^2 + 2\overline{K}_3. \end{aligned}$$

However,

$$\begin{aligned} k_F x_{k-1}^T P_{k-1} x_{k-1} + g z_{k-1} &= \overline{\gamma} W_{k-1} + k_F(1 - \overline{\gamma}) x_{k-1}^T P_{k-1} x_{k-1} + (g - \overline{\gamma}) z_{k-1} \\ &\leq \overline{\gamma} W_{k-1} + k_F(1 - \overline{\gamma})\lambda_m \|x_{k-1}\|^2, \end{aligned}$$

and so,

$$W_k \leq \overline{\gamma} W_{k-1} + k_F x_{k-1}^T (P_{k-1} - P_{k-1}) x_{k-1} + K_{ee} e_k^2 + K_l \|\ell_k\|^2 + 2\overline{K}_3.$$

Now focusing on $x_{k-1}^T (P_k - P_{k-1}) x_{k-1}$, we have

$$\begin{aligned} x_{k-1}^T (P_k - P_{k-1}) x_{k-1} &= x_{k-1}^T \left[\sum_{i=0}^{\infty} (F_k^T)^i F_k^i - \sum_{i=0}^{\infty} (F_{k-1}^T)^i F_{k-1}^i \right] x_{k-1} \\ &= \sum_{i=0}^{\infty} \left(\|F_k^i x_{k-1}\|^2 - \|F_{k-1}^i x_{k-1}\|^2 \right) \\ &= \sum_{i=0}^{\infty} (\|F_k^i x_{k-1}\| + \|F_{k-1}^i x_{k-1}\|)(\|F_k^i x_{k-1}\| - \|F_{k-1}^i x_{k-1}\|) \\ &\leq \sum_{i=0}^{\infty} 2\epsilon_F \sigma^{i/2} \|x_{k-1}\| (\|F_k^i x_{k-1}\| - \|F_{k-1}^i x_{k-1}\|). \end{aligned}$$

Now note that

$$\begin{aligned} \|F_k^i x_{k-1}\| - \|F_{k-1}^i x_{k-1}\| &\leq \|(F_k^i - F_{k-1}^i) x_{k-1}\| \leq \|F_k^i - F_{k-1}^i\| \|x_{k-1}\| \\ &= \|F_k^i - F_k^{i-1} F_{k-1} + F_k^{i-1} F_{k-1} - F_k^{i-2} F_{k-1}^2 \\ &\qquad + F_k^{i-2} F_{k-1}^2 \ldots - F_{k-1}^i\| \|x_{k-1}\| \\ &\leq \sum_{j=0}^{i-1} \|F_k^{i-j} F_{k-1}^j - F_k^{i-1-j} F_{k-1}^{j+1}\| \|x_{k-1}\| \\ &\leq \sum_{j=0}^{i-1} \|F_k^{i-1-j}\| \|F_{k-1}^j\| \|F_k - F_{k-1}\| \|x_{k-1}\| \\ &\leq \epsilon_F^2 \sum_{j=0}^{i-1} \sigma^{(i-1-j)/2} \sigma^{j/2} K \|\delta_k\| \|x_{k-1}\| \\ &= i\sigma^{(i-1)/2} \epsilon_F^2 K \|\delta_k\| \|x_{k-1}\|. \end{aligned}$$

Hence

$$x_{k-1}^T(P_k - P_{k-1})x_{k-1} \le K\epsilon_F^3 \sum_{i=0}^{\infty} 2i\sigma^{(2i-1)/2}\|\delta_k\|\|x_{k-1}\|^2 \le K\epsilon_F^3\|\delta_k\|\|x_{k-1}\|^2$$
$$\le K\epsilon_F^3\|\delta_k\|z_{k-1}.$$

Thus we have

$$W_k \le \overline{\gamma}W_{k-1} + Kk_F\epsilon_F^3\|\delta_k\|z_{k-1} + K_{ee}e_k^2 + K_l\|l_k\|^2 + 2\overline{K}_3$$
$$\le \left[\overline{\gamma} + Kk_F\epsilon_F^3\|\delta_k\| + K_{ee}\frac{e_k^2}{z_{k-1}} + K_l\frac{\|l_k\|^2}{z_{k-1}}\right]W_{k-1} + 2\overline{K}_3,$$

i.e., $W_k \le g_k W_{k-1} + 2\overline{K}_3$, where g_k is the term in the square brackets above.

From this recursive bound, we obtain

$$W_b \le \left(\prod_{j=a+1}^{b} g_j\right)W_a + 2\overline{K}_3\sum_{t=a+1}^{b}\left(\prod_{j=t+1}^{b} g_j\right),$$

and so

$$\frac{W_b}{W_a} \le \left(\prod_{j=a+1}^{b} g_j\right)\left[1 + \frac{2\overline{K}_3}{W_a}\sum_{t=a+1}^{b}\left(\prod_{j=a+1}^{t} g_j\right)^{-1}\right]$$
$$\le e^{\sum_{j=a+1}^{b}\ln g_j}\left[1 + \frac{\overline{K}_3}{K_{wz}L}\sum_{t=a+1}^{b}\left(\frac{1}{\overline{\gamma}}\right)^{t-a}\right] \tag{42}$$
$$\le e^{\sum_{j=a+1}^{b}\ln g_j}\left[1 + \frac{\overline{K}_3}{K_{wz}L}\frac{1}{\overline{\gamma}^{b-a}(1-\overline{\gamma})}\right].$$

Now, since $\log x \le x - 1$ for all $x > 0$,

$$\sum_{j=a+1}^{b}\ln g_j \le \sum_{j=a+1}^{b} g_j - (b-a)$$
$$\le (\overline{\gamma} - 1)(b-a) + Kk_F\epsilon_F^3\sum_{j=a+1}^{b}\|\delta_j\|$$
$$+K_{ee}\sum_{j=a+1}^{b}\frac{e_j^2}{z_{j-1}} + K_l\sum_{j=a+1}^{b}\frac{\|l_j\|^2}{z_{j-1}}.$$

Using (vii) of Lemma 3 and the fact that $z_{j-i} \ge L$, we obtain $z_{j-1} \ge K_\rho\rho_{j-1}$ for all $j \in [a+1, b]$. Now noting that $\sum_{j=a+1}^{b}\|\delta_j\| \le K_\delta + k_\delta(b-a)$, $z_{j-1} \ge K_\rho\rho_{j-1}$, and

$$\frac{\|l_j\|^2}{z_{j-1}} = \frac{\|b_{v,j}v_j + b_{r,j}r'_{j-1}\|^2}{z_{j-1}} \le 2\left[\frac{v_j^2}{b_{\min}^2 z_{j-1}} + \frac{\|b_{r,j}\|^2 r'^2_{j-1}}{z_{j-1}}\right]$$
$$\le 2\frac{K_{vz}}{b_{\min}^2} + \frac{2k_{vz}}{b_{\min}^2 L} + \frac{2K(1+K_\theta^2)}{L},$$

we obtain

$$\sum_{j=a+1}^{b} \ln g_j \le -(1-\overline{\gamma}-Kk_F\epsilon_F^3k_\delta)(b-a)+Kk_F\epsilon_F^3K_\delta+\frac{K_{ee}}{K_\rho}\sum_{j=a+1}^{b}\frac{e_j^2}{\rho_{j-1}}$$
$$+K_l\left[\frac{2K_{vz}}{b_{\min}^2}+\frac{2k_{vz}}{b_{\min}^2L}+\frac{2K(1+K_{\hat{\theta}}^2)}{L}\right](b-a).$$

From (17), using Lemma 4(ii), we obtain

$$\sum_{j=a+1}^{b}\frac{e_j^2}{\rho_{j-1}} \le K_{ev}\sum_{j=a+1}^{b}\frac{v_j^2}{\rho_{j-1}}+k_e(b-a)+K_e$$
$$\le \frac{K_{ev}}{\bar{k}(N)}(b-a)\left[K_{vz}+\frac{k_{vz}}{L}\right]+k_e(b-a)+K_e.$$

Hence,

$$\sum_{j=a+1}^{b}\ln g_j \le -(1-\overline{\gamma}-Kk_F\epsilon_F^3k_\delta)(b-a)+Kk_F\epsilon_F^3K_\delta$$
$$+\frac{K_{ee}}{K_\rho}\frac{K_{ev}}{\bar{k}(N)}\left[K_{vz}+\frac{k_{vz}}{L}\right](b-a)+\frac{K_{ee}}{K_\rho}k_e(b-a)$$
$$+\frac{K_{ee}K_e}{K_\rho}+K_l\left[\frac{2K_{vz}}{b_{\min}^2}+\frac{2k_{vz}}{b_{\min}^2L}+\frac{2K(1+K_{\hat{\theta}}^2)}{L}\right](b-a)$$
$$=-(b-a)\left(1-\overline{\gamma}-Kk_F\epsilon_F^3k_\delta-\frac{K_{ee}}{K_\rho}\frac{K_{ev}}{\bar{k}(N)}\left[K_{vz}+\frac{k_{vz}}{L}\right]\right.$$
$$\left.-\frac{K_{ee}k_e}{K_\rho}-K_l\left[\frac{2K_{vz}}{b_{\min}^2}+\frac{2k_{vz}}{b_{\min}^2L}+\frac{2K(1+K_{\hat{\theta}}^2)}{L}\right]\right)$$
$$+Kk_F\epsilon_F^3K_\delta+\frac{K_{ee}K_e}{K_\rho}$$
$$=-(b-a)\left[1-\overline{\gamma}-Kk_F\epsilon_F^3k_\delta-K_{vz}\left(\frac{K_{ee}}{K_\rho}\frac{K_{ev}}{\bar{k}(N)}+\frac{2K_l}{b_{\min}^2}\right)\right.$$
$$\left.-\frac{1}{L}\left(\frac{K_{ee}}{K_\rho}\frac{K_{ev}}{\bar{k}(N)}k_{vz}+\frac{2k_{vz}K_l}{b_{\min}^2}+\frac{2K(1+K_{\hat{\theta}}^2)K_l}{L}\right)-\frac{K_{ee}k_e}{K_\rho}\right]$$
$$+Kk_F\epsilon_F^3K_\delta+\frac{K_{ee}K_e}{K_\rho}.$$

By choosing K_v, k_δ small enough and L large enough, we obtain

$$\sum_{j=a+1}^{b}\ln g_j \le -(b-a)\lambda+\frac{K_eK_{ee}}{K_\rho}+Kk_F\epsilon_F^3K_\delta \tag{43}$$

for some $0<\lambda<1$.

We thus obtain

$$\frac{W_b}{W_a} \le e^{-(b-a)\lambda}\left[1+\frac{\overline{K}_3}{K_{wz}L(1-\overline{\gamma})\overline{\gamma}^{b-a}}\right]e^{\left(K_eK_{ee}/K_\rho+Kk_F\epsilon_F^3K_\delta\right)}.$$

Now we turn to the case $d > 1$. The essential difference between the cases $d = 1$ and $d > 1$ is that we replace $b_{e,k}e_k$ by $b_1 l_{1,k}$, and replace l_k by $b_{1,k-d}$ in (39). Hence we need to bound $l_{1,k}^2$ and $l_{2,k}^2$ in terms of $e_{(\cdot)}^2$, $v_{(\cdot)}^2$, and $\|\delta_{(\cdot)}\|$. This is done as in the proof of Lemma 3, except that $A^*(B_{k-1} - B_{k-d})u_{k-d}$, which is part of $\Delta_{5,k}u_{k-d}$ has to be handled differently. It gives rise to terms involving $\|\delta_j\|$.

Let us consider how we overbound $l_{1,k}^2$. As in the proof of Lemma 3(ii), we obtain

$$(\Delta_{1,k}u_{k-d})^2 \le K\sum_{j=1}^{d-1} e_{k-j}^2 + K,$$

$$(\Delta_{2,k}u_{k-d})^2 \le K\sum_{j=1}^{d-1} e_{k-j}^2 + K,$$

$$(\Delta_{3,k}y_k)^2 \le K\sum_{j=1}^{d-1} e_{k-j}^2 + K, \quad \text{and}$$

$$(\Delta_{4,k}y_{k-d})^2 \le K\sum_{j=1}^{d-1} e_{k-j}^2 + K.$$

Combining these yields $l_{1,k}^2 \le K\sum_{j=1}^{d-1} e_{k-j}^2 + K$.

Next, consider $l_{2,k}^2$. First, note that we have the following inequality, whose counterpart in the $d = 1$ case is (40):

$$u_{k-d}^2 \le K\|x_{k-1}\|^2 + K. \tag{44}$$

This is obtained by applying the control law (9), using the boundedness of the parameter estimates, and the boundedness from below, by $b_{\min} > 0$, of $\hat{b}_{1,k-d}$.

For the term $A^*(B_{k-1} - B_{k-d})u_{k-d}$, we have

$$\begin{aligned}
(A^*(B_{k-1} - B_{k-d})u_{k-d})^2 &= ((B_{k-1} - B_{k-d})A^*u_{k-d})^2 \\
&= \left(\sum_{j=1}^{d-1}(B_{k-j} - B_{k-j-1})A^*u_{k-d}\right)^2 \\
&= \left(\sum_{j=1}^{d-1}\sum_{i=0}^{\ell-1}\sum_{i'=1}^{p'}(b_{i+1,k-j} - b_{i+1,k-j-1})a_{i'}^*u_{k-d-i-i'}\right)^2 \\
&\le K\left(\sum_{j=1}^{d-1}\|\theta_{k-j} - \theta_{k-j-1}\|\sum_{i=0}^{p'+\ell-1}|u_{k-d-i}|\right)^2 \\
&\le K\sum_{j=1}^{d-1}\|\theta_{k-j} - \theta_{k-j-1}\|^2(u_{k-d}^2 + \cdots + u_{k-d-p'-\ell+1}^2) \\
&\le (K\|x_{k-1}\|^2 + K)\left(\sum_{j=1}^{d-1}\|\theta_{k-j} - \theta_{k-j-1}\|^2\right) \\
&\le (Kz_{k-1} + K)\left(\sum_{j=1}^{d-1}\|\theta_{k-j} - \theta_{k-j-1}\|^2\right),
\end{aligned}$$

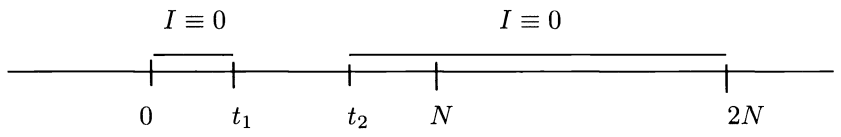


FIG. 1. *Illustration of Case* 2.

where the last two inequalities follow from (44) and the definition of $z(\cdot)$. The remaining terms can be handled similarly. □

LEMMA 7 (the Contraction Lemma). Consider $0 < \gamma^* < 1$. Then there exist N, L large enough, and $K_{v\max}, k_{\delta\max} > 0$ so that if $K_v \in [0, K_{v\max}]$ and $k_\delta \in [0, k_{\delta\max}]$, and

$$W_k \geq 2K_{wz}L \quad \text{for all } k \in [l-1, l+2N],$$

then

$$W_{l+2N} \leq \gamma^* W_{l-1}. \tag{45}$$

Proof. There are four cases to consider.

Case 1. Suppose $I_{t-1} = 0$ for all $t \in [l, l+2N]$. From part (iii) of Lemma 6,

$$W_{l+2N} \leq 2K_{wz}\left(g^{2N} + \frac{2\overline{K}_3}{(1-g)L}\right) W_l.$$

Hence, for N, L large enough, and $K_{v\max}$ and $k_{\delta\max}$ correspondingly small, we have (45).

Case 2. Suppose $0 \leq t_1 \leq t_2 \leq N$, where $t_1 = \min\{t \in [0, 2N] : I_{t+l-1} = 1\}$ and $t_2 = \max\{t \in [t_1, 2N] : I_{t+l-1} = 1\}$. (See Fig. 1.)

Note that this implies that $I_{k-1} = 0$ for all $k \in [l, l+t_1-1]$, and $I_{k-1} = 0$ for all $k \in [l+t_2+1, l+2N]$. First suppose $t_2 > 0$. Using parts (ii), (iii) and (iv) of Lemma 6, we have

$$\begin{aligned}
W_{l+2N} &\leq 2K_{wz}\left(g^{2N-t_2-1} + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l+t_2+1} \\
&\leq 2K_{wz}K_{ww}^2\left(g^{2N-t_2-1} + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l+t_2-1} \\
&\leq 2K_{wz}K_{ww}^2\left(g^{N-1} + \frac{2\overline{K}_3}{(1-g)L}\right) K\exp[-\lambda t_2]\left(1 + \frac{K}{L\overline{\gamma}^{t_2}}\right) W_{l-1} \\
&\leq 2K_{wz}K_{ww}^2 K\left(g^{N-1} + \frac{2\overline{K}_3}{(1-g)L}\right)\left(1 + \frac{K}{L\overline{\gamma}^N}\right) W_{l-1},
\end{aligned}$$

yielding (45) when N, L are large enough, and $K_{v\max}$ and $k_{\delta\max}$ appropriately small. If $t_2 = 0$, then by parts (ii) and (iii) of Lemma 6,

$$\begin{aligned}
W_{l+2N} &\leq 2K_{wz}\left(g^{2N-1} + \frac{2\overline{\overline{K}}_3}{(1-g)L}\right) W_{l+1} \\
&\leq 2K_{wz}K_{ww}^2\left(g^{2N-1} + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l-1},
\end{aligned}$$

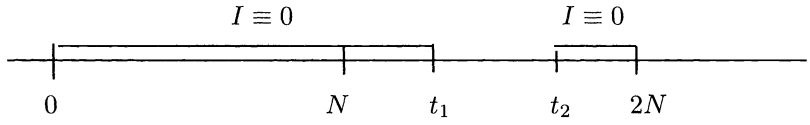


FIG. 2. *Illustration of Case* 3.

hence giving (45).

Case 3. Suppose $N < t_1 \leq t_2 \leq 2N$. First suppose $t_2 < 2N$. (See Fig. 2.)

Using parts (ii), (iii), and (iv) of Lemma 6, we have

$$W_{l+2N} \leq 2K_{wz}\left(g^{2N-t_2-1} + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l+t_2+1},$$

$$W_{l+t_2+1} \leq K_{ww}^2 W_{l+t_2-1},$$

$$W_{l+t_2-1} \leq K \exp\left[-\lambda(t_2 - 1 - N)\right]\left[1 + \frac{K}{L\overline{\gamma}^{t_2-1-N}}\right] W_{l+N},$$

and using parts (ii) and (iii) of Lemma 6 gives

$$W_{l+N} \leq 2K_{wz}K_{ww}\left(g^N + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l-1}.$$

This yields

$$W_{l+2N} \leq 4KK_{wz}^2K_{ww}^3\left(g^N + \frac{2\overline{K}_3}{(1-g)L}\right)\left(1 + \frac{2\overline{K}_3}{(1-g)L}\right)\left(1 + \frac{K}{L\overline{\gamma}^{N-1}}\right) W_{l-1},$$

and hence (45). If $t_2 = 2N$, the first two inequalities are replaced by

$$W_{l+2N} \leq K_{ww}W_{l+2N-1} = K_{ww}W_{l+t_2-1}.$$

This gives

$$W_{l+2N} \leq 2KK_{wz}K_{ww}^2\left(g^N + \frac{2\overline{K}_3}{(1-g)L}\right)\left(1 + \frac{K}{L\overline{\gamma}^{N-1}}\right) W_{l-1},$$

hence giving (45).

Case 4. Suppose $0 \leq t_1 \leq N < t_2 \leq 2N$. Define $t_3 := \min\{t \in [N, t_2] : I_{t+l-1} = 1\}$ and $t_4 := \max\{t \in [t_1, N] : I_{t+l-1} = 1\}$. Note that this implies that $I_{k-1} = 0$ for all $k \in [l + t_4 + 1, l + t_3 - 1]$. (See Fig. 3.)

Case 4a. Suppose $t_3 - t_4 < N$. First suppose $t_2 < 2N$. Using parts (ii), (iii), and (iv) of Lemma 6, we have

$$W_{l+2N} \leq 2K_{wz}\left(g^{2N-t_2-1} + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l+t_2+1},$$

$$W_{l+t_2+1} \leq K_{ww}^2 W_{l+t_2-1},$$

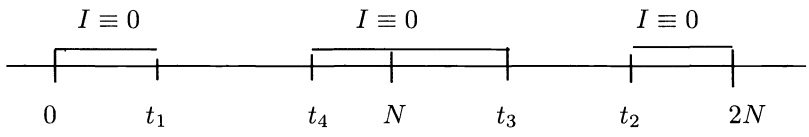


FIG. 3. *Illustration of Case* 4.

and

$$W_{l+t_2-1} \leq K \exp[-\lambda t_2] \left(1 + \frac{K}{L\bar{\gamma}^{t_2}}\right) W_{l-1},$$

which gives

$$W_{l+2N} \leq 2KK_{wz}K_{ww}^2 \left(1 + \frac{2\overline{K}_3}{(1-g)L}\right) \exp(-\lambda N) \left(1 + \frac{K}{L\bar{\gamma}^{2N}}\right) W_{l-1},$$

and hence (45). If $t_2 = 2N$, the first two inequalities are replaced by

$$W_{l+2N} \leq K_{ww} W_{l+2N-1} = K_{ww} W_{l+t_2-1}.$$

This gives

$$W_{l+2N} \leq KK_{ww} \exp(-2\lambda N) \left(1 + \frac{K}{L\bar{\gamma}^{2N}}\right) W_{l-1},$$

and hence (45).

Case 4b. Suppose $t_3 - t_4 \geq N$. Again, first suppose that $t_2 < 2N$. Using parts (ii), (iii), and (iv) of Lemma 6, we have

$$W_{l+2N} \leq 2K_{wz} \left(g^{2N-t_2-1} + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l+t_2+1},$$

$$W_{l+t_2+1} \leq K_{ww}^2 W_{l+t_2-1},$$

$$W_{l+t_2-1} \leq K \exp(-\lambda(t_2 - t_3)) \left(1 + \frac{K}{L\bar{\gamma}^{t_2-t_3}}\right) W_{l+t_3-1},$$

$$W_{l+t_3-1} \leq 2K_{wz} \left(g^{t_3-t_4-2} + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l+t_4+1},$$

$$W_{l+t_4+1} \leq K_{ww}^2 W_{l+t_4-1},$$

and

$$W_{l+t_4-1} \leq K \exp(-\lambda t_4) \left(1 + \frac{K}{L\bar{\gamma}^{t_4}}\right) W_{l-1},$$

which gives

$$W_{l+2N} \leq 4K^2 K_{wz}^2 K_{ww}^4 \left(g^{N-2} + \frac{2\overline{K}_3}{(1-g)L}\right) \left(1 + \frac{K}{L\bar{\gamma}^N}\right)^2 \left(1 + \frac{2\overline{K}_3}{(1-g)L}\right) W_{l-1},$$

and hence (45). If $t_2 = 2N$, the first two inequalities are replaced by

$$W_{l+2N} \leq K_{ww} W_{l+2N-1} = K_{ww} W_{l+t_2-1}.$$

This gives

$$W_{l+2N} \leq 2K^2 K_{wz} K_{ww}^3 \left(g^{N-2} + \frac{2\overline{K}_3}{(1-g)L} \right) \left(1 + \frac{K}{L\bar{\gamma}^N} \right)^2 W_{l-1},$$

thus establishing (45) in all cases. □

THEOREM 1 (Robust Boundedness Theorem). *Consider the adaptive control system, when the plant satisfies Assumptions 1–5. Then all signals in the closed-loop adaptive system are bounded, whenever K_v and k_δ are small enough.*

Proof. As a consequence of the Bounded Growth Rate Lemma (BGRL) and the Contraction Lemma (CL), W is bounded. To see this, note that by BGRL, W cannot have a finite escape time. Hence, it can only become infinite by growing over an unbounded length of time. However, CL disallows this by establishing that if W stays above a certain value ($2K_{wz}L$) for a certain time interval ($2N+1$ samples), it must contract. That is, W cannot continue to grow for more than $2N+1$ consecutive samples at a time, once it has grown larger that $2K_{wz}L$. Hence, W is bounded.

Since W bounds all other signals through z and x, we conclude that all closed-loop signals are bounded. □

Remark. These results extend easily to recursive least-squares-based schemes that keep the condition number of the covariance matrix bounded.

8. Performance for a nominal system. In this section we consider the performance that can be achieved in the absence of unmodeled dynamics, and when the parameter variations go to zero asymptotically, specifically when $\|\delta_k\| \in \ell_1$. We show that the desired objective is met.

THEOREM 2. *If $K_v = k_v = k_\delta = 0$, then*

(a) $\lim_{k\to\infty} e_k = 0$,

(b) $\lim_{k\to\infty} (A^*(q^{-1})y_k - q^{-d}B^*(q^{-1})r_k) = 0$.

Remark. Note that we *do not* need the nominal plant parameters to be time-invariant; we only require them to satisfy $\|\delta_k\| \in \ell_1$. That is,

$$\sum_{k=t+1}^{\infty} \|\theta_k - \theta_{k-1}\| \leq K_\delta, \qquad \forall t.$$

Proof. (a) Recall from Lemma 1(iii) that

$$\sum_{k=t+1}^{t+T} \frac{e_k^2}{\rho_{k-1}} \leq K_{ev} \sum_{k=t+1}^{t+T} \frac{v_k^2}{\rho_{k-1}} + k_e T + K_e.$$

Since $K_v = K_v = k_\delta = 0$, from Assumption 4 and Lemma 1(iii) we have $v_k \equiv 0$ and $k_e = 0$, which gives

$$\sum_{k=t+1}^{t+T} \frac{e_k^2}{\rho_{k-1}} \leq K_e \qquad \forall t, T.$$

Recalling that $\rho_{(\cdot)}$ is uniformly bounded by Theorem 1 and letting $\rho_{\max} := \max_{k\geq 0}\{\rho_k\}$, we obtain

$$\sum_{k=t+1}^{t+T} e_k^2 \leq K_e \rho_{\max} \qquad \forall t, T.$$

Fixing t and letting T go to infinity gives $e \in \ell_2$ and hence $e_k \to 0$ as $k \to \infty$.

(b) Recalling (10), the fact that $\|\widehat{\theta}_k - \widehat{\theta}_{k-1}\| \leq \|\widehat{\theta}'_k - \widehat{\theta}_{k-1}\|$ due to parameter projection and using the Cauchy–Schwarz inequality, we obtain

$$\|\widehat{\theta}_k - \widehat{\theta}_{k-j}\| \in \ell_2, \quad \text{for all finite } j.$$

Using this in (27), the closed-loop boundedness of all signals, and using (a) yields the desired result. □

9. Robust performance. We now show that the performance of the adaptive controller, as measured by the mean square output prediction error, is robust in that it is linear (hence also continuous) in the magnitude of the unmodeled dynamics, bounded disturbances, and average rate of parameter variations.

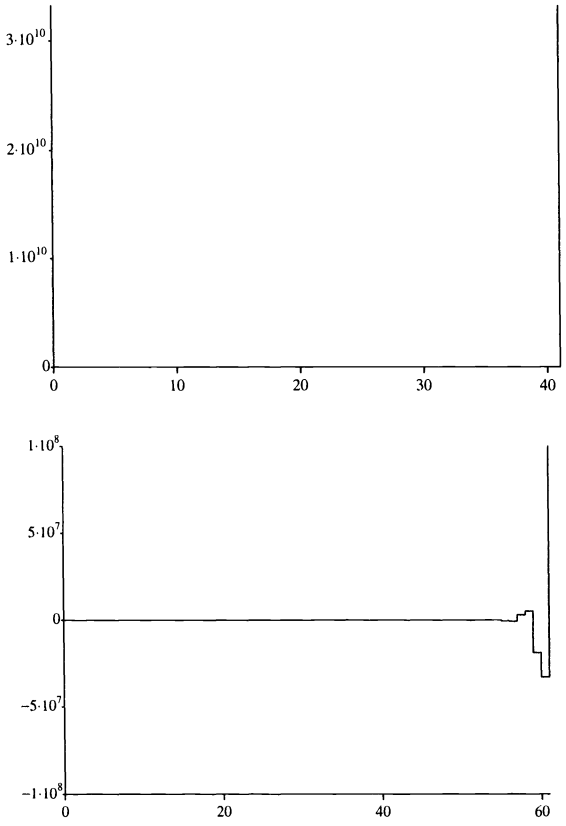


FIG. 4. (a) *Unmodified (LMS-type) gradient estimator: y blows up*; (b) *gradient estimator with parameter projection: y blows up.*

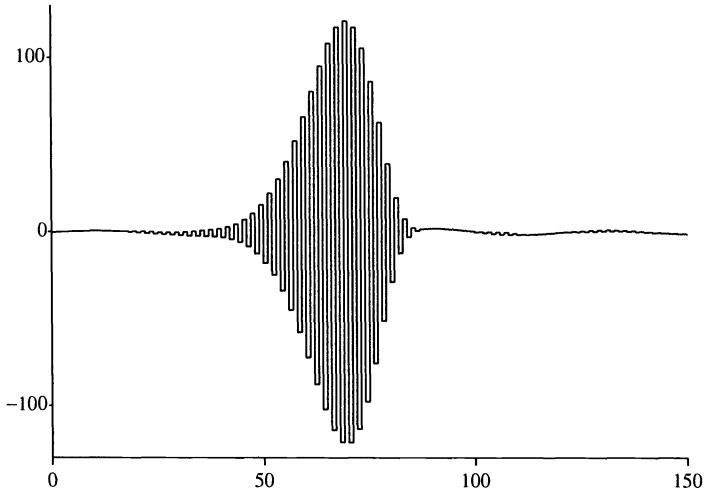


FIG. 5. *Normalized gradient estimator: y is not well behaved.*

THEOREM 3. *We have*

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{k=t+1}^{t+T} e_k^2 \leq c_1 K_v + c_2 k_v + c_3 k_\delta,$$

where c_1, c_2, c_3 are generic constants that can only decrease (or remain constant) as K_v, k_v, k_δ decrease.

Proof. Using Assumption 4 and Lemma 1(iv), and recalling that $\{m_k\}$ and $\{\rho_k\}$ are uniformly bounded by Theorem 1, we obtain

$$\begin{aligned}\sum_{k=t+1}^{t+T} \frac{e_k^2}{\rho_{k-1}} &\leq K_{ev} \sum_{k=t+1}^{t+T} \frac{K_v m_{k-1} + k_v}{\rho_{k-1}} + k_e T + K_e, \qquad \forall t, T \\ &\leq K_{ev}(KK_v + Kk_v)T + k_e T + K_e, \qquad \forall T.\end{aligned}$$

This implies

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{k=t+1}^{t+T} e_k^2 \leq KK_{ev}(KK_v + Kk_v) + Kk_e,$$

which yields the desired result after using Lemma 1(iv). □

10. Simulation example. We now demonstrate the advantage of the suggested adaptive control algorithm through a simulation example. The system is modelled as

$$y_k = a_k y_{k-1} + b_k u_{k-1},$$

where a_k and b_k are unknown, possibly time-varying parameters. The true unknown system is, however, given by the following:

$$y_k = 1.5 \sin\left(\frac{\pi k}{1000}\right) y_{k-1} + \left[1 + 0.4 \cos\left(\frac{\pi k}{1500}\right)\right] u_{k-1} + 0.2 \sum_{j=0}^{k-2} (0.5)^j y_{k-2-j} + d_k,$$

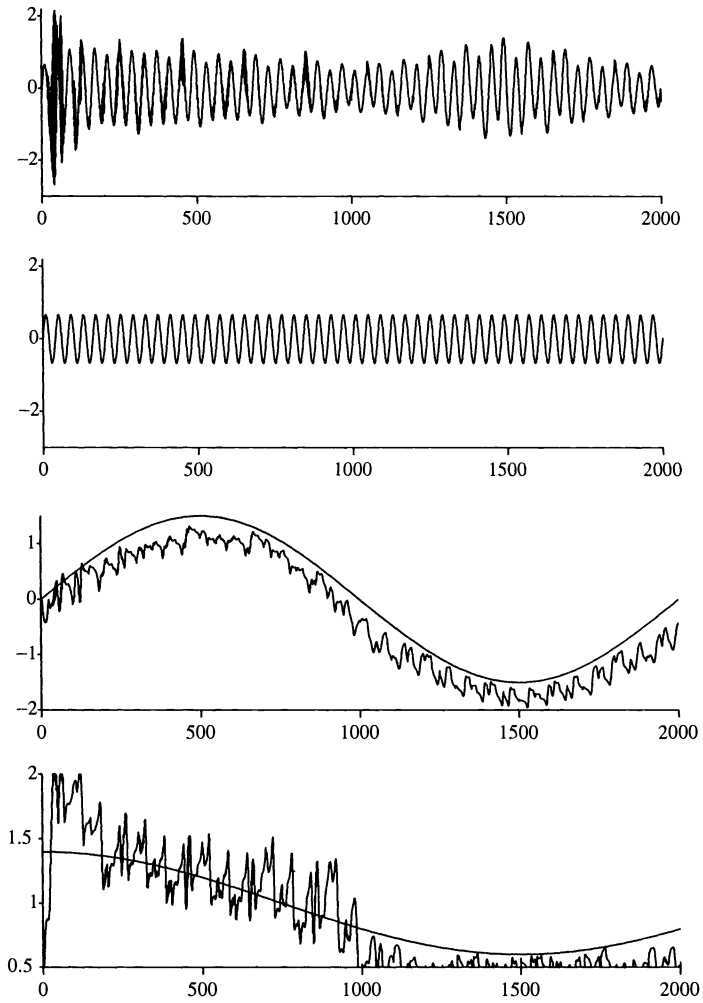


FIG. 6. *Proposed adaptive control algorithm*: (a) *system output*, y; (b) *model reference output*, y^*; (c) *true and estimated* a_k; (d) *true and estimated* b_k.

where d_k denotes a discrete square wave disturbance of period 100 and amplitude 0.15. The adaptive control is designed to track the following model reference trajectory:

$$y_k^* = -0.5y_{k-1}^* + r_k.$$

As Fig. 4 shows, using an unmodified LMS-type gradient estimator or an LMS-type gradient estimator with parameter projection causes the output to blow up. Figure 5 illustrates the undesirable behavior that results if we use a normalized gradient estimator, as used in the ideal case. Finally, Fig. 6 illustrates the results obtained if we use the adaptive control algorithm with parameter projection, as proposed in this paper. Figures 6(c) and 6(d) also exhibit the nice parameter tracking that is achieved.

11. Concluding remarks. We have presented an indirect adaptive pole-zero placement control law using a simple parameter estimator employing projection. We have shown that this is robust for plants that simultaneously feature unknown slow-

in-the-mean time-variations of the nominal parameters, as well as small unmodeled dynamics and bounded disturbances, without any restriction on the magnitude of the bound. The plant parameters may even make occasional jumps. No special normalization is used. Instead, the signals entering the parameter update law are normalized by the squared norm of an "extended" regressor, which requires neither any a priori system knowledge nor any additional computation.

It is straightforward to extend this analysis to recursive least-squares-based update laws that monitor, and keep bounded, the condition number of the covariance matrix.

Several issues still need to be explored. A major restriction is that we require the frozen nominal plant to be minimum phase at every instant. Transient performance, and the precise sizes of unmodeled dynamics and parameter variations tolerated, are issues that require deeper study.

Acknowledgments. This paper was completed while one of the authors was visiting the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. The authors are grateful to Professor N. Viswanadham for the warm hospitality and facilities provided.

REFERENCES

[1] K. Åström and B. Wittenmark, *Self-tuning controllers based on pole-zero placement*, IEEE Proceedings, 127 (1980), pp. 120–130.

[2] E. G. Vogel and T. F. Edgar, *Application of an adaptive pole-zero placement controller to chemical processes with variable dead-time*, in Proc. Amer. Control Conf., Washington D.C., June 1982.

[3] S. Ananthakrishnan and R. Fullner, *Application of a class of adaptive control algorithms to hydraulic servosystems*, in Proc. Amer. Control Conf., San Diego, CA, May 1990, pp. 1086–1087.

[4] M. Sunwoo and K. C. Cheok, *An application of explicit self tuning control to vehicle active suspension systems*, in Proc. IEEE 29th Conf. on Decision and Control, Honolulu, HI, December 1990, pp. 2251–2257.

[5] P. A. Ioannou and J. Sun, *Theory and design of robust direct and indirect adaptive control schemes*, Internat. J. Control, 47 (1988), pp. 775–813.

[6] B. E. Ydstie, *Stability of discrete* MRAC*-revisited*, Systems Control Lett., 13 (1989), pp. 429–439.

[7] S. M. Naik, P. R. Kumar, and B. E. Ydstie, *Robust continuous time adaptive control by parameter projection*, IEEE Trans. Automat. Control, AC-37 (1992), pp. 182–197.

[8] B. Egardt, *Stability of Adaptive Controllers*, Springer, Berlin, 1979.

[9] V. Solo, *A one step ahead adaptive controller with slowly time varying parameters*, 1991, submitted.

[10] G. Kreisselmeier, *Adaptive control of a class of slowly time-varying plants*, Systems and Control Letters, 8 (1986), pp. 97–103.

[11] K. S. Tsakalis and P. A. Ioannou, *Adaptive control of linear time varying plants: A new model reference controller structure*, IEEE Trans. Automat. Control, 34 (1989), pp. 1038–1046.

[12] L. Guo, *On adaptive stabilization of time-varying stochastic systems*, SIAM J. Control Optim., 28 (1990), pp. 1432–1451.

[13] S. P. Meyn and L. Guo, *Adaptive control of time varying stochastic systems*, in Proc. 11th IFAC World Congress, V. Utkin and O. Jaaksoo, eds., Vol. 3, Tallinn, Estonia, USSR, August 1990, pp. 198–202.

[14] P. de Larminat and H. Raynaud, *A robust solution to the admissibility problem in indirect adaptive control without persistency of excitation*, Internat. J. Adaptive Control Signal Proc., 2 (1988), pp. 95–110.

[15] R. H. Middleton and G. C. Goodwin, *Adaptive control of time-varying linear systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 150–155.

[16] F. GIRI, M. M'SAAD, L. DUGARD, AND J. M. DION, *Robust adaptive regulation with minimal prior knowledge*, IEEE Trans. Automat. Control, 37 (1992), pp. 305–315.

[17] L. PRALY, *Robustness of indirect adaptive control based on pole placement design*, IFAC Workshop on Adaptive Control, June 1983.

[18] F. GIRI, M. M'SAAD, J. M. DION, AND L. DUGARD, *A cautious approach to robust adaptive regulation*, Internat. J. Adaptive Control Signal Processing, 2 (1988), pp. 273–290.

[19] L. PRALY, S. LIN, AND P. R. KUMAR, *A robust adaptive minimum variance controller*, SIAM J. Control Optim., 27 (1989), pp. 235–266.

[20] L. PRALY, *Almost exact modelling assumption in adaptive linear control*, Internat. J. Control, 51 (1990), pp. 643–668.

[21] C. WEN, *A robust adaptive controller with minimal modifications for discrete time varying systems*, in Proc. IEEE 31st Conf. on Decision and Control, Tuscon, AZ, December 1992, pp. 2132–2136.

[22] C. WEN AND D. J. HILL, *Global boundedness of discrete-time adaptive control just using estimator projection*, Automatic—J. IFAC, 28 (1992), pp. 1143–1158.

[23] J. VAN AMERONGEN, *Adaptive steering of ships—a model reference approach to improved maneuvering and economic course keeping*, Ph. D. thesis, Huisdrukkerij, Delft Univ. of Technology, Delft, The Netherlands, 1982.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1726–1751, November 1994

012

INFORMATION STRUCTURES, CAUSALITY, AND NONSEQUENTIAL STOCHASTIC CONTROL II: DESIGN-DEPENDENT PROPERTIES*

MARK S. ANDERSLAND† AND DEMOSTHENIS TENEKETZIS‡

Abstract. In control theory, the usual notion of causality—that, at all times, a system's output (action) only depends on its past and present inputs (observations)—presupposes that all inputs and outputs can be ordered, a priori, in time. In reality, many distributed systems (those subject to deadlock, for instance), are not *sequential* in this sense.

In a previous paper (part I) [*SIAM J. Control Optim.*, 30 (1992), pp. 1447–1475], the relationship between a less restrictive notion of causality, *deadlock-freeness*, and the design-independent properties of a potentially *nonsequential* generic stochastic control problem formulated within the framework of Witsenhausen's intrinsic model was explored. In the present paper (part II) the properties of individual designs are examined. In particular, a property of a design's *information partition* that is necessary and sufficient to ensure its deadlock-freeness is identified and shown to be sufficient to ensure its possession of an expected reward. It is also shown, by example, that there exist nontrivial deadlock-free designs that cannot be associated with any deadlock-free information structure.

The first result provides an intuitive design-dependent characterization of the cause/effect notion of causality and suggests a framework for the optimization of constrained nonsequential stochastic control problems. The second implies that this characterization is finer than existing design-independent characterizations, including properties C (Witsenhausen) and CI (part I).

Key words. information structures, causality, deadlock-freeness, nonsequential stochastic control

AMS subject classifications. 93E03, 93E05, 90D15, 93A15, 93A14, 93E20

1. Introduction.

In control theory the usual notion of causality—that, at all times a system's output (action) only depends on its past and present inputs (observations)—presupposes that all inputs and outputs can be ordered a priori in time. In reality, many controlled systems—including distributed data [5], communication [6], manufacturing [3], and detection networks [2]—need not be *sequential* [10] in this sense.

Consider, for example, a simple detection network in which three decentralized detectors D^1, D^2, and D^3 (perhaps radars or inspectors) each make a noisy observation of the same uncertain event (plane or product). Suppose that each detector forms and transmits a one-bit hypothesis concerning the event (e.g., friend/foe or pass/fail) to a silent coordinator. Moreover, suppose that each detector may elect to monitor the others' transmissions before forming its hypothesis. Then, depending on the detectors' control laws (termed the *design*) and the particular event that occurs, 64 different dependencies are possible, 39 of which deadlock.[1] For instance, D^3 may wait for D^1, and depending on D^1's transmission, perhaps D^2, but D^3 and D^1 may not wait for each other because then neither can act.

* Received by the editors September 9, 1991; accepted for publication (in revised form) May 12, 1993. This research was supported in part by National Science Foundation grant ECS-8517708, Office of Naval Research grant N00014-87-K-0540, Defense Advanced Research Projects Agency grant N00174-91-C-0116, National Science Foundation grant NCR-9204419, and a Hewlett Packard Faculty Development Award.

† Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, Iowa 52242-1595.

‡ Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109-2122.

[1] Each of the three detectors may wait for: none, one, the other, or both detectors; hence there are 4^3 possibilities. By case analysis, 39 of these deadlock.

This example illustrates two key differences between sequential and *nonsequential* systems, namely: i) that the order in which a nonsequential system's actions occur may explicitly depend on the system's uncontrolled inputs and the actions taken, and ii) that when two or more of a nonsequential system's actions are interdependent, no "causal" ordering of the actions is possible. Due to i), deadlock-free designs that exploit a system's nonsequentiality can outperform those that do not (see [2], Appendix A). This should not be surprising; unlike sequential systems, the dependencies among a nonsequential system's actions can change dynamically. Due to ii), the problem of identifying these "good" designs is difficult to formulate as a well-defined stochastic control problem. In particular, a design that deadlocks need not possess an expected reward,[2] and when it does, it may be mathematically optimal despite the fact that it is "not causal." This raises the question: Under what conditions is it possible to pose well-defined nonsequential stochastic control problems?

In a previous paper [2] (part I), we addressed this question by defining a nonsequential system to be "causal" when, independent of its design, it is deadlock-free. We then identified a property of a potentially nonsequential generic stochastic control problem's *information structure* (property CI) that is necessary and sufficient to ensure deadlock-freeness, and sufficient to ensure that *all* of the problem's designs possess expected rewards. This result subsumes Witsenhausen's design-independent causality condition (property C, in [9], [11]) and provides a framework for the recursive optimization of *unconstrained* nonsequential stochastic control problems [1].

In the present paper (part II) we explore the relationship between deadlock-freeness and the properties of individual designs. Our work is motivated by the fact that when the observations available to a nonsequential system's decision-making agents (e.g., the detectors) are specified independently, the resulting information structure need not be causal in the C or CI sense, although many admissible designs may be deadlock-free. This presents systems designers with a dilemma. If the existence of noncausal designs is ignored, formal optimization may not be possible. On the other hand, if the agents' information is constrained to ensure design-independent causality—by forcing sequentiality, for instance—the designer may limit the system's possible performance.

An obvious alternative to either "fix" is to identify necessary and sufficient conditions for individual designs to be causal. Once again, Witsenhausen's intrinsic model [9], [11] provides the framework for our work. Within this framework, we identify design-dependent analogues of the causality properties C and CI. Specifically, we introduce properties of a design's *information partition* (properties C* and CI*) that are necessary and sufficient to ensure that the design is deadlock-free, and sufficient to ensure that it possesses an expected reward. Moreover, we show by example that there exist deadlock-free designs that cannot be associated with any deadlock-free information structure.

The first result provides an intuitive, design-dependent characterization of the cause/effect notion of causality, and suggests a framework for the optimization of *constrained* nonsequential stochastic control problems. The second implies that for $N > 2$ agents, this characterization is finer than existing design-independent characterizations, including properties C and CI. Because our conditions are based on what a nonsequential system's decision-making agents may know as opposed to what they may do, they are substantially different than those derived using event sequence-based

[2] To compute the reward we must break the deadlock, but the reward may vary depending on how this is done (see [2, §2.3]).

representations such as finite-state automata [7], or Petri nets [8].

The remainder of the paper is organized as follows. In §2 we briefly review the structure of Witsenhausen's intrinsic model and our generic stochastic control problem. In §3 we introduce the design-dependent analogues of the deadlock-freeness, well-posedness, and causality properties in [2] and [9], [11], and relate a design's possession of these properties to its deadlock-freeness and possession of an expected reward. In §4 we examine the relationship between the design-independent and dependent properties, and establish, by example, that the design-dependent properties are finer. Section 5 contains our conclusions.

2. Problem formulation. The generic stochastic control problem considered in this paper is identical to that in [2] (part I). As before, the problem is posed within the framework of Witsenhausen's intrinsic model [9], [11]. This model, which is interpreted in [2], has three components.

1. An *information structure* $\mathcal{I} := \{(\Omega, \mathcal{B}), (U^k, \mathcal{U}^k), \mathcal{J}^k : 1 \leq k \leq N\}$ specifies the system's allowable decisions and distinguishable events.

(a) $N \in \mathbb{N}$ denotes the number of control actions to be taken.

(b) $(\Omega, \mathcal{B})$ denotes the measurable space from which a random input ω is drawn.

(c) $(U^k, \mathcal{U}^k)$ denotes the measurable space from which u^k, the kth control action, is selected. $\mathrm{Card}(U^k)$ is assumed to be greater than one, and $\mathcal{U}^k$ is assumed to contain the singletons of U^k. The measurable product space containing the N-tuple of control actions, $u := (u^1, u^2, \ldots, u^N)$, is denoted by $(U, \mathcal{U}) := (\prod_{i=1}^N U^i, \bigotimes_{i=1}^N \mathcal{U}^i)$.

(d) $\mathcal{J}^k \subset \mathcal{B} \otimes \mathcal{U}$ characterizes the maximal information that can be used to select the kth control action.

2. A design constraint set Γ_C constrains N-tuples of control laws $\gamma := (\gamma^1, \gamma^2, \ldots, \gamma^N)$, $\gamma^k : (\Omega \times U, \mathcal{J}^k) \to (U^k, \mathcal{U}^k)$, $k = 1, 2, \ldots, N$, called *designs*, to a nonempty subset of $\Gamma := \prod_{i=1}^N \Gamma^i$, where Γ^k, $k = 1, 2, \ldots, N$, denotes the set of all $\mathcal{J}^k/\mathcal{U}^k$-measurable functions.

3. A probability measure P on $(\Omega, \mathcal{B})$ determines the statistics of the random input.

When posed within this framework the generic problem takes the following form [2].

(P). Given an information structure $\mathcal{I}$, a design constraint set Γ_C, a probability measure P, and a bounded, nonnegative $\mathcal{B} \otimes \mathcal{U}$-measurable reward function V, identify a design γ in Γ_C that achieves $\sup_{\gamma \in \Gamma_C} E_\omega[V(w, u_\omega^\gamma)]$ exactly, or within $\epsilon > 0$.[3]

3. Design-dependent properties. Problem (P) is well defined when it is: i) *causal*, i.e., every $\gamma \in \Gamma_C$ is deadlock-free; and ii) *well posed*, i.e., every $\gamma \in \Gamma_C$ possesses an expected reward. As in part I, our objective is to identify properties necessary and sufficient to ensure that (P) is causal and well-posed. Here, however, we permit the problem's design constraint set $\Gamma_C \subset \Gamma$ to be arbitrary, and focus on developing *design-dependent* properties (properties that may only hold for *specific* $\gamma \in \Gamma$), as opposed to *design-independent* properties (properties that hold for *all* $\gamma \in \Gamma$).

3.1. Deadlock-freeness, solvability, and solvability-measurability: properties DF*, S*, and SM*. The identification of the design-dependent analogues of the deadlock-freeness property DF [2], and the well-posedness properties S (solvability [9]) and SM (solvability-measurability [9]), is straightforward. To ensure the

[3] The notation u_ω^γ indicates that u depends on ω through γ (see Definitions 2 and 3).

deadlock-freeness of the control problem, it is necessary and sufficient to require that each $\gamma \in \Gamma_C$ possess property DF* (cf. [2, Def. 1]).

DEFINITION 1. *A design γ possesses* property DF* (deadlock-freeness) *when for every $\omega \in \Omega$ there exists an ordering of γ's N control laws, say $\gamma^{s_1(\omega)}$, $\gamma^{s_2(\omega)}$, ..., $\gamma^{s_N(\omega)}$, such that no control action depends on itself or the control actions that follow; i.e., $u^{s_i(\omega)}$ does not depend on $u^{s_j(\omega)}$ for $j \geq i$.*

When a design γ possesses property DF*, it is deadlock-free in the sense that, given ω, $u^{s_1(\omega)}$ can be determined; given ω and $u^{s_1(\omega)}$, $u^{s_2(\omega)}$ can be determined; and so on.

To ensure well-posedness, it suffices to require that each $\gamma \in \Gamma_C$ possess properties S* and SM* (cf. [9, §4]).

DEFINITION 2. *A design γ possesses* property S* (solvability) *when for every $\omega \in \Omega$ there exists a unique $u := (u^1, u^2, \ldots, u^N) \in U$ satisfying the system of equations $u^k = \gamma^k(\omega, u)$, $k = 1, 2, \ldots, N$.*

DEFINITION 3. *A design γ possesses* property SM* (solvability-measurability) *when γ possesses property S*, and the solution map $\Sigma^\gamma : \Omega \to U$ induced by the system of equations $u = \gamma(\omega, u)$ (i.e., $\Sigma^\gamma(\omega) = u^\gamma_\omega$, where $u^\gamma_\omega = \gamma(\omega, u^\gamma_\omega)$) is $\mathcal{B}/\mathcal{U}$-measurable.*

Properties S* and SM* ensure that γ's reward $V(\cdot, \Sigma^\gamma(\cdot))$ is $\mathcal{B}$-measurable, and consequently, that $E_\omega[V(w, \Sigma^\gamma(w))]$ is well defined.

3.2. Design-dependent causality: property C*. When all designs $\gamma \in \Gamma_C$ possess property SM*, (P) is well posed. However, just as property SM need not imply property C ([9, Thm. 2]), a design's possession of property SM* need not ensure that it is deadlock-free.

Example 1.[4] Suppose, for instance, that

$$\gamma^1(\omega, u^1, u^2, u^3) = \begin{cases} 1 & \omega \bar{u}^2 u^3 = 1, \\ 0 & \text{else}, \end{cases}$$

$$\gamma^2(\omega, u^1, u^2, u^3) = \begin{cases} 0 & \omega \bar{u}^3 u^1 = 1, \\ 1 & \text{else}, \end{cases} \tag{3.1}$$

and

$$\gamma^3(\omega, u^1, u^2, u^3) = \begin{cases} 1 & \omega \bar{u}^1 u^2 = 1, \\ 0 & \text{else},^5 \end{cases}$$

are the component control laws of an admissible design $\gamma := (\gamma^1, \gamma^2, \gamma^3)$ for a three-agent problem in which

$$\Omega = U^1 = U^2 = U^3 = \{0, 1\}, \tag{3.2}$$

and

$$\mathcal{B} = \mathcal{U}^1 = \mathcal{U}^2 = \mathcal{U}^3 = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}. \tag{3.3}$$

Because

$$\begin{aligned} \mathrm{G}^\gamma &:= \{(\omega, u) : \ \gamma(\omega, u) = u\} \\ &= \{(0, 0, 1, 0), (1, 0, 1, 1)\}, \end{aligned} \tag{3.4}$$

[4] This example is a variation of the example used to prove Theorem 2 of [9].

[5] $\bar{u}$ denotes the binary complement of $u \in \{0, 1\}$, i.e., $\bar{u} = 1 - u$.

γ possesses property SM*. Nonetheless, when $\omega = 1$, γ^1 depends on u^2 and u^3, γ^2 depends on u^3 and u^1, and γ^3 depends on u^1 and u^2. Accordingly, no agent can act without precognition.

Clearly, Witsenhausen's design-independent causality property C [9], [11] provides a condition sufficient to ensure that individual designs $\gamma \in \Gamma$ do not experience such deadlocks. This condition is not necessary, however, because it imposes constraints on all events that the agents can distinguish (i.e., the sets in the information fields $\mathcal{J}^k$, $k = 1, 2, \ldots, N$), not just those distinguishable given a particular design γ (i.e., those in the restriction of the *information partitions* $\mathcal{J}^{\gamma^k} := \{[\gamma^k]^{-1}(u^k) : u^k \in U^k\}$, $k = 1, 2, \ldots, N$, to the *graph* $\mathrm{G}^\gamma := \{(\omega, u) : \gamma(\omega, u) = u\}$ of γ).

These observations suggest that for fixed $\gamma \in \Gamma$, a design-dependent analogue to property C might be constructed by substituting $\mathcal{J}^{\gamma^k}$ for $\mathcal{J}^k$ and G^γ for $\Omega \times U$ in C (cf. [9, §5] or [11, §2]).

DEFINITION 4. *A design $\gamma \in \Gamma$ possesses* property c* *when $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$ and there exists at least one map $\psi : \mathrm{G}^\gamma \to S_N$ such that for all $s := (s_1, s_2, \ldots, s_k) \in S_k$ and $k = 1, 2, \ldots, N$,*

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [T_k^N \circ \psi]^{-1}(s) \subset \mathcal{F}(T_{k-1}^k(s)) \bigcap \mathrm{G}^\gamma. \tag{3.5}$$

Here, as in [2], S_k, $k = 1, 2, \ldots, N$, denotes the set of all k-action orderings (i.e., all injections of $\{1, 2, \ldots, k\}$ into $\{1, 2, \ldots, N\}$); $T_j^k : S_k \to S_j$, $j = 0, 1, \ldots, N$, $k = j, j+1, \ldots, N$, denotes a truncation map that returns the ordering of the first j agents of a k-action ordering (i.e., T_j^k restricts $s \in S_k$ to the domain $\{1, 2, \ldots, j\}$ or to $\emptyset$ when $j = 0$); $\mathcal{P}_s$, $s := (s_1, s_2, \ldots, s_k) \in S_k$, $k = 1, 2, \ldots, N$, denotes the projection of $\Omega \times (\prod_{i=1}^N U^i)$ onto $\Omega \times (\prod_{i=1}^k U^{s_i})$ (i.e.,

$$\mathcal{P}_s(\omega, u) := (\omega, u^{s_1}, u^{s_2}, \ldots, u^{s_k}), \tag{3.6}$$

when $s \neq \emptyset$ and (ω) when $s = \emptyset$); and

$$\mathcal{F}(s) := [\mathcal{P}_s]^{-1} \left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^k \mathcal{U}^{s_i} \right) \right), \tag{3.7}$$

$s := (s_1, s_2, \ldots, s_k) \in S_k$, $k = 1, 2, \ldots, N$, denotes the cylindrical extension of $\mathcal{B} \otimes (\bigotimes_{i=1}^k \mathcal{U}^{s_i})$ to $\Omega \times U$.

To interpret (3.5) note that

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [T_k^N \circ \psi]^{-1}(s) := \left\{ A \bigcap [T_k^N \circ \psi]^{-1}(s) : A \in \mathcal{J}^{\gamma^{s_k}} \right\} \tag{3.8}$$

is the restriction of the set of events distinguishable by agent s_k under γ to the subset of outcomes $(\omega, u) \in \mathrm{G}^\gamma$ that are mapped by ψ into action orders in which the order of the first k agents is $s \in S_k$. Similarly, $\mathcal{F}(T_{k-1}^k(s)) \bigcap \mathrm{G}^\gamma$ is the restriction, to G^γ, of the set of events that can be induced by ω and the actions of the first $k-1$ agents in s. Accordingly, (3.5) asserts that the set of events that agent s_k can distinguish under γ, for *known* G^γ, given that the ordering of the first k agents as determined by ψ is s, must be a subset of the events that can be induced on G^γ by ω and the actions of the first $k-1$ agents in s.

Consider, for instance, the design γ in Example 1. Because for all $k = 1, 2, 3$, and $s \in S_k$, $\mathcal{F}(T_{k-1}^k(s)) \bigcap \mathrm{G}^\gamma$ is the power set of G^γ, all events that can be distinguished by s_k under any $\psi : \mathrm{G}^\gamma \to S_3$ can be induced by $\omega, \ldots, u^{s_{k-1}}$. Hence γ satisfies property c*.

Although a design's possession of property c* implies that it possesses an expected reward (just as $\mathcal{I}$'s possession of property C implies that all designs $\gamma \in \Gamma$ possess expected rewards [9]), property c* does not imply deadlock-freeness.

LEMMA 1. *For fixed $\gamma \in \Gamma$, property c* implies property* SM*, *although property c* need not imply property* DF*.

Proof. See Appendix A. □

The proof that c* implies SM* parallels the proof that C implies SM in [9, Thm. 1]. Property c*'s failure to ensure deadlock-freeness can be explained as follows. Property C is too restrictive to characterize the deadlock-freeness of individual designs because it requires that there exist a causal ordering for all outcomes in $\Omega \times U$, not just those that can occur (i.e., the outcomes in G^γ). Property c* is not restrictive enough because, for fixed $s \in S_k$, it implicitly permits the s_kth agent to possess information about its own action and the actions of its successors in s—i.e., because the domain of ψ is G^γ, $\mathcal{J}^{\gamma^{s_k}} \bigcap [T_k^N \circ \psi]^{-1}(s)$ unavoidably constrains $\mathcal{J}^{\gamma^{s_k}}$ along axes corresponding to agents that are not among the first $k-1$ agents in s. For instance, as previously noted, the design in Example 1 trivially satisfies property c* although it is not deadlock-free.

One compromise between these extremes is to continue to restrict the domain of ψ to G^γ. However, another is to only require, for all $s \in S_k$ and $k = 1, 2, \ldots, N$, that the inclusion in (3.5) hold when $\mathcal{J}^{\gamma^{s_k}}$ and $\mathcal{F}(T_{k-1}^k(s))$ are restricted to, respectively,

$$[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \tag{3.9}$$

and

$$[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(G^\gamma)), \tag{3.10}$$

the smallest subsets of $\Omega \times U$ containing $[T_k^N \circ \psi]^{-1}(s)$ and G^γ that can be constructed without knowledge of the decisions of agents that are not among the first $k-1$ agents in s.

DEFINITION 5. *A design $\gamma \in \Gamma$ possesses* property C* (causality) *when $\mathcal{P}_\emptyset(G^\gamma) = \Omega$, and there exists at least one map $\psi : G^\gamma \to S_N$ such that for all $s := (s_1, s_2, \ldots, s_k) \in S_k$ and $k = 1, 2, \ldots, N$,*

$$\begin{aligned} &\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \\ &\qquad \subset \mathcal{F}(T_{k-1}^k(s)) \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(G^\gamma)). \end{aligned} \tag{3.11}$$

Because the restriction of $\mathcal{J}^{\gamma^{s_k}}$ to $[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s)))$ in (3.11) does not provide information to agent s_k concerning its action or the actions of its successors, in addition to ensuring that a design possesses an expected reward, property C* also implies deadlock-freeness.

THEOREM 1. *If a design $\gamma \in \Gamma$ possesses property* C*, *then*

(i) *γ possesses property* SM*, *and*

(ii) *γ possesses property* DF*.

Proof. See Appendix B. □

The proof of (i) follows from Lemma 1 and the fact that property C* implies property c*. Part (ii) is an immediate consequence of C*'s definition.

3.3. Design-dependent causality: property CI*. By Theorem 1, when all $\gamma \in \Gamma_C$ possess property C*, problem (P) is causal and well posed. It is not clear, however, that the converse implication holds. In particular, it would seem that the measurability constraints that property C* imposes on ψ are unnecessary to ensure deadlock-freeness. Regardless of ψ's measurability, γ should be deadlock-free if ψ orders the agents, for all outcomes $(\omega, u) \in \mathrm{G}^\gamma$, such that at (ω, u), each agent's action only depends on ω and its predecessors' actions. This suggests the following design-dependent analogue of property CI.

DEFINITION 6. *A design $\gamma \in \Gamma$ possesses* property CI* (causal implementability) *when $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$ and there exists at least one map $\psi : \mathrm{G}^\gamma \to S_N$ such that for all $k = 1, 2, \ldots, N$, and $(\omega, u) \in \mathrm{G}^\gamma$,*

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\omega, u)) \subset \{\emptyset, [\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\omega, u))\} \tag{3.12}$$

when $s := (s_1, s_2, \ldots, s_N) = \psi(\omega, u)$.

As in property C*, for fixed $\gamma \in \Gamma$, the ψ in property CI* is a function that maps every outcome in G^γ into an N-agent ordering. Unlike property C*, however, this ψ is not constrained to be measurable in any sense. Instead, for all outcomes $(\omega, u) \in \mathrm{G}^\gamma$, the cylinder set

$$[\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\omega, u)) = [\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}}) \tag{3.13}$$

induced on $\Omega \times U$ by ω and the actions of the first $k-1$ agents in $s := (s_1, s_2, \ldots, s_N) = \psi(\omega, u)$ is constrained to be a subset of all events containing (ω, u) in the information partition $\mathcal{J}^{\gamma^{s_k}}$ induced by the s_kth agent's control law γ^{s_k}—i.e., no event in $\mathcal{J}^{\gamma^{s_k}}$ containing (ω, u), may depend on u^{s_k}, $u^{s_{k+1}}$, ..., or u^{s_N} (cf. [2, Def. 2]). Accordingly, property CI* ensures that for all outcomes $(\omega, u) \in \mathrm{G}^\gamma$, there exists an action order $s := (s_1, s_2, \ldots, s_N) = \psi(\omega, u)$ such that for all $k = 1, 2, \ldots, N$, the s_kth agent's action at the point (ω, u) does not depend on itself or the actions of its successors in s.

Clearly, the design in Example 1 does not satisfy this condition—when $\omega = 1$, all three agents' actions are interdependent. Such is not the case in the following three-agent example.

Example 2. Suppose that

$$\Omega = U^1 = U^2 = U^3 = [0, 1], \tag{3.14}$$

$$\mathcal{B} = \mathcal{U}^1 = \mathcal{U}^2 = \mathcal{U}^3 = \mathrm{Borel}[0, 1], \tag{3.15}$$

and

$$\gamma^1(\omega, u^1, u^2, u^3) = \begin{cases} 0 & \text{when } \omega \in [0, \frac{1}{2}), \\ \frac{1}{2} & \text{when } (\omega, u^2) \in [\frac{1}{2}, 1] \times [\frac{1}{2}, 1], \\ 1 & \text{else} \end{cases}$$

$$\gamma^2(\omega, u^1, u^2, u^3) = \begin{cases} 0 & \text{when } \omega \in [\frac{1}{2}, 1], \\ \frac{1}{2} & \text{when } (\omega, u^1) \in [0, \frac{1}{2}] \times [\frac{1}{2}, 1], \\ 1 & \text{else,} \end{cases} \tag{3.16}$$

$$\gamma^3(\omega, u^1, u^2, u^3) = \begin{cases} 0 & \text{when } \omega \in [0, \frac{1}{2}), \\ 1 & \text{else} \end{cases}$$

are the component policies of an admissible design $\gamma = (\gamma^1, \gamma^2, \gamma^3)$. It is straightforward to verify that

$$\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \mathcal{P}_\emptyset\left([0, \tfrac{1}{2}) \times \{(0, 1, 0)\} \bigcup [\tfrac{1}{2}, 1] \times \{(1, 0, 1)\}\right) = \Omega, \tag{3.17}$$

and that (3.12) is satisfied for all $k = 1, 2, 3$ and $(\omega, u) \in \mathrm{G}^\gamma$ when

$$\bar{\psi}(\omega, u^1, u^2, u^3) = \begin{cases} (1,2,3) & \text{when } \omega \in [0, \tfrac{1}{2}), \\ (2,1,3) & \text{else.} \end{cases} \tag{3.18}$$

Hence γ possesses property CI*.

Property CI* is of interest because it implies property SM* and provides a complete characterization of γ's deadlock-freeness.

THEOREM 2. *Let γ be an arbitrary design in Γ. Then*

(i) *γ possesses property* SM* *if γ possesses property* CI*, *and*

(ii) *γ possesses property* DF* *if and only if γ possesses property* CI*.

Proof. See Appendix C. □

Theorem 2 ensures that (P) is causal and well posed if and only if all designs $\gamma \in \Gamma_C$ possess property CI*. Its proof, like that of property CI [2], hinges on the following observation. When ψ is an order function such that γ possesses property CI*, for arbitrary but fixed $(\omega, u) \in \Omega \times U$, and $k = 1, 2, \ldots, N$, (3.12) and the fact that $\mathcal{U}^k$ contains the singletons of U^k imply that, at the point (ω, u), γ^{s_k}, $s = \psi(\omega, u)$, does not depend on the s_kth, s_{k+1}th, or s_Nth components of u. This suggests that for fixed $\gamma \in \Gamma$, a unique $\mathcal{B}$-measurable solution $\Sigma^\gamma : \Omega \to U$ to the closed-loop equation $u = \gamma(\omega, u)$ can be obtained by the following recursion.

Fix $\omega \in \mathcal{P}_\emptyset(\mathrm{G}^\gamma)$ and $u^\gamma_\omega \in \mathrm{G}^\gamma|_\omega$. Let $r \in U$ be an arbitrary reference element, let π_U and π_Ω denote the canonical projections of $\Omega \times U$ onto, respectively, U and Ω, let $L^\gamma : \Omega \times U \to \Omega \times U$ be defined as

$$L^\gamma(\omega, r) := (\omega, \gamma(\omega, r)), \tag{3.19}$$

and let $L^\gamma_k : \Omega \times U \to \Omega \times U$ be a k-fold composition of L^γ—i.e.,

$$L^\gamma_k(\omega, r) := (\underbrace{L^\gamma \circ \cdots \circ L^\gamma}_{k \text{ times}})(\omega, r). \tag{3.20}$$

Although (3.19) and (3.20) are nearly identical to (3.6) and (3.7) of [2], because the domain of ψ is G^γ (as opposed to $\Omega \times U$), the arguments following (3.7) in [2] no longer suffice to ensure $\pi_U \circ L^\gamma_N$ is the closed-loop solution map Σ^γ induced by γ. In particular, because $L^\gamma_k(\omega, r)$ need not belong to G^γ for all $r \in U$ and $k = 1, 2, \ldots, N$, a somewhat different argument is required to show that at least one agent's decision becomes invariant after every iteration. Formally, we have the following.

1. After one iteration, the components of $L^\gamma_1(\omega, r)$ corresponding to agents whose actions at the point (ω, r) do not depend on r become invariant to subsequent iterations. By property CI*, the set $\mathcal{A}_1(\omega) \subset \{1, 2, \ldots, N\}$ indexing (by agent) these components is nonempty since, at the point $(\omega, u^\gamma_\omega)$, at least agent $(\psi(\omega, u^\gamma_\omega))_1$'s action does not depend on r. Moreover, since r is arbitrary,

$$\mathcal{P}_i(L^\gamma_1(\omega, r)) = \mathcal{P}_i(L^\gamma_1(\omega, u^\gamma_\omega)) = \mathcal{P}_i(\omega, u^\gamma_\omega), \tag{3.21}$$

for all $i \in \mathcal{A}_1(\omega)$.

2. After two iterations, the components of $L^\gamma_2(\omega, r)$ corresponding to agents in $\{1, 2, \ldots, N\} \setminus \mathcal{A}_1(\omega)$ whose actions at the point $L^\gamma_1(\omega, r)$ do not depend on the components of agents in $\{1, 2, \ldots, N\} \setminus \mathcal{A}_1(\omega)$ become invariant to subsequent iterations.[6]

[6] For sets $A, B \subset X$, $A \setminus B := \{x \in A : x \notin B\}$, the complement of B relative to A.

By property CI*, the set $\mathcal{A}_2(\omega)$ indexing (by agent) these components is nonempty when $\text{card}(\mathcal{A}_1(\omega)) < N$ since, at the point $(\omega, u^\gamma_\omega)$, at least agent $(\psi(\omega, u^\gamma_\omega))_j$'s action,

$$j = \min\{m \in \{1,2,\dots,N\} : \ (\psi(\omega, u^\gamma_\omega))_m \notin \mathcal{A}_1(\omega)\}, \tag{3.22}$$

does not depend on the components of agents in $\{1,2,\dots,N\} \setminus \mathcal{A}_1(\omega)$, and by (3.21), the remaining components of $(\omega, u^\gamma_\omega)$ are identical to those of $L^\gamma_1(\omega, r)$. As before, since r is arbitrary,

$$\mathcal{P}_i(L^\gamma_2(\omega, r)) = \mathcal{P}_i(L^\gamma_2(\omega, u^\gamma_\omega)) = \mathcal{P}_i(\omega, u^\gamma_\omega) \tag{3.23}$$

for all $i \in \mathcal{A}_1(\omega) \cup \mathcal{A}_2(\omega)$.

$\vdots$

k. After k iterations, the components of $L^\gamma_k(\omega, r)$ corresponding to agents in $\{1,2,\dots,N\} \setminus \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)$ whose decisions at the point $L^\gamma_{k-1}(\omega, r)$ do not depend on the components of agents in $\{1,2,\dots,N\} \setminus \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)$ become invariant to subsequent iterations. By property CI*, the set $\mathcal{A}_k(\omega)$ indexing (by agent) these components is nonempty when $\text{card}(\bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)) < N$ since, at the point $(\omega, u^\gamma_\omega)$, at least agent $(\psi(\omega, u^\gamma_\omega))_j$'s action,

$$j = \min\left\{m \in \{1,2,\dots,N\} : \ (\psi(\omega, u^\gamma_\omega))_m \notin \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)\right\}, \tag{3.24}$$

does not depend on the components of agents in $\{1,2,\dots,N\} \setminus \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)$, and by the preceding iterations (e.g., (3.23)), the remaining components of $(\omega, u^\gamma_\omega)$ are identical to those of $L^\gamma_{k-1}(\omega, r)$. Once again, since r is arbitrary,

$$\mathcal{P}_i(L^\gamma_k(\omega, r)) = \mathcal{P}_i(L^\gamma_k(\omega, u^\gamma_\omega)) = \mathcal{P}_i(\omega, u^\gamma_\omega) \tag{3.25}$$

for all $i \in \bigcup_{i=1}^{k} \mathcal{A}_i(\omega)$.

$\vdots$

And so on....

Because property CI* ensures that, until all agents' components are invariant, at least one new component becomes invariant after every iteration, the recursive procedure must converge in, at most, N iterations—i.e., the unique solution to the closed-loop equation $u = \gamma(\omega, u)$ is $\pi_U(L^\gamma_N(\omega, r))$, where $r \in U$ is an arbitrary "seed" that starts the recursive solution process. Because π_Ω, π_U, and γ are, respectively, $\mathcal{B}\otimes\mathcal{U}/\mathcal{B}$-, $\mathcal{B}\otimes\mathcal{U}/\mathcal{U}$- and $\mathcal{B}\otimes\mathcal{U}/\mathcal{U}$-measurable, L^γ, and by composition, L^γ_k and $\pi_U \circ L^\gamma_N$, are, respectively, $\mathcal{B}\otimes\mathcal{U}/\mathcal{B}\otimes\mathcal{U}$-, $\mathcal{B}\otimes\mathcal{U}/\mathcal{B}\otimes\mathcal{U}$-, and $\mathcal{B}\otimes\mathcal{U}/\mathcal{U}$-measurable. It follows, because all u-sections of $\mathcal{B}\otimes\mathcal{U}/\mathcal{U}$-measurable functions are $\mathcal{B}/\mathcal{U}$-measurable, that the induced solution map $\Sigma^\gamma = \pi_U \circ L^\gamma_N|_r$ is necessarily $\mathcal{B}/\mathcal{U}$-measurable.

The preceding recursion has the same physical interpretation as the recursion in [2]. If for all k we ignore all components of $\pi_U(L^\gamma_k(\omega, r))$ except those corresponding to the agents indexed in $\mathcal{A}_k(\omega)$, the preceding recursion outlines the partial ordering

of agent actions that a passive observer would record, given ω, if the design γ were implemented in a "maximally" concurrent fashion. Although the recursion implicitly demonstrates that property CI* implies property DF*, it is far easier to establish sufficiency by a direct appeal to property CI*. For all $(\omega, u) \in \mathrm{G}^\gamma$ and $k = 1, 2, \ldots, N$, property CI* implies that at the point (ω, u), agent s_k's action does not depend on the s_kth, s_{k+1}th, ..., and s_Nth components of u. Consequently, no agent's action depends on its own action or the actions of its successors—i.e., γ must be deadlock-free.

The fact that γ must deadlock when property CI* fails to hold is also a direct consequence of property CI*'s definition. When $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) \neq \Omega$, for some $\omega \in \Omega$, the closed-loop equation has no solution; consequently, for that ω, γ has no implementation (let alone a deadlock-free implementation). Alternatively, suppose that there exists at least one outcome $(\omega, u) \in \mathrm{G}^\gamma$ such that for all N-agent orderings $s := (s_1, s_2, \ldots, s_N) \in S_N$, (3.12) fails for at least one $k \in \{1, 2, \ldots, N\}$, say k_s. Then, for all orderings $s \in S_N$, the s_{k_s}th agent's action, at the point (ω, u), always depends on itself and or the actions of its successors in s, and once again, γ is not deadlock-free.

3.4. Are properties C* and CI* equivalent? By Theorems 1(ii) and 2(ii), property C* implies DF*, which in turn implies property CI*. Consequently, we have the following.

COROLLARY 1. *Property* C* *implies property* CI*.

Proof. See Appendix D for a direct proof. □

Are properties C* and CI* equivalent? When $N = 1$, the answer is yes (this follows from Definition 7 and Theorem 3). When $N > 1$, it is not known (in general) whether property CI* implies property C*. In particular, attempts to establish a design-dependent analogue of Corollary 2 in [2] (i.e., that CI* implies C* when $N = 2$) are complicated by the fact that S* need not imply CI* (or C*) under any circumstances (cf. [9, Thm. 2]). Consider, for instance, the following one-agent example.

Example 3. Suppose that $\Omega = \{0, 1\}$ and $U = \{0, 1, 2\}$, and let

$$\gamma(\omega, u) = \begin{cases} 2 & \text{if } (\omega, u) \in \{(1,1), (1,2)\}, \\ 1 & \text{if } (\omega, u) = (1,0), \\ 0 & \text{else.} \end{cases} \tag{3.26}$$

Because

$$\begin{aligned} \mathrm{G}^\gamma &:= \{(\omega, u) : \gamma(\omega, u) = u\} \\ &= \{(0,0), (1,2)\}, \end{aligned} \tag{3.27}$$

γ possesses property S*. But $[\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset(1,2)) = \{(1,0), (1,1), (1,2)\}$ and $[\gamma]^{-1}(1) = \{(1,0)\}$; consequently,

$$[\gamma]^{-1}(1) \bigcap [\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset(1,2)) = \{(1,0)\} \notin \{\emptyset, \{(1,0), (1,1), (1,2)\}\}. \tag{3.28}$$

Hence γ does not process property CI*.

Properties CI* and C* are equivalent in at least two important cases: when γ is *sequential* (Theorem 3), and when the measurable structure underlying (P) is *discrete*, i.e., when $\mathcal{B} \otimes \mathcal{U}$ contains the singletons of $\Omega \times U$ and $\Omega \times U$ is a countable set (Theorem 4).

DEFINITION 7. *A design $\gamma \in \Gamma$ is said to be* sequential *when property* CI* *holds for some constant order function ψ.*

THEOREM 3. *All constant order functions ψ such that a design $\gamma \in \Gamma$ possesses property* CI* *are order functions such that γ possesses property* C*.

Proof. See Appendix E. □

THEOREM 4. *When Ω and U^k, $k = 1, 2, \ldots, N$, are countable sets, and $\mathcal{B}$ contains the singletons of Ω, all order functions ψ such that a design $\gamma \in \Gamma$ possesses property* CI* *are order functions such that γ possesses property* C*.

Proof. See Appendix F. □

When $\gamma \in \Gamma$ is nonsequential and (P)'s measurable structure is not discrete, it is far more difficult to prove that property CI* implies property C* because, even if γ possesses property C*, order functions for which γ possesses property CI* need not be order functions for which γ possesses property C*.

Example 4. Consider again the three-agent design of Example 2. Although the design γ defined in (3.16) possesses properties CI* and C*, when A is any nonmeasurable subset of $[0, \frac{1}{2})$ (such a set always exists [4]),

$$\psi(\omega, u^1, u^2, u^3) = \begin{cases} (1,2,3) & \text{when } \omega \in [0, \tfrac{1}{2})/A, \\ (3,1,2) & \text{when } \omega \in A, \\ (2,1,3) & \text{else} \end{cases} \tag{3.29}$$

is an order function such that γ possesses property CI*, but not property C*. To see this, note that (3.12) holds for all $k = 1, 2, 3$, and $s \in S_k$, whereas (3.11) fails, for instance, when $k = 1$ and $s = 3 \in S_1$, since

$$\begin{aligned} [\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset([T_1^3 \circ \psi]^{-1}(3))) &= A \times U \\ &\notin \mathcal{F}(\emptyset) \bigcap [\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset(\mathrm{G}^\gamma)) \\ &= \mathcal{B} \otimes \{\emptyset, U\}. \end{aligned} \tag{3.30}$$

The fact that there exist nonsequential designs $\gamma \in \Gamma$ and order functions ψ such that γ possesses property CI*, but not property C*, implies that general proofs that property CI* implies property C* (if such exist) must be constructive—i.e., to prove that property CI* implies property C*, given a ψ such that γ possesses property CI*, but not property C*, we must be able to construct a new order function $\widehat{\psi}$ (obviously distinct from ψ), such that γ possesses property C*. To date, no such constructions are known.

4. Design-independence vs. design-dependence. In this section we briefly examine the relationships between the design-independent properties introduced in [2] (part I) and the design-dependent properties introduced here (part II).

4.1. Basic relationships.

THEOREM 5. *Let $\mathcal{I}$ be an arbitrary information structure. Then*

(i) *all $\gamma \in \Gamma$ possess property* S* *if and only if $\mathcal{I}$ possesses property* S,

(ii) *all $\gamma \in \Gamma$ possess property* SM* *if and only if $\mathcal{I}$ possesses property* SM,

(iii) *all $\gamma \in \Gamma$ possess property* CI* *if and only if $\mathcal{I}$ possesses property* CI, *and*

(iv) *all $\gamma \in \Gamma$ possess property* C* *if $\mathcal{I}$ possesses property* C.

Proof. See Appendix G. □

Parts (i) and (ii) are immediate consequences of the definitions of properties S, SM, S*, and SM*. Part (iii) follows from the fact that properties CI and CI* are

necessary and sufficient conditions for, respectively, all designs $\gamma \in \Gamma$ and particular designs $\gamma \in \Gamma$, to be deadlock-free. If properties C and C* were known to provide necessary and sufficient conditions for, respectively, all designs and particular designs to be deadlock-free (as is the case, for instance, when Ω and U are countable sets and $\mathcal{B}$ contains the singletons of Ω), the proof of part (iv), with the *if* replaced by *if and only if*, would also be immediate. In the absence of such knowledge it is necessary to prove (iv)—and if possible, the converse of (iv)—constructively. The forward construction is straightforward. Given a ψ such that $\mathcal{I}$ possesses property C, simply let $\psi^\gamma = \psi|_{G^\gamma}$ (the restriction of ψ to G^γ) for each $\gamma \in \Gamma$. The reverse construction (if such exists) is not obvious since there does not seem to be any way of relating the set of order functions

$$\bigcup_{\gamma\in\Gamma}\{\psi^\gamma : \gamma \text{ possesses property C* given } \psi^\gamma\} \tag{4.1}$$

to an order function ψ such that $\mathcal{I}$ possesses property C.

4.2. Design-dependent characterizations are finer. By Theorem 5, an information structure $\mathcal{I}$ cannot possess the design-independent property CI (respectively, C, SM, or S) if any one of its designs $\gamma \in \Gamma$ fails to possess the design-dependent property CI* (respectively, C*, SM*, or S*). This suggests that the design-dependent properties provide a finer characterization of a design's closed-loop solvability and deadlock-freeness, than the design-independent properties.

THEOREM 6. *For $N > 2$, there exist designs possessing property* C* (*and consequently, properties* CI*, SM*, *and* S*) *that cannot be associated with any deadlock-free information structure possessing property* S, *let alone properties* SM, CI, *or* C.

Proof. Since C* $\Rightarrow$ CI* $\Rightarrow$ SM* $\Rightarrow$ S* (by Corollary 1, Theorem 2, and Definition 3), and since C $\Rightarrow$ CI $\Rightarrow$ SM $\Rightarrow$ S (by [2, Cor. 1 and Thm. 2] and [9, §4]), it suffices to construct a design possessing property C* that cannot be associated with any information structure possessing property S.

Example 5. Consider a nonsequential $\mathcal{I}$ of the following form:

$$\begin{aligned}
&N = 3,\\
&\Omega = U^1 = U^2 = U^3 = \{0,1\},\\
&\mathcal{B} = \mathcal{U}^1 = \mathcal{U}^2 = \mathcal{U}^3 = \Big\{\emptyset, \{0\}, \{1\}, \{0,1\}\Big\},\\
&\mathcal{J}^1 = \Big\{\emptyset, \{(\omega,u) : \omega = 0\}, \{(\omega,u) : \omega = 1\}, \Omega \times U\Big\},\\
&\mathcal{J}^2 = \Big\{\emptyset, \{(\omega,u) : \max(\bar{\omega}\bar{u}^1\bar{u}^3, u^1u^3) = 0\},\\
&\qquad\qquad \{(\omega,u) : \max(\bar{\omega}\bar{u}^1\bar{u}^3, u^1u^3) = 1\}, \Omega \times U\Big\},
\end{aligned} \tag{4.2}$$

(4.3)

and

$$\mathcal{J}^3 = \Big\{\emptyset, \{(\omega,u) : \omega u^2 = 0\}, \{(\omega,u) : \omega u^2 = 1\}, \Omega \times U\Big\}.$$

Since the closed-loop equations for the design $\hat{\gamma} = (\hat{\gamma}^1, \hat{\gamma}^2, \hat{\gamma}^3)$,

$$\begin{aligned}
\hat{\gamma}^1(\omega, u^1, u^2, u^3) &= \begin{cases} 1 & \omega = 1,\\ 0 & \text{else,}\end{cases}\\
\hat{\gamma}^2(\omega, u^1, u^2, u^3) &= \begin{cases} 1 & \max(\bar{\omega}\bar{u}^1\bar{u}^3, u^1u^3) = 1,\\ 0 & \text{else,}\end{cases}
\end{aligned} \tag{4.4}$$

$$\widehat{\gamma}^3(\omega, u^1, u^2, u^3) = \begin{cases} 1 & \omega u^2 = 1, \\ 0 & \text{else}, \end{cases}$$

exhibit two distinct outcomes when $\omega = 1$—i.e.,

$$\begin{aligned} \mathrm{G}^{\widehat{\gamma}} &:= \{(\omega, u) : \ \widehat{\gamma}(\omega, u) = u\} \\ &= \{(0,0,1,0), (1,1,0,0), (1,1,1,1)\} \end{aligned} \tag{4.5}$$

—$\widehat{\gamma}$ does not possess property S*. Consequently, no information structure (including $\mathcal{I}$) that can be associated with $\widehat{\gamma}$ can possess property S (Theorem 5)—i.e., no information structure

$$\widehat{\mathcal{I}} := \{(\Omega, \mathcal{B}), (U^k, \mathcal{U}^k), \widehat{\mathcal{J}}^k : \ 1 \leq k \leq 3\} \tag{4.6}$$

such that

$$[\widehat{\gamma}^k]^{-1}(\mathcal{U}^k) = \mathcal{J}^k \subset \widehat{\mathcal{J}}^k, \qquad 1 \leq k \leq 3 \tag{4.7}$$

can possess property S.

Consider, however, the design $\gamma = (\gamma^1, \gamma^2, \gamma^3)$,

$$\begin{aligned} \gamma^1(\omega, u^1, u^2, u^3) &= \begin{cases} 0 & \omega = 1, \\ 1 & \text{else}, \end{cases} \\ \gamma^2(\omega, u^1, u^2, u^3) &= \widehat{\gamma}^2(\omega, u^1, u^2, u^3), \\ \gamma^3(\omega, u^1, u^2, u^3) &= \widehat{\gamma}^3(\omega, u^1, u^2, u^3). \end{aligned} \tag{4.8}$$

This design possesses property S*—i.e.,

$$\begin{aligned} \mathrm{G}^{\gamma} &:= \{(\omega, u) : \ \gamma(\omega, u) = u\} \\ &= \{(0,1,0,0), (1,0,0,0)\}. \end{aligned} \tag{4.9}$$

Moreover, when

$$\psi(\omega, u^1, u^2, u^3) = \begin{cases} (1,3,2) & (\omega, u^1, u^2, u^3) = (0,1,0,0), \\ (1,2,3) & (\omega, u^1, u^2, u^3) = (1,0,0,0), \end{cases} \tag{4.10}$$

it can also be shown to possess property C*.[7] But for all $k = 1,2,3$, γ^k and $\widehat{\gamma}^k$ both induce the same information subfield $\mathcal{J}^k$ (i.e., $\mathcal{J}^k = [\gamma^k]^{-1}(\mathcal{U}^k) = [\widehat{\gamma}^k]^{-1}(\mathcal{U}^k)$, for all $k = 1,2,3$). Accordingly, even though γ possesses property C*, it cannot be associated with any information structure possessing property S. This proves the theorem. □

Heuristically, the three-agent information structure that appears in the preceding example can be viewed as a synthesis, parameterized by agent 1's $\mathcal{F}(\emptyset)$-measurable decision[8] of two different two-agent information structures for agents 2 and 3. The first of these structures, $\mathcal{I}^C$, corresponds to the restriction of agent 2 and agent 3's information subfields to the u^1-sections of $\mathcal{J}^2$ and $\mathcal{J}^3$ induced when $u^1 = 0$—i.e.,

$$\mathcal{I}^C := \left\{(\Omega, \mathcal{B}), \ (U^i, \mathcal{U}^i), \mathcal{J}^i|_{u^1=0} : \ 2 \leq i \leq 3\right\}, \tag{4.11}$$

[7] In this case it is somewhat easier to check property CI* and then apply Theorem 4.

[8] By (4.2), $\mathcal{J}^1 \subset \mathcal{F}(\emptyset) := \{\emptyset, \{0\} \times U, \{1\} \times U, \Omega \times U\}$.

where

$$\mathcal{J}^2|_{u^1=0} = \Big\{\emptyset, \{(\omega, u^2, u^3) : \ \bar{\omega}\bar{u}^3 = 0\}, \{(\omega, u^2, u^3) : \ \bar{\omega}\bar{u}^3 = 1\}, \Omega \times U^2 \times U^3\Big\} \tag{4.12}$$

and

$$\mathcal{J}^3|_{u^1=0} = \Big\{\emptyset, \{(\omega, u^2, u^3) : \ \omega u^2 = 0\}, \{(\omega, u^2, u^3) : \ \omega u^2 = 1\}, \Omega \times U^2 \times U^3\Big\}. \tag{4.13}$$

The second of these structures, $\mathcal{I}^{NS}$, corresponds to the restriction of agent 2 and agent 3's information subfields to the u^1-sections of $\mathcal{J}^2$ and $\mathcal{J}^3$ induced when $u^1 = 1$—i.e.,

$$\mathcal{I}^{NS} := \Big\{(\Omega, \mathcal{B}), (U^i, \mathcal{U}^i), \mathcal{J}^i|_{u^1=1} : \ 2 \leq i \leq 3\Big\}, \tag{4.14}$$

where

$$\mathcal{J}^2|_{u^1=1} = \Big\{\emptyset, \{(\omega, u^2, u^3) : \ u^3 = 0\}, \{(\omega, u^2, u^3) : \ u^3 = 1\}, \Omega \times U^2 \times U^3\Big\} \tag{4.15}$$

and

$$\mathcal{J}^3|_{u^1=1} = \Big\{\emptyset, \{(\omega, u^2, u^3) : \ \omega u^2 = 0\}, \{(\omega, u^2, u^3) : \ \omega u^2 = 1\}, \Omega \times U^2 \times U^3\Big\}. \tag{4.16}$$

It is not difficult to verify that $\mathcal{I}^C$ possesses property C when

$$\psi(\omega, u^2, u^3) = \begin{cases} (3,2) & \omega = 0, \\ (2,3) & \text{else.} \end{cases} \tag{4.17}$$

To see this note that

$$\mathcal{J}^2|_{u^1=0} \bigcap [T_1^2 \circ \psi]^{-1}(2) = \Big\{\emptyset, \{0\} \times U^2 \times U^3\Big\} \subset \mathcal{F}(\emptyset)|_{u^1=0} \tag{4.18}$$

and

$$\mathcal{J}^3\big|_{u^1=0} \bigcap [T_1^2 \circ \psi]^{-1}(3) = \Big\{\emptyset, \{1\} \times U^2 \times U^3\Big\} \subset \mathcal{F}(\emptyset)|_{u^1=0}\,. \tag{4.19}$$

$\mathcal{I}^{NS}$, however, does not even possess property S. For instance,

$$\begin{aligned} \mathrm{G}^\gamma &:= \{(\omega, u^2, u^3) : \ \gamma(\omega, u^2, u^3) = (u^2, u^3)\} \\ &= \{(0,0,0), (1,0,0), (1,1,1)\} \end{aligned} \tag{4.20}$$

when

$$\begin{aligned} \gamma^2(\omega, u^2, u^3) &= \begin{cases} 1 & u^3 = 1, \\ 0 & \text{else,} \end{cases} \\ \gamma^3(\omega, u^2, u^3) &= \begin{cases} 1 & \omega u^2 = 1, \\ 0 & \text{else.} \end{cases} \end{aligned} \tag{4.21}$$

It follows, because agent 1's decision determines whether agent 2 and agent 3's interdependence is characterized by $\mathcal{I}^C$ $(u^1 = 0)$ or $\mathcal{I}^{NS}(u^1 = 1)$, that agent 1's control law determines whether nontrivial designs for the synthesized system (4.4) possess property C* or do not possess property S*. Specifically, all designs such that

$$\gamma^1(\omega, u^1, u^2, u^3) = 1 \tag{4.22}$$

or

$$\gamma^1(\omega, u^1, u^2, u^3) = \begin{cases} 1 & \omega = 1, \\ 0 & \text{else}, \end{cases} \tag{4.23}$$

and neither γ^2 nor γ^3 is a constant policy (there are 8 such designs since $\text{card}(\mathcal{J}^k) > 2$ and $\text{card}(U^k) = 2$ for $k = 2, 3$), do not possess property S*. All remaining designs (there are 56) possess property C*.

Clearly, the preceding heuristic can be used to synthesize far more complicated information structures that fail to possess property S, but nonetheless admit nontrivial designs possessing property C*. For instance, noncausal and causal 2-agent information structures can be combined, when parameterized by two additional agents' decisions, to form a 4-agent information structure that fails to possess property S but admits nontrivial designs possessing property C*; similarly, this 4-agent information structure and a second 4-agent information structure can be combined, when parameterized by three additional agents' decisions, to form a 7-agent information structure that fails to possess property S but admits nontrivial designs possessing property C*; and so on. It follows that there exist a myriad of designs whose deadlock-freeness and closed-loop solvability can not be characterized using any design-independent property.

5. Conclusions. In this paper we have introduced conditions (properties C* and CI*) necessary and sufficient to ensure the deadlock-freeness (property DF*) and measurable closed-loop solvability (property SM*) of a nonsequential design γ represented within the framework of Witsenhausen's intrinsic model. We have also shown, by example, that there exist nontrivial, deadlock-free designs that cannot be associated with any deadlock-free information structure.

Our conditions, which are the design-dependent analogues of conditions in [2] and [9] (properties CI and C), provide an intuitive characterization of the cause/effect notion of causality in terms of the events that a system's decision-making agents can distinguish, and suggest a framework for the optimization of constrained nonsequential stochastic control problems.

The existence of deadlock-free designs that cannot be associated with any deadlock-free information structure is not surprising. Many network routing, flow, and concurrency control systems are seen to be deadlock-free under some designs and deadlock-prone under others. In fact, unless the specification of a nonsequential system's agents' information subfields is coordinated (in practice physical constraints, complexity and/or cost may preclude such coordination) it is unlikely that the system's information structure will possess any design-independent property. Moreover, as illustrated by Example 5, the deadlock-freeness and closed-loop solvability of the admissible designs for such systems may hinge on the control laws of a small fraction of the agents. The only difference between the designs $\widehat{\gamma}$ and γ of Example 5, for instance, is that $\widehat{\gamma}^1$'s decision is the binary complement of γ^1's decision. Nonetheless, although $\widehat{\gamma}$ does not possess any design-dependent property, γ possesses all of the

known design-dependent properties. Simply put, the inappropriate use of information by a single agent can give rise to deadlocks.

One final note. In [9, p. 159] it is remarked that the "physical interpretation" of information structures possessing property SM, but not property C, "appears difficult" (the difficulty being the host of paradoxes that arise when effects precede their causes). In light of Example 5, it would seem, rather, that it is the physical interpretation of *designs* possessing property SM* but not property CI* that may be difficult.

Appendix A.

Proof of Lemma 1. Fix $\gamma \in \Gamma$ and suppose that $\psi : \mathrm{G}^\gamma \to S_N$ is an order function such that γ possesses property c*. Except for the restriction of ψ's domain to G^γ, the proof that c* implies SM* parallels the proof that C implies SM in [9, Thm. 1]. Note, however, that unlike Witsenhausen's kth umpire update map [9, §7], the analogous update map, $M_k^\gamma : \mathrm{G}^\gamma \to \mathrm{G}^\gamma$ with

$$\text{(A.1)}\quad \mathcal{P}_\alpha(M_k^\gamma(\omega,u)) := \begin{cases} (\omega, \gamma^\alpha(\omega,u)) & \text{when } \alpha = (\psi(\omega,u))_k, \\ \mathcal{P}_\alpha(\omega, u_\omega^\gamma) & \text{when } \alpha = (\psi(\omega,u))_j, \quad j = k+1,\dots,N, \\ \mathcal{P}_\alpha(\omega,u) & \text{otherwise} \end{cases}$$

for all $\alpha \in \{\emptyset, 1, \dots, N\}$, cannot be used to establish γ's deadlock-freeness because the restriction of M_k^γ to G^γ permits the umpire to know the actions of agents before they have acted.

To see that c* need not imply property DF*, note that although the design in Example 1 is not deadlock-free, for all $\psi : \mathrm{G}^\gamma \to S_3$, it trivially satisfies property c* because, as pointed out in §3.2, for all $k = 1, 2, 3$, and $s \in S_k$, $\mathcal{F}(T_{k-1}^k(s)) \bigcap \mathrm{G}^\gamma$ is the power set of G^γ (see (3.4)). □

Appendix B. Proof of Theorem 1. To prove Theorem 1 we need the following facts.

FACT 1. *For all $s \in S_k$, $k = 1, 2, \dots, N$, if $\mathcal{P}_s(\omega,u) = \mathcal{P}_s(\bar\omega,\bar u)$ for some (ω,u) and $(\bar\omega,\bar u) \in \mathrm{G}^\gamma$, then no set in $\mathcal{F}(s) \bigcap \mathrm{G}^\gamma$ contains (ω,u) but not $(\bar\omega,\bar u)$.*

Proof of Fact 1. Suppose that the fact fails for some $s \in S_k$, $(\omega,u) \in \mathrm{G}^\gamma$, and $(\bar\omega,\bar u) \in \mathrm{G}^\gamma$. Then because

$$\text{(B.1)}\qquad \mathcal{F}(s) := [\mathcal{P}_s]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^k \mathcal{U}^{s_i}\right)\right),$$

there exists a set $A \in \mathcal{B} \otimes (\bigotimes_{i=1}^k \mathcal{U}^{s_i})$ such that

$$\text{(B.2)}\qquad (\omega,u) \in [\mathcal{P}_s]^{-1}(A) \bigcap \mathrm{G}^\gamma$$

and

$$\text{(B.3)}\qquad (\bar\omega,\bar u) \notin [\mathcal{P}_s]^{-1}(A) \bigcap \mathrm{G}^\gamma.$$

It follows that ω and $\bar\omega$, or least one of the s_1th through s_kth components of u and $\bar u$, must differ. But $\mathcal{P}_s(\omega,u) = \mathcal{P}_s(\bar\omega,\bar u)$, a contradiction. Accordingly, the fact holds. □

FACT 2. *Property* C* *implies property* c*.

Proof of Fact 2. Fix $\gamma \in \Gamma$ and suppose that ψ is an order function such that γ possesses property C*. Because property C* ensures that $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$, it suffices to show that ψ is also an order function such that γ possesses property c*.

The restriction of (3.11) to G^γ yields the desired result—(3.5) of property c*—if, for all $k = 1, 2, \ldots, N$, and $s \in S_k$,

$$[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \bigcap G^\gamma = [T_k^N \circ \psi]^{-1}(s). \tag{B.4}$$

By construction, the right side of (B.4) is a subset of the left. Suppose that the converse inclusion fails for some $k \in \{1, 2, \ldots, N\}$ and $s \in S_k$. Then there exists an outcome

$$(\bar{\omega}, \bar{u}) \in [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \bigcap G^\gamma \tag{B.5}$$

that is not in $[T_k^N \circ \psi]^{-1}(s)$. Moreover,

$$\mathcal{P}_{T_{k-1}^k(s)}(\bar{\omega}, \bar{u}) \in \mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s)). \tag{B.6}$$

It follows from (B.6) that there exists an outcome $(\omega, u) \in [T_k^N \circ \psi]^{-1}(s)$ such that

$$\mathcal{P}_{T_{k-1}^k(s)}(\omega, u) = \mathcal{P}_{T_{k-1}^k(s)}(\bar{\omega}, \bar{u}), \tag{B.7}$$

and (ω, u) and $(\bar{\omega}, \bar{u})$ differ in one or more of the components of u not indexed in $T_{k-1}^k(s)$.

But this is impossible. By property C*, the sets

$$[\gamma^{s_k}]^{-1}(\bar{u}^{s_k}) \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \bigcap G^\gamma \tag{B.8}$$

and

$$[\gamma^{s_k}]^{-1}(u^{s_k}) \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \bigcap G^\gamma \tag{B.9}$$

are elements of $\mathcal{F}(T_{k-1}^k(s)) \bigcap G^\gamma$. If $\bar{u}^{s_k} \neq u^{s_k}$, $[\gamma^{s_k}]^{-1}(\bar{u}^{s_k})$ and $[\gamma^{s_k}]^{-1}(u^{s_k})$, and consequently the sets in (B.8) and (B.9), are disjoint. Since $(\bar{\omega}, \bar{u})$ and (ω, u) satisfy (B.7), this contradicts Fact 1.

Similarly, if, for $\bar{s} = T_j^N(\psi(\omega, u))$, $j > k$,

$$\mathcal{P}_{T_{j-1}^j(\bar{s})}(\omega, u) = \mathcal{P}_{T_{j-1}^j(\bar{s})}(\bar{\omega}, \bar{u}), \tag{B.10}$$

and $\bar{u}^{\bar{s}_j} \neq u^{\bar{s}_j}$, then by property C*,

$$[\gamma^{\bar{s}_j}]^{-1}(\bar{u}^{\bar{s}_j}) \bigcap [\mathcal{P}_{T_{j-1}^j(\bar{s})}]^{-1}(\mathcal{P}_{T_{j-1}^j(\bar{s})}([T_j^N \circ \psi]^{-1}(\bar{s}))) \bigcap G^\gamma \tag{B.11}$$

and

$$[\gamma^{\bar{s}_j}]^{-1}(u^{\bar{s}_j}) \bigcap [\mathcal{P}_{T_{j-1}^j(\bar{s})}]^{-1}(\mathcal{P}_{T_{j-1}^j(\bar{s})}([T_j^N \circ \psi]^{-1}(\bar{s}))) \bigcap G^\gamma \tag{B.12}$$

are disjoint sets in $\mathcal{F}(T_{j-1}^j(\bar{s})) \bigcap G^\gamma$. Since $(\bar{\omega}, \bar{u})$ and (ω, u) satisfy (B.10), Fact 1 is once again contradicted. It follows, by induction, that $(\omega, u) = (\bar{\omega}, \bar{u})$. Hence, Fact 2 is proved. □

Proof of Theorem 1. Fix $\gamma \in \Gamma$ and suppose that $\psi : G^\gamma \to S_N$ is an order function such that γ possesses property C*. The proof of (i) follows from Lemma 1 and Fact 2.

To prove (ii) it suffices to show that all agents can act without precognition for all outcomes in G^γ. Fix $(\omega, u) \in \mathrm{G}^\gamma$. The first agent to act under ψ is agent $s_1 = T_1^N(\psi(\omega, u))$. Since $T_0^1(s_1) = \emptyset$, property C* implies that

$$\mathcal{J}^{\gamma^{s_1}} \bigcap [\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset([T_1^N \circ \psi]^{-1}(s_1))) \subset \mathcal{F}(\emptyset) \bigcap [\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset(\mathrm{G}^\gamma)). \tag{B.13}$$

Because

$$\{\omega\} \times U \in [\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset([T_1^N \circ \psi]^{-1}(s_1))) \subset [\mathcal{P}_\emptyset]^{-1}(\mathcal{P}_\emptyset(\mathrm{G}^\gamma)), \tag{B.14}$$

and

$$\begin{aligned} \mathcal{F}(\emptyset) \bigcap (\{\omega\} \times U) &= (\mathcal{B} \otimes \{\emptyset, U\}) \bigcap (\{\omega\} \times U) \\ &= \{\emptyset, \{\omega\} \times U\}, \end{aligned} \tag{B.15}$$

the restriction of (B.13) to $\{\omega\} \times U$ can be rewritten as

$$\mathcal{J}^{\gamma^{s_1}} \bigcap (\{\omega\} \times U) \subset \{\emptyset, \{\omega\} \times U\}. \tag{B.16}$$

But (B.16) implies that at the point (ω, u), γ^{s_1} does not depend on u (recall that $\mathcal{J}^{\gamma^{s_1}} := [\gamma^{s_1}]^{-1}(U^{s_1})$); consequently, given ω, agent s_1 acts without precognition.

Now, suppose that $k-1$ agents (agents $s_1, s_2, \ldots, s_{k-1}$) have acted without precognition and in accordance with ψ (i.e., $s = T_{k-1}^N(\psi(\omega, u))$). The kth agent to act under ψ is agent $s_k = (T_k^N(\psi(\omega, u)))_k$. Since $T_{k-1}^k(s, s_k) = s$, property C* implies that

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_s]^{-1}(\mathcal{P}_s([T_k^N \circ \psi]^{-1}(s, s_k))) \subset \mathcal{F}(s) \bigcap [\mathcal{P}_s]^{-1}(\mathcal{P}_s(\mathrm{G}^\gamma)). \tag{B.17}$$

Because

$$\begin{aligned} [\mathcal{P}_s]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}}) &\in [\mathcal{P}_s]^{-1}(\mathcal{P}_s([T_k^N \circ \psi]^{-1}(s, s_k))) \\ &\subset [\mathcal{P}_s]^{-1}(\mathcal{P}_s(\mathrm{G}^\gamma)) \end{aligned} \tag{B.18}$$

and

$$\begin{aligned} &\mathcal{F}(s) \bigcap [\mathcal{P}_s]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}}) \\ &\quad = [\mathcal{P}_s]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i}\right)\right) \bigcap [\mathcal{P}_s]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}}) \\ &\quad = \{\emptyset, [\mathcal{P}_s]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}})\}, \end{aligned} \tag{B.19}$$

the restriction of (B.17) to $[\mathcal{P}_s]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}})$ can be rewritten as

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_s]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}}) \subset \{\emptyset, [\mathcal{P}_s]^{-1}(\omega, u^{s_1}, \ldots, u^{s_{k-1}})\}. \tag{B.20}$$

But (B.20) implies that at the point (ω, u), γ^{s_k} does not depend on the s_kth, s_{k+1}th, ..., or s_Nth components of u; consequently, when nature and agents $s_1, s_2, \ldots, s_{k-1}$'s actions are $(\omega, u^{s_1}, \ldots, u^{s_{k-1}})$, agent s_k acts without precognition. It follows, by induction, that all agents act without precognition. Thus γ possesses property DF* and the theorem is proved. □

Appendix C.

Proof of Theorem 2. (i). Fix $\gamma \in \Gamma$, let $\mathrm{G}^\gamma := \{(\omega, u) \in \Omega \times U : \gamma(\omega, u) = u\}$, and suppose that ψ is an order function such that γ possesses property CI*. By assumption, the closed-loop equation $\gamma(\omega, u) = u$ admits at least one solution $u_\omega^\gamma \in \mathrm{G}^\gamma|_\omega := \{u \in U : \gamma(\omega, u) = u\}$ for all $\omega \in \Omega$ (i.e., $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$); hence, to prove that γ possesses property SM*, it suffices to show, for each $\omega \in \Omega$, that this solution is unique, and that the mapping $\Sigma^\gamma : \Omega \to U$ induced by these solutions (i.e., $\Sigma^\gamma(\omega) = u_\omega^\gamma$) is $\mathcal{B}/\mathcal{U}$-measurable (cf. Definitions 2 and 3).

Uniqueness. Fix $\omega \in \Omega$ and $u_\omega^\gamma \in \mathrm{G}^\gamma|_\omega$, and let $s := (s_1, s_2, \ldots, s_N) = \psi(\omega, u_\omega^\gamma)$. Let π_U denote the canonical projection of $\Omega \times U$ onto U, let $L^\gamma : \Omega \times U \to \Omega \times U$ be defined as in (3.19) and (3.20). Clearly, $u = \pi_U(L_N^\gamma(\omega, u))$ for all $u \in \mathrm{G}^\gamma|_\omega$, including u_ω^γ. Accordingly, to establish the uniqueness of u_ω^γ, it suffices to show that

$$(\omega, u_\omega^\gamma) = L_N^\gamma(\omega, r) \tag{C.1}$$

for all $r \in U$ (since $\mathrm{G}^\gamma|_\omega \subset U$), or equivalently, that

$$\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma) = \mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r)) \tag{C.2}$$

when $k = N + 1$.

Fix $r \in U$. When $k = 1$, $T_{k-1}^N(s) = \emptyset$ and

$$\begin{aligned} \mathcal{P}_\emptyset(\omega, u_\omega^\gamma) &= (\omega) \\ &= \mathcal{P}_\emptyset(\omega, \gamma(L_{N-1}^\gamma(\omega, r))) \\ &= \mathcal{P}_\emptyset(L_N^\gamma(\omega, r)). \end{aligned} \tag{C.3}$$

Suppose, for $k > 1$, that (C.2) holds. Since property CI* holds with order function ψ,

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma)) \subset \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma))\}. \tag{C.4}$$

Equation (C.4) and the fact that $\mathcal{U}^{s_k}$ contains the singletons of U^{s_k} (§2, 1(c)), implies that at the point $(\omega, u_\omega^\gamma) \in \Omega \times U$, γ^{s_k} does not depend on the s_kth, s_{k+1}th, ..., or s_Nth components of u_ω^γ (recall that $\mathcal{J}^{\gamma^{s_k}} := [\gamma^{s_k}]^{-1}(U^{s_k})$). Accordingly, (C.2) implies that

$$\gamma^{s_k}(\omega, u_\omega^\gamma) = \gamma^{s_k}(L_N^\gamma(\omega, r)), \tag{C.5}$$

and consequently, that

$$\begin{aligned} \mathcal{P}_{T_k^N(s)}(\omega, u_\omega^\gamma) &= (\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma), \gamma^{s_k}(\omega, u_\omega^\gamma)) \\ &= (\mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r)), \gamma^{s_k}(L_N^\gamma(\omega, r))) \\ &= \mathcal{P}_{T_k^N(s)}(L_N^\gamma(\omega, r)). \end{aligned} \tag{C.6}$$

It follows, by induction, that (C.2) holds for all $k = 1, 2, \ldots, N + 1$; hence, $(\omega, u_\omega^\gamma) = L_N^\gamma(\omega, r)$ for all $r \in U$, and consequently, the unique solution u_ω^γ to the closed-loop equation $u = \gamma(\omega, u)$ is $\pi_U(L_N^\gamma(\omega, r))$, where $r \in U$ is the (arbitrary) "seed" that starts the recursive solution process.

Measurability. Fix $r \in U$ and let π_U and π_Ω denote, respectively, the canonical projections of $\Omega \times U$ onto U and Ω. To establish the $\mathcal{B}/\mathcal{U}$-measurability of the induced closed-loop solution map $\Sigma^\gamma : \Omega \to U$, it suffices to show that the u-section of $\pi_U \circ L_N^\gamma$, $\pi_U \circ L_N^\gamma|_r$, is $\mathcal{B}/\mathcal{U}$-measurable because, for fixed r,

$$\Sigma^\gamma(\omega) = (\pi_U \circ L_N^\gamma|_r)(\omega) := (\pi_U \circ L_N^\gamma)(\omega, r). \tag{C.7}$$

To begin, note that (3.19) implies that

$$L^\gamma(\omega, r) = (\pi_\Omega(\omega, u), \gamma(\omega, r)). \tag{C.8}$$

By definition, π_Ω and π_U are, respectively, $\mathcal{B} \otimes \mathcal{U}/B$- and $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$-measurable. Likewise, γ^k, $k = 1, 2, \ldots, N$, is $\mathcal{J}^k/\mathcal{U}^k$-measurable. Accordingly, $\gamma := (\gamma^1, \gamma^2, \ldots, \gamma^N)$ is $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$-measurable (since $\mathcal{J}^k \subset \mathcal{B} \otimes \mathcal{U}$ for all k). It follows that L^γ and, by composition [4, Thm. 13.1], L_k^γ and $\pi_U \circ L_N^\gamma$ are, respectively, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B} \otimes \mathcal{U}$-, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B} \otimes \mathcal{U}$-, and $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$-measurable. But all u-sections of $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$-measurable functions are $\mathcal{B}/\mathcal{U}$-measurable [4, Thm. 18.1]; consequently, $\Sigma^\gamma = \pi_U \circ L_N^\gamma|_r$ is $\mathcal{B}/\mathcal{U}$-measurable.

(ii). *Sufficiency.* Fix $\gamma \in \Gamma$, and suppose that ψ is an order function such that γ possesses property CI*. To prove that γ is deadlock-free it suffices to show that for each $\omega \in \Omega$, the agents can be ordered such that no agent's action depends on itself, or actions of its successors.

Fix $\omega \in \Omega$. By (i), the closed-loop equation $u = \gamma(\omega, u)$ possesses a unique solution $u_\omega^\gamma \in U$. Let

$$s := (s_1, s_2, \ldots, s_N) = \psi(\omega, u_\omega^\gamma). \tag{C.9}$$

Since property CI* holds with order function ψ, for all $k = 1, 2, \ldots N$,

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma)) \subset \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma))\}. \tag{C.10}$$

But (C.10) implies that at the point $(\omega, u_\omega^\gamma) \in \mathrm{G}^\gamma$, γ^{s_k} does not depend on the s_kth, s_{k+1}th, ..., or s_Nth components of $(\omega, u_\omega^\gamma)$ (recall that $\mathcal{J}^{\gamma^{s_k}} := [\gamma^{s_k}]^{-1}(U^{s_k})$); consequently, for all $k = 1, 2, \ldots, N$, the s_kth agent's action does not depend on the actions of agents $s_k, s_{k+1}, \ldots,$ and s_N. This proves sufficiency.

Necessity. Fix $\gamma \in \Gamma$, and suppose that γ does not possess property CI* for any order function ψ. Then $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) \neq \Omega$, or there exists at least one outcome in G^γ, say (ω^*, u^*), such that for all N-agent orderings $s := (s_1, s_2, \ldots, s_N) \in S_N$, the inclusion

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega^*, u^*)) \subset \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega^*, u^*))\} \tag{C.11}$$

fails for at least one $k \in \{1, 2, \ldots, N\}$. To prove necessity, it suffices to demonstrate that γ is not deadlock-free in either case.

When $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) \neq \Omega$, for some $\omega \in \Omega$, the closed-loop equation $\gamma(\omega, u) = u$ has no solution; consequently, for that ω, γ has no implementation (let alone a deadlock-free implementation). When there exists an outcome $(\omega^*, u^*) \in \mathrm{G}^\gamma$ such that for every N-agent ordering $s \in S_N$, (C.11) fails for at least one $k \in \{1, 2, \ldots, N\}$ for fixed $s \in S_N$,

$$\begin{aligned} &\mathcal{J}^{\gamma^{s_{k^*}}} \bigcap [\mathcal{P}_{T_{k^*-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k^*-1}^N(s)}(\omega^*, u^*)) \\ &\qquad \not\subset \{\emptyset, [\mathcal{P}_{T_{k^*-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k^*-1}^N(s)}(\omega^*, u^*))\} \end{aligned} \tag{C.12}$$

for some $k^* \in \{1,2,\ldots,N\}$. But (C.12) implies that at the point (ω^*, u^*), $\gamma^{s_{k^*}}$ depends on the actions of agents that have yet to act under s; consequently, agent s_{k^*} cannot act without precognition under s. Since the same argument applies for all $s \in S_N$, γ must deadlock. This proves necessity. □

Appendix D.

Proof of Corollary 1. Although this corollary is an immediate consequence of Theorems 1(ii) and 2(ii) (property C* ⇒ property DF* ⇒ property CI*), it is instructive to prove it directly.

Fix $\gamma \in \Gamma$ and suppose that ψ is an order function such that γ possesses property C*. Since property C* ensures that $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$, it suffices to show that ψ is also an order function such that γ possesses property CI*—i.e., that (3.11) of property C* (with $s = T_k^N(\psi(\omega,u)) \in S_k$), implies (3.12) of property CI* (with $s = \psi(\omega,u) \in S_N$) for all $(\omega,u) \in \mathrm{G}^\gamma$ and $k = 1,2,\ldots,N$.

Fix $(\omega,u) \in \mathrm{G}^\gamma$ and $k \in \{1,2,\ldots,N\}$, and let

$$s := (s_1, s_2, \ldots, s_N) = \psi(\omega,u). \tag{D.1}$$

Since $T_k^N(s) \in S_k$ and $T_{k-1}^N = T_{k-1}^k \circ T_k^N$, (3.11) of property C* implies that

$$\begin{aligned} &\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}([T_k^N \circ \psi]^{-1}(T_k^N(s)))) \\ &\qquad \subset \mathcal{F}(T_{k-1}^N(s)) \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\mathrm{G}^\gamma)). \end{aligned} \tag{D.2}$$

Restricting both sides of (D.2) to

$$[\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega,u)) \tag{D.3}$$

yields the desired result—(3.12) of property CI*—if

$$\begin{aligned} &[\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}([T_k^N \circ \psi]^{-1}(T_k^N(s)))) \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega,u)) \\ &\qquad = [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega,u)) \end{aligned} \tag{D.4}$$

and

$$\begin{aligned} &\mathcal{F}(T_{k-1}^N(s)) \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\mathrm{G}^\gamma)) \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega,u)) \\ &\qquad = \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega,u))\}. \end{aligned} \tag{D.5}$$

Equation (D.5) follows from the definition of $\mathcal{F}(T_{k-1}^N(s))$,

$$\mathcal{F}(T_{k-1}^N(s)) := [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i}\right)\right), \tag{D.6}$$

the fact that inverse images preserve intersections—i.e.,

$$\begin{aligned} &[\mathcal{P}_{T_{k-1}^N(s)}]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i}\right)\right) \bigcap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega,u)) \\ &\qquad = \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega,u))\} \end{aligned} \tag{D.7}$$

—and the fact that

$$[\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\mathrm{G}^\gamma)) \bigcap [\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\omega,u)) = [\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\omega,u)) \tag{D.8}$$

since

$$[\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\omega,u)) \in [\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(\mathrm{G}^\gamma)) \tag{D.9}$$

when $(\omega,u) \in \mathrm{G}^\gamma$.

Equation (D.4) follows from the observation that

$$[\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}(w,u)) \subset [\mathcal{P}_{T^N_{k-1}(s)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s)}([T^N_k \circ \psi]^{-1}(T^N_k(s)))) \tag{D.10}$$

by (D.1). □

Appendix E.

Proof of Theorem 3. Fix $\gamma \in \Gamma$ and suppose that γ is sequential. Then there exists a *constant* order function ψ such that γ possesses property CI*. Since property CI* ensures that $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$, it suffices to show that ψ is also an order function such that γ possesses property C*—i.e., that for all $k = 1, 2, \ldots, N$, the fact that (3.12) of property CI* holds for all $(\omega,u) \in \mathrm{G}^\gamma$ with $s = s^* \in S_N$ constant implies that (3.11) of property C* holds for all $s \in S_k$.

Fix $k \in \{1, 2, \ldots, N\}$ and let

$$s^* := (s^*_1, s^*_2, \ldots, s^*_N) \tag{E.1}$$

denote the constant order induced by ψ. Since $T^N_{k-1} = T^k_{k-1} \circ T^N_k$, and since for all $s \in S_k$,

$$[\mathcal{P}_{T^k_{k-1}(s)}]^{-1}(\mathcal{P}_{T^k_{k-1}(s)}([T^N_k \circ \psi]^{-1}(s))) = \begin{cases} [\mathcal{P}_{T^N_{k-1}(s*)}]^{-1}(\mathcal{P}_{T^N_{k-1}(s*)}(\mathrm{G}^\gamma)) & \text{when } s = T^N_k(s^*), \\ \emptyset & \text{else,} \end{cases} \tag{E.2}$$

it suffices to show that

$$\mathcal{J}^{\gamma^{s^*_k}} \subset \mathcal{F}(T^N_{k-1}(s^*)) \tag{E.3}$$

for all $k = 1, 2, \ldots, N$.

By definition (§2, 1(d)), $\mathcal{J}^{\gamma^{s^*_k}} \subset \mathcal{J}^{s^*_k}$ is a subset of

$$\mathcal{B} \otimes \mathcal{U} = [\mathcal{P}_{T^N_N(s^*)}]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^N \mathcal{U}^{s^*_i}\right)\right). \tag{E.4}$$

Since (3.12) holds for all $(\omega,u) \in \mathrm{G}^\gamma$ when $s = s^*$, all events in $\mathcal{J}^{\gamma^{s^*_k}}$ must be of the form

$$[\mathcal{P}_{T^N_N(s^*)}]^{-1}\left(A \times \left(\prod_{i=k}^N U^{s^*_i}\right)\right), \tag{E.5}$$

where $A \subset \Omega \times (\Pi_{i=1}^{k-1} U^{s_i^*})$; accordingly, $\mathcal{J}^{\gamma^{s_k^*}}$ is also a subset of

$$\mathcal{C}_{s^*} := \sigma\left([\mathcal{P}_{T_N^N(s^*)}]^{-1}\left(A \times \left(\prod_{i=k}^{N} U^{s_i^*} \right) \right) : \; A \subset \Omega \times \left(\prod_{i=1}^{k-1} U^{s_1^*} \right) \right)$$

$$= \sigma\left([\mathcal{P}_{T_{k-1}^N(s^*)}]^{-1}(A) : \; A \subset \Omega \times \left(\prod_{i=1}^{k-1} U^{s_1^*} \right) \right) \tag{E.6}$$

—the cylindrical extension of the power set of $\Omega \times (\prod_{i=1}^{k-1} U^{s_i^*})$ to $\Omega \times U$. But

$$(\mathcal{B} \otimes \mathcal{U}) \bigcap \mathcal{C}_{s^*} = [\mathcal{P}_{T_{k-1}^N(s^*)}]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i^*} \right) \right)$$

$$:= \mathcal{F}(T_{k-1}^N(s^*)). \tag{E.7}$$

Consequently, $\mathcal{J}^{\gamma^{s_k^*}} \subset \mathcal{F}(T_{k-1}^N(s^*))$. □

Appendix F.

Proof of Theorem 4. Fix $\gamma \in \Gamma$ and suppose that ψ is an order function such that γ possesses property CI*. Since property CI* ensures that $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$, it suffices to show that ψ is also an order function such that γ possesses property C*—i.e., that for all $k = 1, 2, \ldots, N$, the fact that (3.12) holds for all $(\omega, u) \in \mathrm{G}^\gamma$ with order function ψ implies that (3.11) holds for all $s \in S_k$ with order function ψ.

By assumption, the σ-fields $\mathcal{B}$ and $\mathcal{U}^k$, $k = 1, 2, \ldots, N$, contain, respectively, the singletons of the countable sets Ω and U^k, $k = 1, 2, \ldots, N$ ($\mathcal{U}^k$ contains the singletons of U^k due to (§2, 1(c)). Accordingly, for all $s := (s_1, s_2, \ldots, s_k) \in S_k$, $k = 1, 2, \ldots, N$, the product field $\mathcal{B} \otimes (\bigotimes_{i=1}^{k} \mathcal{U}^{s_i})$ contains the singletons of the countable set $\Omega \times (\Pi_{i=1}^{k} U^{s_i})$, implying that $\mathcal{B} \otimes (\bigotimes_{i=1}^{k} \mathcal{U}^{s_1})$ is the power set of $\Omega \times (\Pi_{i=1}^{k} U^{s_i})$. It follows, for all $s \in S_k$, $k = 1, 2, \ldots, N$, that

$$\mathcal{F}(T_{k-1}^k(s)) := [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i} \right) \right)$$

$$= \sigma\left([\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(A) : \; A \subset \Omega \times \left(\prod_{i=1}^{k-1} U^{s_i} \right) \right) \tag{F.1}$$

—i.e., that $\mathcal{F}(T_{k-1}^k(s))$ is the cylindrical extension of the power set of $\Omega \times (\Pi_{i=1}^{k-1} U^{s_i})$ to $\Omega \times U$.

Fix $k \in \{1, 2, \ldots, N\}$ and $s \in S_k$. Since property CI* holds with order function ψ, (3.12) and (F.1) imply that for all $(\omega, u) \in [T_k^N \circ \psi]^{-1}(s)$ and $A \in \mathcal{J}^{\gamma^{s_k}}$,

$$A \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\omega, u)) \in \{\emptyset, [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\omega, u))\}$$

$$\subset \mathcal{F}(T_{k-1}^k(s)). \tag{F.2}$$

Since $[T_k^N \circ \psi]^{-1}(s) \in \mathrm{G}^\gamma$ is a countable set, and since inverse and direct images preserve unions, it follows by (F.2) that

$$A \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s)))$$

$$= A \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}\left(\mathcal{P}_{T_{k-1}^k(s)}\left(\bigcup_{(\omega,u)\in[T_k^N\circ\psi]^{-1}(s)}(\omega,u)\right)\right)$$

$$= \bigcup_{(\omega,u)\in[T_k^N\circ\psi]^{-1}(s)} (A \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\omega,u))) \tag{F.3}$$

$$\in \mathcal{F}(T_{k-1}^k(s)).$$

But because $[T_k^N \circ \psi]^{-1}(s) \subset \mathrm{G}^\gamma$,

$$[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \subset [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\mathrm{G}^\gamma)); \tag{F.4}$$

hence (F.3) implies that

$$A \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s)))$$

$$\in \mathcal{F}(T_{k-1}^k(s)) \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\mathrm{G}^\gamma)). \tag{F.5}$$

Since (F.5) holds for all $A \in \mathcal{J}^{\gamma^{s_k}}$,

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s)))$$

$$\subset \mathcal{F}(T_{k-1}^k(s)) \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\mathrm{G}^\gamma)). \tag{F.6}$$

This proves the theorem. □

Appendix G.

Proof of Theorem 5. (i) and (ii). Properties S and SM are, by definition, specializations of properties S* and SM* to all $\gamma \in \Gamma$ (cf. Definitions 1 and 2, and [9, §4, Definitions]). Accordingly, all $\gamma \in \Gamma$ possess property S* (respectively SM*) if and only if $\mathcal{I}$ possesses property S (respectively SM).

(iii). By Theorem 1(ii) of [2], $\mathcal{I}$'s possession of property CI is a necessary and sufficient condition for all $\gamma \in \Gamma$ to possess property DF*. By Theorem 2(ii), $\gamma \in \Gamma$ possesses property DF* if and only if γ possesses property CI*. Accordingly, all $\gamma \in \Gamma$ possess property CI* if and only if $\mathcal{I}$ possesses property CI.

(iv). To prove that all $\gamma \in \Gamma$ possess property C* when $\mathcal{I}$ possesses property C, let $\psi : \Omega \times U \to S_N$ be an order function for which $\mathcal{I}$ possesses property C, fix $\gamma \in \Gamma$, and let ψ^γ denote the restriction of ψ to G^γ. Since γ induces a unique $\mathcal{B}/\mathcal{U}$-measurable mapping $\Sigma^\gamma : \Omega \to U$ with graph G^γ [9, Thm. 1], $\mathcal{P}_\emptyset(\mathrm{G}^\gamma) = \Omega$; accordingly, to establish that γ possesses property C*, it suffices to show that (3.11) of property C* holds with order function ψ^γ for all $s := (s_1, s_2, \ldots, s_k) \in S_k$ and $k = 1, 2, \ldots, N$.

Since $\Omega \times U \in \mathcal{J}^{s_k}$, property C [9, Lem. 1] implies that

$$[T_k^N \circ \psi]^{-1}(s) \in \mathcal{F}(T_{k-1}^k(s)) := [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}\left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i}\right)\right). \tag{G.1}$$

Consequently,

$$[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) = [T_k^N \circ \psi]^{-1}(s). \tag{G.2}$$

Since $\mathcal{J}^{\gamma^{s_k}} \subset \mathcal{J}^{s_k}$, property C [9] also implies that

$$\mathcal{J}^{\gamma^{s_k}} \bigcap [T_k^N \circ \psi]^{-1}(s) \subset \mathcal{F}(T_{k-1}^k(s)). \tag{G.3}$$

Substitute (G.2) into (G.3), and restrict both sides of the result to

$$[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\mathrm{G}^\gamma)). \tag{G.4}$$

The desired result—(3.11) of property C*—follows if

$$\begin{aligned}&[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \bigcap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\mathrm{G}^\gamma))\\ &\quad = [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi^\gamma]^{-1}(s))).\end{aligned} \tag{G.5}$$

To verify (G.5), note that

$$\begin{aligned}[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\mathrm{G}^\gamma)) &= \bigcup_{(\omega,u)\in \mathrm{G}^\gamma} [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\omega,u))\\ &= \bigcup_{s'\in S_k} [\mathcal{P}_{T_{k-1}^k(s')}]^{-1}(\mathcal{P}_{T_{k-1}^k(s')}([T_k^N \circ \psi^\gamma]^{-1}(s'))).\end{aligned} \tag{G.6}$$

Since ψ^γ is the restriction of ψ to G^γ, and since direct and inverse images preserve inclusions,

$$\begin{aligned}&[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi^\gamma]^{-1}(s)))\\ &\quad \subset [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))).\end{aligned} \tag{G.7}$$

But

$$\begin{aligned}&\{[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) : s \in S_k\}\\ &\quad = \{[T_k^N \circ \psi]^{-1}(s) : s \in S_k\}\end{aligned} \tag{G.8}$$

partitions $\Omega \times U$ (cf. (G.2)). Hence, by (G.7), the restriction of (G.6) to

$$[\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s))) \tag{G.9}$$

is (G.5), and thus γ possesses property C*. □

Acknowledgments. The authors thank the referees and editors for their helpful comments.

REFERENCES

[1] M. S. Andersland, *Decoupling nonsequential stochastic control problems*, Systems Control Lett., 16 (1991), pp. 65–69.

[2] M. S. Andersland and D. Teneketzis, *Information structures, causality, and nonsequential stochastic control* I: *design-independent properties*, SIAM J. Control Optim., 30 (1992), pp. 1447–1475.

[3] Z. Banaszak and B. H. Krogh, *Deadlock avoidance in flexible manufacturing systems with concurrently competing process flows*, IEEE Trans. Robotics Automation, RA-6 (1990), pp. 1434–1449.

[4] P. Billingsley, *Probability and Measure*, 2nd ed., John Wiley, New York, 1986.

[5] E. Chen and S. Lafortune, *Dealing with blocking in supervisory control of discrete-event systems*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 724–735.

[6] W. S. Lai, *Protocol traps in computer networks–a catalog*, IEEE Trans. Comm., COM-30 (1982), pp. 1434–1449.

[7] Y. Li and W. M. Wonham, *Deadlock issues in supervisory control of discrete-event systems*, in Proc. 22nd Conf. on Infor. Sci. and Syst., Princeton, NJ, March 1988, pp. 57–63.

[8] T. MURATA, *Petri nets: properties, analysis and applications*, Proc. of the IEEE, 77 (1989), pp. 541–580.

[9] H. S. WITSENHAUSEN, *On information structures, feedback and causality*, SIAM J. Control, 9 (1971), pp. 149–160.

[10] ———, *A standard form for sequential stochastic control*, Math. Systems Theory, 7 (1973), pp. 5–11.

[11] ———, *The intrinsic model for discrete stochastic control: some open problems*, Lecture Notes in Econ. and Math. Sys., Vol. 107, Springer-Verlag, Berlin, 1975, pp. 322–335.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1752–1762, November 1994

013

ON CONVERGENCE OF ITERATED RANDOM MAPS*

JOHN R. LIUKKONEN† AND ARNOLD LEVINE†

Abstract. Let K be a compact subset of $\mathbf{R}^N$ consisting of an open subset of $\mathbf{R}^N$ and smooth boundary. Many stochastic optimization algorithms can be viewed as iterations of independent and identically distributed random elements of the continuous self-maps of such a K. In case the target of the algorithm is a single point $\mathbf{p}$, a near dichotomy for the convergence to $\mathbf{p}$ is given; roughly, if $\mathbf{M}_1$ is one of the random elements, then there will be no convergence in probability if $E(\log \|\mathbf{M}_1'(\mathbf{p})\|) > 0$, but there will be almost sure convergence if $E(\log \|M_1'(\mathbf{p})\|) < 0$ and an accessibility condition holds. Similar conclusions are reached if the algorithm has an entire closed set of target points. Illustrative applications of these algorithms to the analysis of algorithms are given, and stochastic order of convergence and expected number of steps to stopping are discussed.

Key words. stochastic optimization, iterated random maps, almost sure convergence, convergence in probability

AMS subject classifications. 93E20, 93E23, 60J20, 62L20

1. Introduction.

Numerical optimization or root finding for functions $g : \mathbf{R}^N \to \mathbf{R}^m$ is accomplished commonly through the following iterative algorithm:

$$x_n = M_n(x_{n-1}, \ldots, x_1), \qquad n = 1, 2, \ldots, \tag{1.1}$$

where $\{x_n\}, \{M_n\}$ are a sequence of solution estimates and operators on the estimates. Major problems faced by such algorithms include unacceptably slow convergence or failure to converge at all. Adding noise in a controlled fashion to such algorithms can yield solutions to problems intractable by the classical deterministic methods. Much work has been published on simulated annealing (see Hajek [3], most notably), but the technique of randomizing classical algorithms in Euclidean space is also available. In Joseph, Levine, and Liukkonen [5] and [6], for example, it is shown that by adding noise in a controlled fashion to Newton–Raphson we can obtain convergence in probability to a minimum or root for a wider class of functions and initial guesses than that for which usual deterministic Newton–Raphson works.

In this paper we develop general conditions for almost sure convergence for a wide class of such randomized algorithms. We also develop companion conditions for instances when these algorithms will not converge even in probability. We also discuss the stochastic order of convergence and the expected time to solution. We restrict attention to time-homogeneous Markov algorithms: the random solution estimate x_{n+1} is to depend solely on the previous estimate x_n and in particular not on the time variable n. Our rationale for this is the following: For some types of stochastic optimization, such as simulated annealing, one tacitly assumes that local minima are easy to come by but one does not know when one has arrived at the global minimum; the fear is that one will settle prematurely for a local minimum that, on a global scale, is in fact not very good. One guards against this by requiring the noise level to decrease slowly enough with time. By contrast, in other situations there may be few local minima and one may know immediately when one has arrived at a correct answer,

*Received by the editors October 7, 1992; accepted for publication (in revised form) May 21, 1993.

†Department of Mathematics, Tulane University, New Orleans, Louisiana 70118.

but there may be difficulty even getting near a local minimum through traditional means. To deal with these situations, we want the random part of the algorithm to be prominent during the first phase, during which we search through an uninformative wilderness for a neighborhood of a minimum, and during which we must avoid traps and cycles. But we also want the random part of the algorithm to get out of the way when we finally arrive in a region where the traditional algorithm is effective. For algorithms aimed at these latter situations we tie the noise level to current performance and not to time.

In this introductory section we set out our mathematical framework. In §2 we give conditions for almost sure convergence and conditions for the failure even of convergence in probability of our random algorithms. Section 3 is devoted to examples illustrating how these results can be used in the analysis of a random algorithm, and in §4 we take up the issues of order of convergence and expected stopping time for algorithms that meet our convergence criteria.

A simple motivating example for our mathematical setup is the following: Our starting value x_0 is given, and then for each n we define $X_{n+1} = f(x_n) + e_{n+1}(x_n)$, where the nth state x_n is a realization of the random variable X_n, f is a deterministic function, and $e_1(x_0), e_2(x_1), \ldots$, is the sequence of controlled noise terms. Each $e_{n+1}(x_n)$ depends continuously on the previous state. Moreover, as function-valued random variables the noise terms $e_1(\cdot), e_2(\cdot), \ldots$, are to be independent and identically distributed. For example the sequence $\{e_n(x_{n-1})\}$ might be obtained from a sequence $\{V_n\}$ of independent and identically distributed random variables by a continuous scale funtion ϕ : for each n, $e_{n+1}(x_n) = \phi(x_n)V_{n+1}$. In any case, our basic object of study is the map-valued random variable M given by $M(t) = f(t) + e(t)$.

In general we assume that we have a compact set K in $\mathbf{R}^N$ consisting of an open subset of $\mathbf{R}^N$ and a smooth boundary. We let $T \subset K$ be the set of target points of the algorithm; we assume T is closed. We consider the space $C(K, K)$ of continuous maps $M : K \to K$, equipped with the uniform topology and the resulting Borel structure. Let $C_T(K, K)$ denote the closed subspace of $C(K, K)$ consisting of those maps leaving all target points fixed. Letting $M_1 \circ M_2$ denote the composition of the maps M_1 and M_2 we have that the maps $(M_1, M_2) \to M_1 \circ M_2 : C_T(K, K) \times C_T(K, K) \to C_T(K, K)$ and $(M, \mathbf{t}) \to M(\mathbf{t}) : C_T(K, K) \times K \to K$ are continuous.

Any probability measure P on $C_T(K, K)$ defines a random element of $C_T(K, K)$. If we have two independent random elements M_1, M_2 of $C_T(K, K)$ defined, say, by the probability measures P_1, P_2, respectively, then their composition may be defined as a random element of $C_T(K, K)$ based on $(C_T(K, K) \times C_T(K, K), P_1 \times P_2)$. The probability measure on $C_T(K, K)$ giving the distribution of this composition is the convolution of probability measures on the topological semigroup (under composition) $C_T(K, K)$. We can similarly obtain the random element $M(k)$ of K, given independent random elements $M \in C_T(K, K)$ and $k \in K$, and a probability measure on K serving as the distribution of $M(k)$.

DEFINITION 1.1. *Let $\mathbf{p} \in T$ and let M be a random element of $C_T(K, K)$. We say $M'(\mathbf{p})$ exists in probability if there is a random $N \times N$ matrix $M'(\mathbf{p})$ such that*

$$M(\mathbf{t}) = \mathbf{p} + M'(\mathbf{p})(\mathbf{t} - \mathbf{p}) + o_P(\mathbf{t} - \mathbf{p}). \tag{1.2}$$

We say that $M'(\mathbf{p})$ exists in power if for some $\beta > 0$ and some random $N \times N$ matrix $M'(\mathbf{p})$ we have $\|M'(\mathbf{p})\| \in L^\beta$ and

$$M(\mathbf{t}) = \mathbf{p} + M'(\mathbf{p})(\mathbf{t} - \mathbf{p}) + o_\beta(\mathbf{t} - \mathbf{p}). \tag{1.3}$$

Here $o_\beta(\mathbf{t}-\mathbf{p})$ *denotes a random term such that* $E(\|o_\beta(\mathbf{t}-\mathbf{p})\|^\beta/\|\mathbf{t}-\mathbf{p}\|^\beta) \to 0$ *as* $\mathbf{t} \to \mathbf{p}$. *Note that if* (1.3) *holds for* $\beta > 0$ *and if* $0 < \beta' < \beta$, *then* (1.3) *also holds for* β'.

Given a problem with a single target point $\mathbf{p}$, an iterative algorithm based on a sequence of independent copies of M, and reasonable accessibility to neighborhoods of $\mathbf{p}$ we show that convergence to $\mathbf{p}$ depends essentially on whether $E(\log\|M'(\mathbf{p})\|) < 0$ or > 0. For problems with multiple target points we establish similar conditions guaranteeing convergence to the target set as a whole and further conditions guaranteeing convergence to a single unspecified point in the target set.

2. Convergence theorems. In this section we discuss several kinds of convergence, and we must distinguish among them carefully. We are given an independent and identically distributed sequence $\{M_n\}$ of random elements of $C_T(K,K)$. For each initial distribution ν on K we write the corresponding Markov random variables $X_n(\nu) = M_n \circ \cdots \circ M_1(\nu)$; letting $\delta_\mathbf{t}$ be point mass at $\mathbf{t}$ we write $X_n(\mathbf{t})$ for $X_n(\delta_\mathbf{t})$. For any closed subset S of T and $\mathbf{t} \in K$ we let $d_S(\mathbf{t})$ be the distance from $\mathbf{t}$ to S; we write $d(\mathbf{t})$ for $d_T(\mathbf{t})$. We say $X_n(\mathbf{t})$ converges in probability to S if $d_S(X_n(\mathbf{t})) \to 0$ in probability. We say $X_n(\mathbf{t})$ converges almost surely to S if $d_S(X_n(\mathbf{t})) \to 0$ almost surely. This leaves open the possibility of an oscillatory approach to S, and so we need two further definitions. We say $X_n(\mathbf{t})$ converges almost surely to some element $\mathbf{p} \in S$ if $X_n(\mathbf{t}) \to \mathbf{p}$ almost surely. We say $X_n(\mathbf{t})$ almost surely has a limit in S if $P(\exists\, \mathbf{p} \in S : X_n(\mathbf{t}) \to \mathbf{p}) = 1$.

We first develop a simple condition guaranteeing that the algorithm will not converge in probability to a given subset $S \subset T$. We place a mild restriction on our random map M : we assume that $\log[d_S(M(\mathbf{t}))/d_S(\mathbf{t})] \geq -Y$ for all $\mathbf{t}$, for some integrable nonnegative random variable Y. Then we say that the (scalar-valued) random variable L is a lower derivate in probability for M at S if

$$d_S(M(\mathbf{t})) \geq L d_S(\mathbf{t}) + o_P(d_S(\mathbf{t})) \quad \text{as } d_S(\mathbf{t}) \to 0. \tag{2.1}$$

Now we assume that we have an independent and identically distributed sequence $\{L_n\}$ of such lower derivates. To be precise we view $\{M_n\}$ as coordinate maps of the countable product of copies of a probability space $(C(K,K),\mu_0)$, and for each n, L_n is based on the same factor as M_n. Our first basic question is when will these sequences fail to converge to S in probability for all starting values $\mathbf{t}$?

THEOREM 2.1 (Negative convergence theorem). *Assume we have an independent and identically distributed sequence* $\{M_n\}$ *of random elements of* $C_T(K,K)$ *with independent and identically distributed lower derivates* $\{L_n\}$ *in probability at* S. *Suppose*

(i) $E(\log(L_1)) > 0$, *and*

(ii) *there is a nonnegative random variable* Y *with* $E(Y) < \infty$ *and* $\log[d_S(M_1(\mathbf{t}))/d_S(\mathbf{t})] \geq -Y$ *for all* $\mathbf{t}$.

Then for any $\mathbf{t} \notin S$, $X_n(\mathbf{t})$ *does not converge to* S *in probability.*

Proof. By (ii) we have for each $\mathbf{t}$,

$$E[\log d_S(M_n(\mathbf{t})) - \log d_S(\mathbf{t})] \geq -E(Y) > -\infty. \tag{2.2}$$

It follows that $E(\log d_S(X_{n+1})) - E(\log d_S(X_n)) > -\infty$ for all n, so that $E(\log d_S(X_n)) > -\infty$ for all n. However, also by condition (ii) and the convergence in probability version of Fatou's lemma

$$\liminf_{\mathbf{t}\to\mathbf{p}} E(\log d_S(M_n(\mathbf{t}))) - \log d(\mathbf{t}) \geq E(\log(L_1)). \tag{2.3}$$

By (i), this last expected value is positive, and so for $d_S(\mathbf{t})$ small we have $E[\log d_S(M_n(\mathbf{t}))] - \log d_S(\mathbf{t}) > a$ for some $a > 0$. It follows that for some $r_0 > 0$, we have

$$E(\log d_S(X_{n+1}) - \log d_S(X_n) \mid d_S(X_n) \leq r_0) > a \tag{2.4}$$

for every n. Now

(2.5)

$$\begin{aligned} E(\log d_S(X_{n+1}) - \log d_S(X_n)) =& E([\log d_S(X_{n+1}) - \log d_S(X_n)]I_{[d_S(X_n)\leq r_0]}) \\ &+ E([\log d_S(X_{n+1}) - \log d_S(X_n)]I_{[d_S(X_n)>r_0]}). \end{aligned}$$

By (ii), the negative part of $\log d_S(X_{n+1}) - \log d_S(X_n)$ is uniformly integrable. Therefore if $P(d_S(X_n) > r_0) \to 0$, then the negative part of the second summand in (2.5) also tends to 0, and we have all together

$$E(\log d_S(X_{n+1}) - \log d_S(X_n)) > aP(d_S(X_n) \leq r_0) - \psi(P(d_S(X_n) > r_0)), \tag{2.6}$$

where ψ is a function tending to 0 at 0.

Now assume that X_n converges to S in probability. Then $P(d_S(X_n) \leq r_0) \to 1$ and we have from (2.6) that $E(\log d_S(X_n))$ cannot converge to $-\infty$. But if $X_n \to S$ in probability we must have $E(\log d_S(X_n)) \to -\infty$. □

We now establish sufficient conditions for the almost sure convergence of our algorithm. To develop our conditions we impose a different mild restriction on the random map M: We assume $d(M(\mathbf{t}))/d(\mathbf{t}) \leq Y$ for all $\mathbf{t}$, where Y is some scalar-valued random variable whose βth moment exists for some $\beta > 0$. Note that this implies that $E(d(M(\mathbf{t}))^\beta) < \infty$ for each $\mathbf{t}$.

DEFINITION 2.2. *Let M be a random element of $C_T(K, K)$, let $S \subset T$ be a closed subset, and let U be a nonnegative scalar-valued random variable. We say that U is an upper derivate at S for M in power if for some $\beta > 0$,*

(i) $E(U^\beta) < \infty$ *and*

(ii) $d_S(M(\mathbf{t})) \leq U d_S(\mathbf{t}) + o_\beta(d_S(\mathbf{t}))$.

Observe that an upper derivate in power is also an upper derivate in probability.

PROPOSITION 2.3. *Suppose we have q independent random elements $M_1, \ldots, M_q$ in $C_T(K, K)$ with independent and identically distributed upper derivates $\{U_j\}$ in power at S. Then the composition $M_q \circ \cdots \circ M_1$ has upper derivate $U_q \cdot \cdots \cdot U_1$ in power at S.*

Proof. This is a straightforward real analysis argument using the Hölder inequality and the fact that for $0 < \beta \leq 1$ and nonnegative random variables Y_1, Y_2, $E([Y_1 + Y_2]^\beta) \leq E(Y_1^\beta) + E(Y_2^\beta)$. □

PROPOSITION 2.4. *Let M be a random element of $C_T(K, K)$ with upper derivate U in power at S. Suppose $E(\log U) < 0$. Then for some sufficiently small $\beta > 0$, some $r_0 > 0$, and some $C < 1$, we have*

$$E(d_S(M(\mathbf{t}))^\beta) \leq C d_S(\mathbf{t})^\beta \tag{2.7}$$

whenever $d_S(\mathbf{t}) \leq r_0$.

Proof. Let us observe that

$$\limsup_{d(\mathbf{t})\to 0} E(d(M(\mathbf{t}))^\beta)/d(\mathbf{t})^\beta \leq E(U^\beta) < \infty \tag{2.8}$$

for sufficiently small $\beta > 0$. Choose $B > 0$ such that $E(\log U\, I_{[-B \le \log U]}) < 0$. Then by the dominated convergence theorem we may differentiate $g_B(\beta) = E(U^\beta I_{[-B \le \log U]})$ through the expectation to obtain $g'_B(0) = E(\log U\, I_{[-B \le \log U]}) < 0$. It follows that for small positive β we have $g_B(\beta) < g_B(0) = P(-B \le \log U)$. Since $E(U^\beta I_{[-B > \log U]}) \le e^{-\beta B} P(-B > \log U)$, we obtain $E(U^\beta) < 1$ for $\beta > 0$ sufficiently small. □

THEOREM 2.5 (Positive convergence theorem). *Suppose given a sequence $\{M_n\}$ of independent and identically distributed random elements of $C_T(K, K)$ with independent and identically distributed upper derivates $\{U_n\}$ in power at T. Suppose there is a nonnegative random variable Y such that $d(M_1(\mathbf{t}))/d(\mathbf{t}) \le Y$ for all $\mathbf{t}$ and such that for some $\beta > 0$ the β moment of Y exists. Suppose also that*

(i) *for every $\mathbf{t} \in K$ and every neighborhood V of T there is an n such that $P(X_n(\mathbf{t}) \in V) > 0$, and*

(ii) $E(\log U_1) < 0$.

Then for any Markov chain generated by the sequence $\{M_n\}$, the invariant probability measures are precisely the probability measures supported on T. Moreover, for each $\mathbf{t} \in K$, $X_n(\mathbf{t})) \to T$ almost surely.

Proof. Let $PM(K)$ denote the probability measures on K. Using the common probability distribution μ_0 of $M_1, M_2, \ldots$, on $C_T(K, K)$, the Markov transition $L : PM(K) \to PM(K)$ may be defined by

$$L(\mu)(f) = \int \int f(M(\mathbf{t}))\, d\mu(\mathbf{t})\, d\mu_0(M) \tag{2.9}$$

for all $f \in C(K, \mathbf{R})$. Then L is continuous in the weak topology, and it is standard (see [8, Chap. 6]) that given any $\mu \in PM(K)$, a subsequence of $\{\Sigma_{k=1}^n L^n(\mu)/n\}$ converges weakly to an invariant probability measure ν on K. We claim that such a ν must be supported on T.

To prove the claim, consider a Markov chain $\{Z_n\}$ on K with initial distribution ν and defined by iterates of independent copies of M. Then every Z_n has distribution ν. By Proposition 2.4 we get $\beta > 0, r_0 > 0$, and $C < 1$ such that

$$E(d(M(\mathbf{t}))^\beta) \le C d(\mathbf{t})^\beta \tag{2.10}$$

whenever $d(\mathbf{t}) \le r_0$. For every n let $Y_n = d(Z_n)^\beta \wedge r_0^\beta$. Observe that the $\{Y_n\}$ are identically distributed and $P(Y_{n+1} = 0 \mid Y_n = 0) = 1$. From (2.10) we get $P(0 < Y_n < r_0^\beta) = 0$ for all n. Assume that $P(Y_n = r_0^\beta) > 0$. Then using assumption (i) we can find a positive integer k such that $P(Y_{n+k} < r_0^\beta \mid Y_n = r_0^\beta) > 0$. But then we get $P(Y_{n+k} < r_0^\beta) > P(Y_n < r_0^\beta)$, which is impossible. Thus $P(Y_n = 0) = 1$ for all n, and ν must be supported on T. The claim is proved.

Fix $\mathbf{t} \in K$, and let $X_n = X_n(\mathbf{t})$ for all n. We can now say that given $\epsilon > 0$ and $r_1 > 0$ there is an arbitrarily large n_1 with $(1/n_1)\Sigma_{k=1}^{n_1} P(d(X_k) \le r_1) > 1 - \epsilon$. Hence for at least one $n_0, P(d(X_{n_0}) \le r_1) > 1 - \epsilon$. Now choose $r_0 > 0$ and an upper derivate W for M at T in power such that $E(\log W) < 0$ and $d(M(\mathbf{t}))/d(\mathbf{t}) \le W$ whenever $d(\mathbf{t}) \le r_0$. Let $S_n(r_1)$ denote the random walk starting at $\log r_1$ with step $\log W$. Since $E(\log W) < 0$, we can show by [1, Thm. 8.3.4] that if we choose r_1 small enough, then $P(S_n(r_1) < \log r_0$ for all $n) > 1 - \epsilon$. Comparing $\log X_n$ stochastically with $S_n(r_1)$ for all n, we get $P(d(X_n) < r_0$ for all $n \ge n_0 \mid d(X_{n_0}) < r_1) > 1 - \epsilon$. Moreover, given $d(X_n) < r_0$ for all $n \ge n_0$, we get $E(d(X_{n+1})^\beta) < CE(d(X_n)^\beta)$ for all $n \ge n_0$, for some $C < 1$. A standard application of the Borel–Cantelli lemma (see for example [1, §4.2]) shows that given $d(X_n) \le r_0$ for all $n \ge n_0$, $d(X_n) \to 0$ almost surely. So given

an event with probability $> 1 - 2\epsilon$, the conditional probability that $d(X_n) \to 0$ is 1. Since $\epsilon > 0$ is arbitrary, $d(X_n) \to 0$ almost surely. □

COROLLARY 2.6. *Assume the hypotheses of Theorem* 2.5, *and in addition that there is a random variable* V *with*

$$\limsup_{d(\mathbf{t})\to 0} \frac{\|\mathbf{t} - M(\mathbf{t})\|}{d(\mathbf{t})} \leq V \tag{2.11}$$

in power and $E(V^\beta) < \infty$ *for some* $\beta > 0$. *Then for each* $\mathbf{t} \in K, X_n(\mathbf{t})$ *almost surely has a limit in* T.

Proof. The hypotheses of this corollary imply that for some $\beta > 0$ and $r_0 > 0$ sufficiently small,

$$E\left(\frac{\|\mathbf{t} - M(\mathbf{t})\|^\beta}{d(\mathbf{t})^\beta}\right) \leq D, \tag{2.12}$$

whenever $d(\mathbf{t}) \leq r_0$, for some finite D. From Theorem 2.5, $d(X_n) \leq r_0$ eventually, with probability 1. Moreover, for r_0 small enough, conditioned on $d(X_n) \leq r_0$, we have $E(d(X_{m+1})^\beta) \leq CE(d(X_m)^\beta)$ for all $m \geq n$, for some $C < 1$. Thus with probability 1, $\Sigma_{n=1}^\infty E(d(X_n)^\beta) < \infty$. But then also with probability 1,

$$E(\Sigma_{n=1}^\infty \|X_{n-1} - X_n\|^\beta) = \Sigma_{n=1}^\infty E(\|X_{n-1} - X_n\|^\beta) < \infty \tag{2.13}$$

by comparison with $\Sigma E(d(X_n)^\beta)$. Thus with probability 1, $\Sigma_{n=1}^\infty \|X_{n-1} - X_n\|^\beta < \infty$, and X_n converges to some $\mathbf{p} \in T$. □

COROLLARY 2.7. *Assume condition* (i) *of Theorem* 2.5, *and in addition that* $T = \{\mathbf{p_1}, \ldots, \mathbf{p_m}\}$ *is finite. Assume further that for each* $i = 1, \ldots, m$ *there is an upper derivate* U_i *at* p_i *for* M *in power, with* $E(\log U_i) < 0$. *Then for each* $\mathbf{t} \in K, X_n(\mathbf{t})$ *almost surely has a limit in* T.

Proof. We sketch a proof parallel to that of Theorem 2.5. Let $d_i(\mathbf{t}) = \|\mathbf{t} - \mathbf{p_i}\|$ for each $i = 1, \ldots, m$. By the hypotheses on the U_i's and Proposition 2.4 we get some $\beta > 0, r_0 > 0$, and $C < 1$ such that $E(d_i(M(\mathbf{t}))^\beta) \leq Cd_i(\mathbf{t})^\beta$ whenever $d_i(\mathbf{t}) \leq r_0$, for each i. Reduce r_0 if necessary so that $r_0 < \|p_i - p_j\|/2$ whenever $i \neq j$. Then for this same β, r_0, and C, (2.10) holds whenever $d(\mathbf{t}) \leq r_0$. We check the proof of Theorem 2.5 to see that the main steps go through. First, from condition (i) of Theorem 2.5 and (2.10), we have that the invariant measures are supported on T. Next we see that given $\epsilon > 0$ there is with probability at least $1 - \epsilon$ an n_0 and an i such that $d_i(X_{n_0}) < s$, where $s > 0$ is chosen so that for each i, the random walk starting at $\log s$ with step $\log W_i$ never exceeds $\log r_0$ except with probability $< \epsilon$. Given a fixed i and the event $d_i(X_n) \leq r_0$ for all $n \geq n_0$ for some n_0, we apply the fact that $d_i(X_{n+1})^\beta \leq Cd_i(X_n)^\beta$ for all $n \geq n_0$ to conclude via the Borel–Cantelli lemma that given this event, $X_n \to p_i$ almost surely. All together we have that given an event of probability $> 1 - 2\epsilon$, the conditional probability is 1 that for some $i, X_n \to p_i$. Since $\epsilon > 0$ is arbitrary this proves the corollary. □

3. Illustrative applications. In this section we illustrate the application of our theorems to the analysis of algorithms. The first example is patterned after the motivating example in the Introduction. It assumes complete information about the algorithm. In that sense it is not a realistic example of the analysis of a stochastic optimization algorithm, but it is a good introduction to the use of our theorems. The second example assumes only information typically available and so is a true-to-life analysis of a stochastic optimization algorithm.

Example A. For our first illustrative example let K be the closed unit ball in $\mathbf{R}^N$. Consider a fixed $f \in C(K, K)$ such that $f(\mathbf{0}) = \mathbf{0}$ and $f(\mathbf{t}) = o(\mathbf{t})$ as $\mathbf{t} \to \mathbf{0}$. Let X be an $\mathbf{R}^N$-valued random variable with bounded density and some finite β moment. Define the random map M by $M(\mathbf{t}) = f(\mathbf{t}) + \phi(\|\mathbf{t}\|)X$, where $\phi : [0,1] \to [0,1]$ is a continuous scale factor with sole root 0 and whose right-hand derivative at 0 exists (we allow $+\infty$ as a possible value for this derivative). Of course this definition must be adjusted to cope with the possibility that for some $\mathbf{t}$, $M(\mathbf{t})$ lands outside K with positive probability. Two common remedies are (i) reset (we project back onto the nearest point of K) and (ii) restart (we take $M(\mathbf{t})$ to be a randomly chosen point of K). We will adopt remedy (i) (reset).

From these assumptions we can verify the following: (a) the nonnegative integrable Y required for the positive convergence theorem exists provided $\phi'(0) < \infty$; (b) condition (ii) of the negative convergence theorem holds provided $\phi'(0) > 0$. A simple way to guarantee condition (i) of the positive convergence theorem is to require that the support of the density of X be all of $\mathbf{R}^N$; however condition (i) of that theorem might easily hold without this requirement.

Now $M'(\mathbf{0}) = \phi'(0)X$, in any sense that we choose. In case $\phi'(0) = +\infty$ we have a lower derivate in the sense we choose. To apply the positive and negative convergence theorems above we observe that

$$E(\log \|M'(0)\|) = \log \phi'(0) + E(\log \|X\|), \tag{3.1}$$

so that for a sequence of random variables generated by iterating independent copies of the map M, we obtain convergence almost surely starting from any nonzero element of K if $\phi'(0) < \exp\{-E(\log \|X\|)\}$, and we fail to have convergence in probability starting from any nonzero element of K if this last inequality is reversed. By Jensen's inequality, $E(\log \|X\|) < \log E(\|X\|)$, so it is enough to have $\phi'(0) < E(\|X\|)^{-1}$ in order to have convergence.

When the distribution of X is known we can say more. For example, if X is uniform on the ball of radius R centered at the origin in $\mathbf{R}^N$ we get $E(\log \|X\|) = \log R - 1/N$, with convergence when $\phi'(0) < \exp(1/N)/R$. In this case the accessibility condition is fulfilled provided that for some $C < 1$, $\|f(\mathbf{t}) + \phi(\|\mathbf{t}\|)X\|/\|\mathbf{t}\| \le C$ with positive probability for each $\mathbf{t} \in K$, and this holds if and only if

$$\sup_{\mathbf{t}} \frac{\|f(\mathbf{t})\|}{\|\mathbf{t}\|} - R\frac{\phi(\|\mathbf{t}\|)}{\|\mathbf{t}\|} < 1. \tag{3.2}$$

If on the other hand X is multivariate normal with mean $\vec{0}$ and covariance matrix I, then accessibility is automatic and $E(\log \|X\|) = \frac{1}{2}E(\log \|X\|^2)$. Now $\|X\|^2$ has the chi-squared distribution with N degrees of freedom and so $E(\log \|X\|^2) = \log 2 - \gamma + [1 + \frac{1}{2} + \cdots + 1/(\frac{N}{2} - 1)]$ for N even and $= -\gamma - \log 2 + 2 + 2/3 + \cdots + 2/(N-2)$ for N odd. (See [4, Chap. 16].)

Let us extend this example. Suppose we have a closed submanifold T of the unit ball K, and $h \in C(K, K)$ such that $h(\mathbf{t}) = \mathbf{t}$ for all $\mathbf{t} \in T$ and $d(h(\mathbf{t})) = o(d(\mathbf{t}))$ as $d(\mathbf{t}) \to 0$. Suppose further that (as might happen in a descent algorithm) $\|h(\mathbf{t}) - \mathbf{t}\| \le Rd(\mathbf{t})$ for all $\mathbf{t} \in K$, for some constant R. Let $M(\mathbf{t}) = h(\mathbf{t}) + [\phi(d(\mathbf{t})) + o(d(\mathbf{t}))]X$, where $\phi : [0,1] \to [0,1]$ is as above, X is an $\mathbf{R}^N$-valued random variable with distribution to be assigned, and we adopt the reset remedy as needed. Then (2.11) of Corollary 2.6 holds, and

$$\lim_{d(\mathbf{t}) \to 0} \frac{d(M(\mathbf{t}))}{d(\mathbf{t})} = \phi'(0)X. \tag{3.3}$$

Our discussion above carries over, and for the algorithm generated by independent copies of M, from Corollary 2.6 we almost surely have a limit in T if $\log \phi'(0) + E(\log \|X\|) < 0$, and from Theorem 2.1 we fail to have convergence in probability to T if $\log \phi'(0) + E(\log \|X\|) > 0$.

Example B. For our second example let us suppose that an optimization problem leads us to seek roots of a continuously differentiable function g on $[-1, 1]$, and that we decide to devise a randomized version of Newton–Raphson. Suppose that g has finitely many roots $p_1, \ldots, p_m$, and that root p_i has order k_i for each i. To cope with obstructions that arise but also to allow Newton–Raphson to proceed when we do make progress, we allow for varying combinations of "signal" and "noise" in our algorithm, depending on the current state. Thus we might base our algorithm on iterates of independent copies of the following map:

$$M(t) = t - \lambda(t)\frac{g(t)}{g'(t)} + (1 - \lambda(t))X, \tag{3.4}$$

where X is a random variable with distribution symmetric about 0 , and $\lambda : [-1, 1] \to [0, 1]$ is a function yet to be specified. Again we adopt the reset remedy when this algorithm jumps out of $[-1, 1]$.

First we want $\lambda(p_i) = 1$ for $i = 1, \ldots, k$, and $\lambda(x) = 0$ whenever x is a root of g' other than $p_1, \ldots, p_m$. One way to accomplish this is to set $\lambda(x) = \psi(|g(x)|/|g'(x)|)$ for some function $\psi : [0, \infty] \to [0, 1]$ such that $\psi(0) = 1$ and $\psi(\infty) = 0$. Then for each $i = 1, \ldots, m$ we have

$$\lim_{t \to p_i} \frac{|M(t) - p_i|}{|t - p_i|} = \left|1 - \frac{1}{k_i} - \frac{1}{k_i}\psi'(0)X\right|. \tag{3.5}$$

The accessibility condition (i) of the positive convergence theorem will hold provided we take X to have density with support all of $\mathbf{R}$. We must also deal with the local convergence condition (ii). We can take $\psi(x) = \exp(-cx)$ for all $x \in [0, \infty]$, for some $c > 0$. Then the right-hand side of (3.3) becomes $|a + bX|$, where $a = 1 - 1/k_i$ and $b = c/k_i$. Corollary 2.7 now gives convergence to some root of g provided

$$E(\log |a + bX|) < 0. \tag{3.6}$$

Since we control the density of X we can find a and b for which (3.4) holds.

If X is standard normal, then $2E(\log |a + bX|) = 2 \log b + E(\log(\eta + X)^2)$, where $(\eta + X)^2$ has the noncentral chi-squared distribution with one degree of freedom and noncentrality parameter $\eta = a/b$. If we assume only simple roots, then $\eta = 0$ and

$$2E(\log |a + bX|) = 2 \log b - \gamma - \log 2, \tag{3.7}$$

where $\gamma = .5772157\ldots$ is the Euler constant. So in this case we need $b < \exp((\gamma + \log 2)/2) = 1.89$. To handle roots of any order, we set $a = 1$ to cover all cases, and by reference to [4, Chap. 28] (see also [2, p. 15], for special functions background), we find

(3.8)

$$2E(\log |1 + bX|) = 2 \log b - \gamma - \log 2 + \exp\left(-\frac{1}{2b}\right) \sum_{j=1}^{\infty} \frac{(2b)^{-j}}{j!} \left[2 + \frac{2}{3} + \cdots + \frac{2}{2j - 1}\right],$$

from which we determine that we need $b < 1.36$.

On the other hand, we may want a heavier tailed distribution for X. If X has the Cauchy density $(\pi(1+x^2))^{-1}$, $-\infty < x < \infty$, then we see that $E(\log|a+bX|) = \log\sqrt{a^2+b^2}$. Thus if simple roots are our only concern we can take $b < 1$, but otherwise we must judge the maximum order k of root that will occur, set $a = 1 - 1/k$, and choose $b < \sqrt{1-a^2}$.

If we redefine ψ to be 0 whenever $|g/g'| \geq M$ and 1 whenever $|g/g'| \leq m$, and the latter condition is inside the region where Newton–Raphson is working, then we could take X to be uniform on an interval symmetric about 0 and determined so that

$$\psi\left(\frac{|g|}{|g'|}\right)\frac{g}{g'} + \left(1 - \psi\left(\frac{|g|}{|g'|}\right)\right)X \tag{3.9}$$

always has support containing $[-\epsilon, \epsilon]$ for some $\epsilon > 0$. This last definition, if successful, has the virtue that the final stages of convergence will be pure Newton–Raphson, and hence optimally fast.

Finally, we remark that if we wish to exclude a known root p_i from our target set, we simply modify λ so that $\lambda(p_i) = 0$, and this has the effect of removing p_i from our fixed point set almost surely.

4. On order of convergence and expected time to the stopping set. Our results have interpretations in terms of order of convergence. Suppose that the hypotheses of Theorem 2.5 hold; then $X_n \to T$ almost surely and by Proposition 2.4

$$\limsup_{d(\mathbf{t})\to 0} \frac{E(d(M(\mathbf{t}))^\beta)}{d(\mathbf{t})^\beta} < 1 \quad \text{for some } \beta > 0. \tag{4.1}$$

Let $\delta > 0$ and $C < 1$ such that $E(d(M(\mathbf{t}))^\beta) < Cd(\mathbf{t})^\beta$ whenever $d(\mathbf{t}) \leq \delta$. Let J be the (random) minimum integer such that $d(X_n) \leq \delta$ for all $n \geq J$; then $J < \infty$ almost surely. Also,

$$E(d(X_{J+k+1})^\beta) < CE(d(X_{J+k})^\beta) \tag{4.2}$$

for all k. If we now put a new metric on K so that the distance from $\mathbf{x}$ to $\mathbf{p}$ in the new metric is just $\|\mathbf{x}-\mathbf{p}\|^\beta$, then almost surely we have eventual mean C-linear convergence to $\mathbf{p}$.

Now suppose again that the hypotheses in our positive convergence theorem hold, and in addition $E(\log M_\eta) < 0$ for some $\eta > 0$, where

$$M_\eta = \limsup_{d(\mathbf{t})\to 0} \frac{d(M(\mathbf{t}))}{d(\mathbf{t})^\eta}. \tag{4.3}$$

Then we again have that $X_n \to T$ almost surely and the same arguments as in Proposition 2.4 can be used to show

$$E(d(X_{n+1})^\beta) < CE(d(X_n)^{\beta\eta}) \tag{4.4}$$

for large n, for some positive $\beta \leq 1$, and some $C < \infty$. Recoordinatizing K as above we have in this case almost sure eventual mean order η convergence.

To obtain realistic estimates of the expected number of steps to a stopping set we must first realize that a successful Markov algorithm will have two broad phases

of convergence. The first stage will be more or less random searching for the zone of rapid convergence. Let us first assume that for some positive integer q such that the probability of reaching this zone in q steps starting from any $x \in K$ has the lower bound $b > 0$ uniformly in x; then replacing the original Markov algorithm by q-fold iterates, we may assume the probability of reaching the zone of rapid convergence in one step has the uniform lower bound b over K. Then the expected number of steps to this zone is bounded above by the expected value of a geometric random variable whose stopping probability on a single trial is b. We finally obtain q/b as an upper bound for the expected number of steps of the original algorithm to the zone of rapid convergence. We should note that the more uniform in x the probability of reaching the zone of rapid convergence is, the more realistic will be this estimate.

The second stage of convergence is that which occurs after the random variable enters the zone of rapid convergence. First assume eventual linear mean convergence, and (possibly as a worst case) that

$$\frac{d(M(\mathbf{t}))}{d(\mathbf{t})} \simeq U, \tag{4.5}$$

a nondegenerate random variable independent of $\mathbf{t}$. It follows that $\log d(X_{n+1}) - \log d(X_n) \simeq \log U_{n+1}$ is approximately a sequence of independent and identically distributed random variables and that we are essentially dealing with a general random walk, familiar from the theory of sequential analysis. By assumption, $E(\log U_n) = a < 0$. Taking the stopping set to be of the form $\log d(X_n) < -C$, then as a consequence of Wald's identity we have that the expected number of steps to stopping is $\simeq C/a$, provided we are willing as usual to neglect the excess probability over boundary. (See [7, Prop. 2.18].)

Finally let us consider the case of eventual mean convergence of order $\eta > 1$. Assume

$$\frac{d(M(\mathbf{t}))}{d(\mathbf{t})^{\eta}} \simeq U, \tag{4.6}$$

a nondegenerate random variable independent of $\mathbf{t}$. Then $\log d(X_{n+1}) - \eta \log d(X_n) \simeq \log U_{n+1}$ defines an approximately independent and identically distributed sequence of random variables. For each n, let $Y_n = -\log U_n$ and $W_n = -\log d(X_n)$. Then $Y_1, \ldots, Y_n, \ldots$, are independent and identically distributed, $E(Y_n) > 0$, and $W_{n+1} = Y_{n+1} + \eta W_n = Y_{n+1} + \eta Y_n + \cdots + \eta^{n+1} Y_0$ for each n. The Wald analysis does not seem to apply immediately to this situation, but with additional reasonable assumptions we can still get an upper bound for the expected number of steps to stopping. We assume that we are far enough into rapid convergence that $Y_n \geq 0$ for each n, and also that we have a uniform upper bound B for the density of Y_n. Let $W_n \geq C$ be the stopping criterion and let J be the stopping time. Then $J = 1 + \Sigma_{n=1}^{\infty} I_{[W_1, \ldots, W_n < C]}$, so that

$$E(J) = 1 + \Sigma_{n=1}^{\infty} P(W_1, \ldots, W_n < C) = 1 + \Sigma_{n=1}^{\infty} P(W_n < C). \tag{4.7}$$

However,

$$P(W_n < C) = P(Y_0 + \cdots + \eta^{-n} Y_n < C\eta^{-n}). \tag{4.8}$$

Now we have three upper bounds for $P(W_n < C)$:

(a) $P(W_n < C) \leq 1$;
(b) $P(W_n < C) \leq P(Y_0 < C\eta^{-n}) \leq BC\eta^{-n}$; and
(c) $P(W_n < C) < B^{n+1}C^{n+1}[\eta^{n(n+1)/2}(n+1)!]^{-1}$.

To see (c) we first observe that the density for each $\eta^{-k}Y_k$ is bounded above by $B\eta^k$ and so

(4.9)
$$P(Y_0 + \cdots + \eta^{-n}Y_n < C\eta^{-n}) \leq \frac{B^{n+1}\eta^{n(n+1)/2}(C\eta^{-n})^{n+1}}{(n+1)!} = \frac{B^{n+1}C^{n+1}\eta^{-n(n+1)/2}}{(n+1)!}.$$

For each n let $B_n = 1 \wedge BC\eta^{-n} \wedge B^{n+1}C^{n+1}\eta^{-n(n+1)/2}/(n+1)!$. Then $E(J) \leq 1 + \Sigma_{n=1}^{\infty} B_n$. This series converges quickly: for example, if $B = 1, C = 10$, and $\eta = 2$, we get $E(J) \leq 4.67$.

Acknowledgment. We would like to thank Ian Dinwoodie for many helpful suggestions and discussions about this work.

REFERENCES

[1] K. L. CHUNG, *A Course in Probability Theory*, 2nd ed., Academic Press, New York, 1974.
[2] A. ERDELYI, ED., *Higher Transcendental Functions*, Vol. I, *The Bateman Manuscript Project*, McGraw-Hill, New York, 1953.
[3] B. HAJEK, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.
[4] N. JOHNSON AND S. KOTZ, *Continuous Univariate Distributions - 1, 2, Distributions in Statistics*, Houghton-Mifflin, Boston, 1970.
[5] G. JOSEPH, A. LEVINE, AND J. LIUKKONEN, *Randomized Newton–Raphson*, Appl. Numer. Math., 6 (1990), pp. 459–469.
[6] ———, *A stochastic numerical method for estimating the thickness of subsurface strata*, in Proc. Offshore Technology Conference, Houston, TX, May, 1991.
[7] D. SIEGMUND, *Sequential Analysis Tests and Confidence Intervals*, Springer-Verlag, New York, 1985.
[8] P. WALTERS, *An Introduction to Ergodic Theory*, Springer-Verlag, New York, 1982.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1763–1768, November 1994

014

MINIMIZING THE EUCLIDEAN CONDITION NUMBER*

RICHARD D. BRAATZ† AND MANFRED MORARI‡

Abstract. This paper considers the problem of determining the row and/or column scaling of a matrix A that minimizes the condition number of the scaled matrix. This problem has been studied by many authors. For the cases of the ∞-norm and the 1-norm, the scaling problem was completely solved in the 1960s. It is the Euclidean norm case that has widespread application in robust control analyses. For example, it is used for integral controllability tests based on steady-state information, for the selection of sensors and actuators based on dynamic information, and for studying the sensitivity of stability to uncertainty in control systems.

Minimizing the scaled Euclidean condition number has been an open question—researchers proposed approaches to solving the problem numerically, but none of the proposed numerical approaches guaranteed convergence to the true minimum. This paper provides a convex optimization procedure to determine the scalings that minimize the Euclidean condition number. This optimization can be solved in polynomial-time with off-the-shelf software.

Key words. scaling, conditioning, condition number

AMS subject classifications. 65F35, 93B35, 93D21

1. Introduction.

Let $V_1 = \mathbf{C}^n$ be the normed complex vector space with Hölder p-norm $\|\cdot\|_p$, $\|x\|_p = (\sum |x_j|^p)^{1/p}$. For an $n \times n$ matrix $A : V_1 \to V_1$, the following induced matrix norm is defined:

$$\|A\|_{ip} = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \tag{1}$$

If the inverse A^{-1} exists, then the condition number subordinate to the norm $\|\cdot\|_p$ is defined by

$$\kappa_p(A) = \|A\|_{ip}\|A^{-1}\|_{ip}. \tag{2}$$

Define $\mathbf{C}^{n\times n}$ to be the set of complex $n \times n$ matrices. Let $\mathbf{D}^{n\times n}$ be the set of all diagonal invertible matrices in $\mathbf{C}^{n\times n}$. If $A \in \mathbf{C}^{n\times n}$ is the matrix defining a system of linear equations $Ax = b$, scaling the rows of this system is equivalent to premultiplying A by a diagonal matrix $D_1 \in \mathbf{D}^{n\times n}$. Scaling the unknowns is equivalent to postmultiplying A by a diagonal matrix $D_2 \in \mathbf{D}^{n\times n}$. The quality of numerical computations is generally better when the condition number of A is small. Since diagonal scalings of A are trivial modifications, researchers in the 1960s–1970s were led to investigate the following minimizations in order to obtain optimal scalings of a matrix:

$$\begin{aligned} &\text{(i)} && \kappa_p^l(A) = \inf_{D_1 \in \mathbf{D}^{n\times n}} \kappa_p(D_1 A), \\ &\text{(ii)} && \kappa_p^r(A) = \inf_{D_2 \in \mathbf{D}^{n\times n}} \kappa_p(A D_2), \\ &\text{(iii)} && \kappa_p^{lr}(A) = \inf_{D_1, D_2 \in \mathbf{D}^{n\times n}} \kappa_p(D_1 A D_2). \end{aligned} \tag{3}$$

* Received by the editors October 13, 1992; accepted for publication (in revised form) June 22, 1993.

† Department of Control and Dynamical Systems, California Institute of Technology, Pasadena, California 91125 (rdb@mozart.caltech.edu). This author was supported by the Fannie and John Hertz Foundation.

‡ Department of Chemical Engineering, California Institute of Technology, Pasadena, California 91125 (mm@imc.caltech.edu).

TABLE 1

Minimized condition numbers. The matrix whose elements are the moduli of the corresponding elements of A is denoted by $|A|$. The spectral radius of A is denoted by $\rho(A)$. The maximum singular value $\overline{\sigma}(A)$ refers to $\|A\|_{i2}$.

	$p = 1, \infty$	$p = 2$
$\inf_{D_1} \kappa_p(D_1 A)$	$\overline{\sigma}\left(\lvert A^{-1}\rvert \cdot \lvert A\rvert\right)$	?
$\inf_{D_2} \kappa_p(A D_2)$	$\overline{\sigma}\left(\lvert A\rvert \cdot \lvert A^{-1}\rvert\right)$	?
$\inf_{D_1, D_2} \kappa_p(D_1 A D_2)$	$\rho(\lvert A\rvert \cdot \lvert A^{-1}\rvert)$	?

Problem (3(i)) was present for example in the error analysis of direct methods for the solution of linear equations [34], [2]. Problem (3(ii)) is important for obtaining the best possible bounds for eigenvalue inclusion theorems [3], and is a natural measure of the linear independence of the column vectors that form A [2]. Problem (3(iii)) was used for decreasing the error in calculation of the matrix inverse A^{-1} [14].

Later, it was realized that the appropriate scalings depend on the error in the matrix, not the elements of the matrix itself [10], [31]. This implied, for example, that the scalings solving problem (3(iii)) are not necessarily the best scalings of A to decrease the error in the calculation of A^{-1}. However, problems (3(i))–(3(iii)) still have widespread application in robust control analyses. For example, the minimized condition number (3(iii)) is used for integral controllability tests based on steady-state information [13], [18], and for the selection of sensors and actuators using dynamic information [24], [19], [20]. The sensitivity of stability to uncertainty in control systems is given in terms of the minimized condition number in [29], [30].

Without loss of generality, for each of these problems we need only consider the infimum over the set of real positive diagonal invertible matrices $\mathbf{D}_+^{n\times n}$. This is because any matrix in $\mathbf{D}^{n\times n}$ can be decomposed into a matrix in $\mathbf{D}_+^{n\times n}$ and a unitary diagonal matrix. The unitary diagonal matrix does not affect the value of the condition number in (2) (see [2] for a simple proof). Conditions for the existence of scaling matrices that achieve the infimum are given by Businger [6].

The minimizations were solved for $p = 1$ and $p = \infty$ by Bauer [2] (the results are in Table 1). Many researchers consider the 2-norm as most important for applications [2], [14], [17]. Solving (3(i))–(3(iii)) for the 2-norm has been an open question [28], [35]. In this paper we solve the minimizations for the 2-norm by transforming the minimizations (3(i))–(3(iii)) so that they can be solved via convex programming.

Nonsquare A [33], block diagonal scalings [12], [27], [9], [11], [35], and cross-condition numbers (with B replacing A^{-1} in (2), see [8], [16], [15]) have also received attention. For ease of notation, the results are derived for square matrices with fully diagonal scalings. The results (and proofs) hold for these other cases with the modifications given after the lemmas.

2. Results. The induced matrix norm for the vector 2-norm is commonly referred to as the maximum singular value, $\overline{\sigma}(A) = \|A\|_{i2}$. To simplify notation, drop the subscript on κ_2, i.e., $\kappa_2 = \kappa$. Let $\mathbf{R}_+$ be the set of real positive scalars. Let I be the $n \times n$ identity matrix.

LEMMA 2.1. *The following equality holds*:

$$\kappa(A) = \inf_{d_1, d_2 \in \mathbf{R}_+} \overline{\sigma}^2 \left(\begin{bmatrix} d_1 I & 0 \\ 0 & d_2 I \end{bmatrix} \begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix} \begin{bmatrix} (d_1)^{-1} I & 0 \\ 0 & (d_2)^{-1} I \end{bmatrix} \right). \tag{4}$$

Proof.

$$\inf_{d_1,d_2\in\mathbf{R}_+} \overline{\sigma}^2\left(\begin{bmatrix} d_1 I & 0 \\ 0 & d_2 I \end{bmatrix}\begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix}\begin{bmatrix} (d_1)^{-1} I & 0 \\ 0 & (d_2)^{-1} I \end{bmatrix}\right) \tag{5}$$

$$= \inf_{d_1,d_2\in\mathbf{R}_+} \overline{\sigma}^2\left(\begin{bmatrix} 0 & \frac{d_1}{d_2}A^{-1} \\ \frac{d_2}{d_1}A & 0 \end{bmatrix}\right) \tag{6}$$

$$= \inf_{d_1,d_2\in\mathbf{R}_+} \max\left\{\overline{\sigma}^2\left(\frac{d_1}{d_2}A^{-1}\right), \overline{\sigma}^2\left(\frac{d_2}{d_1}A\right)\right\} \tag{7}$$

$$= \inf_{d_1,d_2\in\mathbf{R}_+} \max\left\{\frac{d_1^2}{d_2^2}\frac{\overline{\sigma}\left(A^{-1}\right)}{\overline{\sigma}\left(A\right)}, \frac{d_2^2}{d_1^2}\frac{\overline{\sigma}\left(A\right)}{\overline{\sigma}\left(A^{-1}\right)}\right\} \cdot \overline{\sigma}\left(A\right)\overline{\sigma}\left(A^{-1}\right) \tag{8}$$

$$= \inf_{x\in\mathbf{R}_+} \max\left\{x, x^{-1}\right\} \cdot \overline{\sigma}\left(A\right)\overline{\sigma}\left(A^{-1}\right) \tag{9}$$

$$= \overline{\sigma}\left(A\right)\overline{\sigma}\left(A^{-1}\right) = \kappa(A). \tag{10}$$

Note that this proof is similar to a proof in [21]. □

The following lemma gives similar expressions as in (4) for $\kappa^l(A)$, $\kappa^r(A)$, and $\kappa^{lr}(A)$.

LEMMA 2.2. *The following equalities hold:*

$$\kappa^l(A) = \inf_{D_1\in\mathbf{D}_+^{n\times n}} \overline{\sigma}^2\left(\begin{bmatrix} I & 0 \\ 0 & D_1 \end{bmatrix}\begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & D_1^{-1} \end{bmatrix}\right). \tag{11}$$

$$\kappa^r(A) = \inf_{D_2\in\mathbf{D}_+^{n\times n}} \overline{\sigma}^2\left(\begin{bmatrix} D_2^{-1} & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix}\begin{bmatrix} D_2 & 0 \\ 0 & I \end{bmatrix}\right). \tag{12}$$

$$\kappa^{lr}(A) = \inf_{D\in\mathbf{D}_+^{2n\times 2n}} \overline{\sigma}^2\left(D\begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix}D^{-1}\right). \tag{13}$$

Proof. Substituting D_1AD_2 for A in Lemma 2.1 and rearranging gives

$$\kappa(D_1AD_2) = \inf_{d_1,d_2\in\mathbf{R}_+} \overline{\sigma}^2\left(\begin{bmatrix} d_1D_2^{-1} & 0 \\ 0 & d_2D_1 \end{bmatrix}\begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix}\begin{bmatrix} (d_1D_2^{-1})^{-1} & 0 \\ 0 & (d_2D_1)^{-1} \end{bmatrix}\right), \tag{14}$$

where d_1 and d_2 are real positive scalars.

Take the infimum over D_1 and D_2 on both sides to give

$$\kappa^{lr}(A) = \inf_{D_1,D_2\in\mathbf{D}_+^{n\times n}} \inf_{d_1,d_2\in\mathbf{R}_+} \overline{\sigma}^2\left(\begin{bmatrix} d_1D_2^{-1} & 0 \\ 0 & d_2D_1 \end{bmatrix}\begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix}\begin{bmatrix} (d_1D_2^{-1})^{-1} & 0 \\ 0 & (d_2D_1)^{-1} \end{bmatrix}\right) \tag{15}$$

$$= \inf_{D_1,D_2\in\mathbf{D}_+^{n\times n}} \overline{\sigma}^2\left(\begin{bmatrix} D_2^{-1} & 0 \\ 0 & D_1 \end{bmatrix}\begin{bmatrix} 0 & A^{-1} \\ A & 0 \end{bmatrix}\begin{bmatrix} D_2 & 0 \\ 0 & D_1^{-1} \end{bmatrix}\right). \tag{16}$$

Letting $D = \operatorname{diag}\left\{D_2^{-1}, D_1\right\}$ gives (13). Expressions (11) and (12) are proved similarly. □

Let I_r be the $r \times r$ identity matrix. Let $\mathcal{D}^{2n\times 2n} = \operatorname{diag}\{[d_1 I_{r_1}, \ldots, d_m I_{r_m}] : d_j \in \mathbf{R}, r_1 + \cdots + r_m = 2n\}$, and $M \in \mathcal{C}^{2n\times 2n}$. Consider the following lemma.

LEMMA 2.3. *The following optimization is convex:*

$$\inf_{D\in\mathbf{D}^{2n\times 2n}} \overline{\sigma}^2\left(e^D M e^{-D}\right). \tag{17}$$

Proof. See [25]. □

Because $\left\{e^D : D \in \mathcal{D}\right\} = \{D : D \in \mathcal{D}_+\}$, the optimizations in Lemmas 2.1 and 2.2 are equivalent to the optimization in Lemma 2.3. This means that the condition number κ and minimized condition numbers κ^l, κ^r, κ^{lr} can all be calculated through convex programming. Since the optimization (17) is convex, it can have only one minimum.

The optimization (17) has been studied extensively [22], [32], [23], [25], and off-the-shelf software is available for solving these polynomial-time problems (for example, see the program *mu* in [1]). The calculation of the minimized condition numbers is slow, however, since the minimization (17) requires repeated maximum singular value calculations.

The parallelism between expressions (4), (11), (12), and (13) for κ, κ^l, κ^r, and κ^{lr} is interesting. The same optimization can be used for the condition number calculations—the optimizations are just over different "scaling matrices." This is nice theoretically, since κ^l, κ^r, and κ^{lr} are just the *scaled* condition numbers.

Remark 2.4. Conditions for the existence of scaling matrices that achieve the infimum are given by Businger [6]. When the infinum is achieved, any algorithm that solves (17) provides the minimizing scaling matrices for the condition number. When the infinum is not achieved, the algorithm provides scaling matrices such that the infinum is approached with arbitrary closeness.

Remark 2.5. To generalize to nonsquare A, replace every occurrence of A^{-1} with the respective right or left inverse. More specifically, if $A \in \mathbf{C}^{m\times n}$ and has full row rank with $m < n$, then replace A^{-1} with $A^T(AA^T)^{-1}$ in all proofs and lemmas. For $m > n$ with A having full column rank, replace A^{-1} with $(A^TA)^{-1}A^T$.

Remark 2.6. The Euclidean cross-condition number is defined by

$$\hat{\kappa}(A, B) := \overline{\sigma}(A)\,\overline{\sigma}(B). \tag{18}$$

Minimized cross-condition numbers can be defined similarly as in (3), for example,

$$\hat{\kappa}^{lr}(A, B) := \inf_{D_1, D_2\in\mathbf{D}^{n\times n}} \hat{\kappa}(D_1 A D_2, D_2^{-1} B D_1^{-1}). \tag{19}$$

Lemmas 2.1 and 2.2 follow with B, replacing A^{-1}. This problem is important for testing stability of systems with element-by-element uncertainty [7], [8], [16], [15].

Remark 2.7. For block-diagonal scaling matrices, without loss of generality we can take each block to be positive definite Hermitian. This is because any nonsingular complex matrix can be decomposed into a positive definite Hermitian matrix and a unitary matrix [4], and the unitary matrix does not affect the value of the Euclidean condition number. The proofs of Lemmas 2.1 and 2.2 follow exactly as for the fully diagonal case. Lemma 2.3 does not hold for block-diagonal scalings. For block-diagonal scalings it is better to convert the singular value minimizations in Lemma 2.2

into generalized eigenvalue minimizations, as follows:

$$\inf_{D\in\mathbf{D}_+} \overline{\sigma}(DMD^{-1}) = \inf_{D^2\in\mathbf{D}_+} \left\{\beta \,\middle|\, M^*D^2M - \beta D^2 < 0\right\}. \tag{20}$$

The condition $M^*D^2M - \beta D^2 < 0$ is convex in D^2, so any local minimum is global, and off-the-shelf software is available [1]. Many researchers are working to develop improved computational approaches for these polynomial-time problems (for example, see [5] and the literature cited therein).

3. Conclusions. We have completed Table 1 in the sense that all values in the table can now be calculated with arbitrary precision.

All entries in the table, including the now-filled entries, require the inverse of A to calculate the minimizing scalings and the minimized condition numbers. There are algorithms for numerically determining the minimized condition numbers without predetermining the matrix inverse [26], [35], but these methods are not guaranteed to converge to the true minima.

REFERENCES

[1] G. J. Balas, A. K. Packard, J. C. Doyle, K. Glover, and R. S. R. Smith, *Development of advanced control design software for researchers and engineers*, in Proc. 1991 American Control Conference, 1991, pp. 996–1001.

[2] F. L. Bauer, *Optimally scaled matrices*, Numer. Math., 5 (1963), pp. 73–87.

[3] F. L. Bauer and A. S. Householder, *Absolute norms and characteristic roots*, Numer. Math., 3 (1961), pp. 241–246.

[4] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[5] S. Boyd and L. E. Ghaoui, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188 (1993), pp. 63–111.

[6] P. A. Businger, *Matrices which can be optimally scaled*, Numer. Math., 12 (1968), pp. 346–348.

[7] J. Chen, M. K. H. Fan, and C. N. Nett, *Robustness analysis with non-diagonally structured uncertainty*, IEEE Trans. Automat. Control, (1994), preprint.

[8] R. W. Daniel, B. Kouvaritakis, and H. Latchman, *Principal direction alignment: A geometric framework for the complete solution to the μ-problem*, IEE Proceedings Part D, 133 (1986), pp. 45–56.

[9] J. W. Demmel, *The condition number of equivalence transformations that block diagonalize matrix pencils*, SIAM J. Numer. Anal., 20 (1983), pp. 599–610.

[10] J. J. Dongarra, C. B. Moler, J. R. Bunch, and G. W. Stewart, *LINPACK User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[11] P. V. Dooren and P. Dewilde, *Minimal cascade factorization of real complex rational transfer matrices*, IEEE Trans. Circuits and Systems, 28 (1981), pp. 390–400.

[12] S. C. Eisenstat, J. W. Lewis, and M. H. Schultz, *Optimal block diagonal scaling of block 2-cyclic matrices*, Linear Algebra Appl., 44 (1982), pp. 181–186.

[13] P. Grosdidier, M. Morari, and B. R. Holt, *Closed-loop properties from steady-state gain information*, Ind. Eng. Chem. Fund., 24 (1985), pp. 221–235.

[14] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964, p. 123.

[15] B. Kouvaritakis and H. Latchman, *Necessary and sufficient stability condition for systems with structured uncertainties: The major principal direction alignment principle*, Internat. J. Control, 42 (1985), pp. 575–598.

[16] ———, *Singular-value and eigenvalue techniques in the analysis of systems with structured perturbations*, Internat. J. Control, 41 (1985), pp. 1381–1412.

[17] C. McCarthy and G. Strang, *Optimal conditioning of matrices*, SIAM J. Numer. Anal., 10 (1973), pp. 370–388.

[18] M. Morari and E. Zafiriou, *Robust Process Control*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[19] C. N. Nett, *Decentralized control system design for a variable-cycle gas turbine engine*, in Proc. 1990 American Control Conference, 1990.

[20] C. N. NETT AND K. D. MINTO, *A quantitative approach to the selection and partitioning of measurements and manipulations for the control of complex systems*, in Proc. 1989 American Control Conference, 1989.

[21] C. N. NETT AND J. A. UTHGENANNT, *An explicit formula and an optimal weight for the 2-block structured singular value interaction measure*, in Proc. 1987 American Control Conference, 1987, pp. 506–511.

[22] E. E. OSBORNE, *On pre-conditioning of matrices*, J. Assoc. Comp. Mach., 7 (1960), pp. 338–345.

[23] A. PACKARD AND J. C. DOYLE, *The complex structured singular value*, Automatica, 29 (1993), pp. 71–109.

[24] D. E. REEVES, C. N. NETT, AND Y. ARKUN, *Control configuration design for complex systems: A practical theory*, Tech. report, Georgia Inst. Tech., Athens, GA, 1992.

[25] R. SEZGINER AND M. OVERTON, *The largest singular value of $e^X A e^{-X}$ is convex on convex sets of commuting matrices*, IEEE Trans. Automat. Control, 35 (1990), pp. 229–230.

[26] A. SHAPIRO, *Weighted minimum trace factor analysis*, Psychometrika, 47 (1982), pp. 243–264.

[27] ———, *Optimal block diagonal l_2 scaling of matrices*, SIAM J. Numer. Anal., 22 (1985), pp. 81–94.

[28] ———, *Upper bounds for nearly optimal diagonal scaling of matrices*, Linear and Multilinear Algebra, 29 (1991), pp. 145–147.

[29] S. SKOGESTAD AND M. MORARI, *Design of resilient processing plants—IX. Effect of model uncertainty on dynamic resilience*, Chem. Eng. Sci., 42 (1987), pp. 1765–1780.

[30] ———, *Implications of large* RGA *elements on control performance*, Ind. Eng. Chem. Res., 26 (1987), pp. 2323–2330.

[31] G. W. STEWART AND J. G. SUN, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.

[32] J. STOER AND C. WITZGALL, *Transformations by diagonal matrices in a normed space*, Numer. Math., 4 (1962), pp. 158–171.

[33] A. VAN DER SLUIS, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

[34] ———, *Condition, equilibration and pivoting in linear algebraic systems*, Numer. Math., 15 (1970), pp. 74–86.

[35] G. A. WATSON, *An algorithm for optimal l_2 scaling of matrices*, IMA J. Numer. Anal., 11 (1991), pp. 481–492.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1769–1781, November 1994

015

CONTINUITY OF BEST HANKEL APPROXIMATION AND CONVERGENCE OF NEAR-BEST APPROXIMANTS*

CHARLES K. CHUI† AND XIN LI‡

Abstract. Consider a bounded Hankel operator Γ with s-numbers $s_0 \geq s_1 \geq \cdots$ and a sequence of bounded Hankel operators Γ_n converging to Γ in the operator norm. In this paper, it is shown that for each k with $s_{k-1} > s_k \geq s_{k+1} \geq \cdots$, the rational symbols of the best rank-k Hankel approximants of Γ_n converge uniformly to the corresponding rational symbol of the best rank-k Hankel approximant of Γ. Based on this continuity result, the convergence of the near-best Hankel approximants corresponding to a fairly general class of truncated Hankel operators is discussed.

Key words. Hankel operators; Adamjun, Arov, and Krein (AAK); rational approximation; near-best approximation; continuity of best approximation

AMS subject classifications. primary 47B05, 41A35; secondary 41A20

1. Introduction.

Studies of Hankel operators and Hankel-type approximation date back to the original work of Carathéodory and Fejér [4], Schur [28], Takagi [29], and Achiezer [3]. A more recent fundamental approach, which may be regarded as a complete extension of the celebrated theorem of Nehari [22], was given by Adamjan, Arov, and Krein [1], [2], and for this reason, it is commonly called the AAK theory. Applications of Hankel operators to signal processing, H^∞-control, and approximation theory can be found in [8], [10], [12], [14], [21].

The objective of this paper is to give a complete answer to the question of "continuity" of best Hankel approximation. This result has important applications to systems theory and H^∞-control, and we also discuss the convergence of near-best Hankel approximants that correspond to a fairly large class of truncated Hankel operators. Similar problems have been studied in the literature. We only mention Helton and Schwartz [20], Hayashi, Trefethen, and Gutknecht [19], and Peller [26] (also cf. [32]).

Let R_k denote the class of all strictly proper rational functions with at most k poles (counting multiplicities), all of which lie inside the unit circle. As usual, we denote by L^∞ the Banach space of all essentially bounded measurable functions on the unit circle $|z| = 1$, and by H^∞, its Hardy subspace of bounded analytic functions in $|z| < 1$. We consider approximation of functions in L^∞ from

$$\widetilde{R}_k := R_k + H^\infty. \tag{1.1}$$

It is well known (cf. [2]) that, for each integer $k \geq 0$ and function $f \in L^\infty$, there exists a unique $g_k \in \widetilde{R}_k$ such that

$$\|f - g_k\|_\infty = \inf_{g \in \widetilde{R}_k} \|f - g\|_\infty. \tag{1.2}$$

*Received by the editors June 10, 1992; accepted for publication (in revised form) July 6, 1993. This research was partially supported by National Science Foundation grant DMS-8901345 and Army Research Office contract DAAL 03-90-G-0091.

†Department of Mathematics, Texas A&M University, College Station, Texas 77843-3368 (cat@math.tamu.edu).

‡Department of Mathematical Sciences, University of Nevada, Las Vegas, Nevada 89154-4020 (xinlixin@nevada.edu). The research of this author was partially supported by the University Research Grants and Fellowship Committee at the University of Nevada, Las Vegas.

The (nonlinear) operator

$$A_k\colon\ L^\infty \to \widetilde{R}_k, \tag{1.3}$$

defined by $A_k(f) = g_k$ as in (1.2), is called the operator of best approximation (see [19], [26]). Hence, the problem of continuity of this operator is to study conditions under which we have

$$\|f_n - f\|_\infty \to 0 \Rightarrow \|A_k(f_n) - A_k(f)\|_\infty \to 0, \quad \text{as } n \to \infty. \tag{1.4}$$

The problem of best approximation in (1.2) happens to be intimately related to the problem of best approximation of bounded Hankel operators by those with (specified) finite ranks. This is the so-called AAK theory, and we give a very brief review of this topic in the next section. For the time being, we only note that corresponding to any $f \in L^\infty$, there corresponds a unique bounded Hankel operator Γ_f, and that the measurement of error of best approximation in (1.2) is given by the singular values (better known as s-numbers in the AAK theory) $s_k\ (\Gamma_f)$ of Γ_f. These values are arranged in nonincreasing order, with $s_0(\Gamma)$ being the largest one.

Returning to the discussion of the continuity of the operator A_k as described by (1.4), we first mention a result in [20] which says that if the largest s-number $s_0(\Gamma_f)$ is simple and $f, f_n \in C^\infty$, then (1.4) holds for $k = 0$.

The Wiener class W of functions

$$f(z) = \sum_{-\infty}^{\infty} c_n z^n, \qquad |z| = 1,$$

defined by $\sum |c_n| < \infty$, was considered in [19]. Using the norm $\|f\|_W := \sum |c_n|$, and considering some extension of $\widetilde{R}_k$ (see [19, pp. 198–199]), the continuity result in [19] is that $A_k\colon\ W \to W$ is continuous at $f \in W$ relative to the norm $\|\ \|_W$, if and only if the s-number $s_k\ (\Gamma_f)$ is simple. This result is a consequence of the theorems of Wiener and AAK (cf. [19] for further discussions).

Some larger classes of functions, such as VMO and Besov classes, were considered in [26] (see properties $(A_1)-(A_4)$ in [26, pp. 143–144]). Let X be such a function space with norm $\|\ \|_X$ and $f \in X$. Then again under the assumption that the s-number $s_k\ (\Gamma_f)$ is simple, it was shown in [26] that the operator A_k is continuous at f relative to the norm $\|\ \|_X$.

In all the papers [19], [20], [26] mentioned above, the continuity of the operator A_k depends on the basic assumption that the s-number $s_k\ (\Gamma_k)$ is simple, namely that

$$s_{k-1}(\Gamma_f) > s_k(\Gamma_f) > s_{k+1}(f) \geq \cdots, \tag{1.5}$$

with $s_{-1}(f) := +\infty$. In fact, some counterexamples and negative results were given in [19], [20] and [26], respectively, to show that the results there do not hold otherwise.

In this paper, we attack the continuity problem by considering the intimate relationship between the approximation problem (1.2) and that of best approximation of bounded Hankel operators by those with (specified) finite ranks. In doing so, we are able to remove the requirement on the simpleness of the s-numbers. That is, the requirement (1.5) can be relaxed to

$$s_{k-1}(\Gamma_f) > s_k(\Gamma_k) \geq s_{k+1}(f) \geq \cdots\ . \tag{1.6}$$

More precisely, let $\mathcal{H}$ denote the Banach space of all bounded Hankel operators with norm $\|\ \|$, and $G^{[k]}$ the class of those operators in $\mathcal{H}$ with rank at most k. Note that to each $f \in L^\infty$ corresponds a unique $\Gamma = \Gamma_f \in \mathcal{H}$. Then the (best) Hankel approximation problem corresponding to the best approximation problem (1.2) is given by

$$\|\Gamma - \Lambda_k\| = \inf_{\Lambda \in G^{[k]}} \|\Gamma - \Lambda\|, \quad \Lambda_k \in G^{[k]}. \tag{1.7}$$

It is well known that Λ_k is unique, and similar to (1.3), we consider the operator

$$B_k\colon\ \mathcal{H} \to G^{[k]} \tag{1.8}$$

defined by $B_k(\Gamma) = \Lambda_k$. Then the continuity problem of this operator corresponding to (1.4) for A_k is to study conditions under which we have

$$\|\Gamma_n - \Gamma\| \to 0 \Rightarrow \|B_k(\Gamma_n) - B_k(\Gamma)\| \to 0, \quad \text{as } n \to \infty. \tag{1.9}$$

The main result in this paper is that (1.9) holds for all $\Gamma \in \mathcal{H}$ and for each k for which (1.6) is satisfied. That is, we don't require $s_k(\Gamma)$ to be simple, but only k to be the smallest index of the (possibly multiple) s-number. We also give a simple example to demonstrate the sharpness of this result in the case of multiple s-numbers, namely, if $s_k(\Gamma) = s_{k+1}(\Gamma)$, then (1.9) does not hold for $k+1$.

The relation between the continuity considerations (1.4) and (1.9) is governed by the AAK theory (to be discussed in §2). It is important to note that if r_k and $r_{n,k}$ are the rational (symbol) functions corresponding to $B_k(\Gamma)$ and $B_k(\Gamma_n)$, respectively, then the continuity result (1.9) yields

$$\|r_{n,k} - r_k\|_\infty \to 0 \quad \text{as } n \to \infty. \tag{1.10}$$

This is seen in §3. Hence, in parallel to the convergence of near-best "rational" approximants,

$$\|g_{n,k} - g_k\|_\infty = \|A_k(f_n) - A_k(f)\|_\infty \to 0, \quad n \to \infty,$$

as given in (1.4), and we have the convergence of near-best rational approximants as described in (1.10). A discussion of the construction of such approximants via "truncations" of the Hankel operator Γ is discussed in §4.

2. A review of Hankel approximation. In this section, we state some well-known results on Hankel operators that are used in the following sections. For any given function $f(z) = \sum_{n=-\infty}^{\infty} h_n z^{-n}$ in L^∞, the Hankel operator Γ_f associated with f is defined by

$$\Gamma_f g = P(fg), \quad g \in H^2. \tag{2.1}$$

Here and throughout, H^2 denotes the hardy Hilbert space of analytic functions in the unit disc and P denotes the orthogonal projection from L^2 onto $H^2_- := L^2 \ominus H^2$, the subspace of the space L^2 of square-integrable functions on $|z| = 1$, complementary to H^2. Let $e_i = (0, \ldots, 0, 1, 0, \ldots)^T$, $1 \le i < \infty$, whose entries are all zeros except the ith one, which is 1, be the standard basis of ℓ^2. Considering the isometric isomorphisms among ℓ^2, H^2, and H^2_-, we can easily verify that Γ_f has an infinite matrix representation Γ on ℓ^2, with respect to the basis e_i, $1 \le i < \infty$, given by

$$\Gamma = \begin{pmatrix} h_1 & h_2 & h_3 & \cdots \\ h_2 & h_3 & \cdots & \\ h_3 & \vdots & & \\ \vdots & & & \end{pmatrix}. \tag{2.2}$$

We call f the symbol function of the Hankel operator Γ_f. For simplicity, we also denote by $\Gamma = \{h_{i+j-1}\}_{1\le i,j<\infty}$ the Hankel matrix in (2.2). The intimate relations between Hankel operators and their symbol functions are well known. In particular, the following classical result due to Kronecker (see, for example, [13, p. 207]) is instrumental to our discussions.

THEOREM A (Kronecker). *The infinite matrix* $\Gamma = \{h_{i+j-1}\}_{1\le i,j<\infty}$ *has finite rank if and only if the symbol function*

$$\sum_{n=1}^{\infty} h_n z^{-n}$$

is a strictly proper rational function in z; *that is,*

$$\sum_{n=1}^{\infty} h_n z^{-n} = \frac{c(z)}{d(z)},$$

where $c(z)$ *and* $d(z)$ *are relatively prime polynomials in* z *with* $\text{degree}(c) < \text{degree}(d)$. *Furthermore, in this situation*

$$\text{rank}(\Gamma) = \text{degree}(d).$$

If $\text{rank}(\Gamma) \le k$ and we write $c(z) = c_1 z^{k-1} + \cdots + c_k$, $d(z) = z^k + d_1 z^{k-1} + \cdots + d_k$, then it can be shown that Γ has finite rank k if and only if

$$\begin{pmatrix} h_1 & \dots & h_k \\ \vdots & \ddots & \vdots \\ h_k & \dots & h_{2k-1} \end{pmatrix}, \tag{2.3}$$

called the principle minor of order k in Γ, is nonsingular. Moreover, the following relations hold:

$$\begin{pmatrix} h_1 & \dots & h_k \\ \vdots & \ddots & \vdots \\ h_k & \dots & h_{2k-1} \end{pmatrix} \begin{pmatrix} -d_k \\ \vdots \\ -d_1 \end{pmatrix} = \begin{pmatrix} h_{k+1} \\ \vdots \\ h_{2k} \end{pmatrix} \tag{2.4}$$

and

$$\begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} = \begin{pmatrix} h_1 & 0 & \dots & 0 \\ h_2 & h_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_k & h_{k-1} & \dots & h_1 \end{pmatrix} \begin{pmatrix} 1 \\ d_1 \\ \vdots \\ d_{k-1} \end{pmatrix}. \tag{2.5}$$

The following theorem due to Nehari [22] is fundamental to H^∞-control (cf. [12]).

THEOREM B (Nehari). Γ *is a bounded Hankel operator on* ℓ^2 *if and only if there is a function* $f \in L^\infty$ *such that* $\Gamma = \Gamma_f$; *moreover,* f *can be so chosen that* $\|\Gamma\| = \|f\|_\infty$.

We now give some notations in order to introduce the main result of AAK in [1], [2]. For a given Hankel operator Γ on ℓ^2, let $\overline{\Gamma}$ denote its adjoint (or complex conjugate). Let $\sigma(|\Gamma|)$ represent the spectrum of $|\Gamma| := (\overline{\Gamma}\Gamma)^{1/2}$, and write $\sigma(|\Gamma|) =$

$\sigma_1 \cup \sigma_2$, where σ_2 is the (possibly empty) set of isolated eigenvalues $s_k := s_k(\Gamma)$ with finite multiplicities and $\sigma_1 = \sigma(|\Gamma|)\backslash\sigma_2$. Set

$$s_\infty = s_\infty(\Gamma) := \sup\{s:\ s \in \sigma_1\},$$

and order the elements in σ_2, listed according to their multiplicities, as follows:

$$s_0 \geq s_1 \geq \cdots \geq s_\infty.$$

If σ_2 contains $m(< \infty)$ members, then we set $s_{m+1} = s_{m+2} = \cdots = s_\infty$. These numbers s_k, $k = 0, 1, \ldots$, are called the s-numbers of Γ. In the case that $s_{k-1} > s_k = \cdots = s_{k+r} > s_{k+r+1}$ for some integers k and $r \geq 0$, there is, associated with k, an $(r+1)$-dimensional linear manifold of Schmidt pairs $\{\xi, \eta\}$, ξ, $\eta \in \ell^2$, such that

$$\Gamma\xi = s_k\eta \quad \text{and} \quad \overline{\Gamma}\eta = s_k\xi.$$

If we write $\xi = (\xi_1, \xi_2, \ldots)^T \in \ell^2$, $\eta = (\eta_1, \eta_2, \ldots)^T \in \ell^2$ and define

$$\xi_+(z) := \sum_{n=1}^{\infty} \xi_n z^{n-1} \quad \text{and} \quad \eta_-(z) := \sum_{n=1}^{\infty} \eta_n z^{-n},$$

then it follows that

$$|\eta_-(z)/\xi_+(z)| = 1, \qquad |z| = 1,$$

and the function $\eta_-(z)/\xi_+(z)$ does not depend on the choice of the Schmidt pairs $\{\xi, \eta\}$ (cf. [1], [2]). Corresponding to the function $\eta_-(z)/\xi_+(z)$, we consider the Hankel operator

$$\Gamma_{s_k} := s_k\Gamma_{\eta_-/\xi_+}. \tag{2.6}$$

The following theorem is the main result of AAK in [1], [2].

THEOREM C (AAK). *Let s_k be an s-number of Γ and suppose that $s_{k-1} > s_k = \cdots = s_{k+r}$ for some $r \geq 0$. Then there exists a unique bounded Hankel operator Λ_k in $G^{[k+r]}$ such that*

$$s_k = \|\Gamma - \Lambda_k\| = \inf_{\Lambda \in G^{[k+r]}} \|\Gamma - \Lambda\|. \tag{2.7}$$

Moreover, Λ_k is given by $\Gamma - \Gamma_{s_k}$ (cf. (2.6)) and has rank k.

A simple observation, by applying Theorem C, is that if $s_{k-1} > s_k = \cdots = s_{k+r}$, then the best approximants of Γ from $G^{[k]}$, $G^{[k+1]}, \ldots, G^{[k+r]}$, respectively, are identical, and are given by $\Lambda_k = \Gamma - \Gamma_{s_k}$ with $\operatorname{rank}(\Lambda_k) = k$. According to Theorems A and B, the result in Theorem C can be described in the L^∞ sense as follows. Suppose that a bounded Hankel operator Γ is generated by some $f \in L^\infty$ and $s_{k-1} > s_k = \cdots = s_{k+r}$. Then

$$s_k = \inf_{g \in \widetilde{R}_{k+r}} \|f - g\|_\infty, \tag{2.8}$$

and moreover, the infimum is attained at $f(z) - s_k \frac{\eta_-(z)}{\xi_+(z)}$. This result (particularly assertions (2.7) and (2.8)) describes the intimate relationship between the two approximation problems (1.2) and (1.7).

3. Continuity of best Hankel approximation. From Theorem C, we see that corresponding to any bounded Hankel operator Γ there exists a unique best approximant Λ_k from $G^{[k]}$ to Γ. The problem we are concerned with in this paper is the feasibility of replacing Λ_k, and consequently its rational function symbol $r_k(z)$, by $\Lambda_{n,k}$, which best approximates a more realistic model Γ_n (in terms of computational implementations) instead of Γ. More precisely, if $\|\Gamma_n - \Gamma\| \to 0$ and $\Lambda_{n,k}$, with rational function symbol $r_{n,k}(z)$ and best approximates Γ_n, we are interested in studying the possibility of $\|\Lambda_{n,k} - \Lambda_k\| \to 0$ and $\|r_{n,k} - r_k\|_\infty \to 0$ as $n \to \infty$. For instance, for computational purposes, if Γ has infinite rank, we choose $\{\Gamma_n\}$ to be a sequence of finite-rank approximants of Γ, by truncations, say. This is called the problem of "continuity of best Hankel approximation." The main result in this paper is the following theorem.

THEOREM 3.1. *Let Γ be a bounded Hankel operator with s-numbers $s_0 \geq s_1 \geq \cdots$. Suppose that $s_{k-1} > s_k \geq \cdots$ and Γ_n are bounded Hankel operators satisfying*

$$\|\Gamma_n - \Gamma\| \to 0.$$

Let $\Lambda_{n,k}$ and Λ_k denote the best approximants from $G^{[k]}$ to Γ_n and Γ, respectively. Also, let $r_{n,k}(z)$ and $r_k(z)$ be their corresponding (strictly proper) rational symbol functions. Then

$$\|\Lambda_{n,k} - \Lambda_k\| \to 0 \quad \textit{and} \quad \|r_{n,k} - r_k\|_\infty \to 0, \quad \textit{as } n \to \infty.$$

Proof. Assume that the conclusion is false, and assume, without loss of generality, by taking a subsequence, if necessary, that $\|\Gamma_n - \Gamma\| \to 0$ but

$$\|\Lambda_{n,k} - \Lambda_k\| \geq \epsilon_0 > 0 \tag{3.1}$$

for all n, where $\Lambda_{n,k}$ and Λ_k are the best approximants from $G^{[k]}$ to Γ_n and Γ, respectively. Write

$$\Lambda_{n,k} = \begin{pmatrix} l_{n,1} & l_{n,2} & l_{n,3} & \cdots \\ l_{n,2} & l_{n,3} & \cdots & \\ l_{n,3} & \vdots & & \\ \vdots & & & \end{pmatrix}.$$

Then

$$\Lambda_{n,k}(e_1) = (l_{n,1}, l_{n,2}, \cdots)^T \in \ell^2.$$

Let $s_{n,0} \geq s_{n,1} \geq \cdots$ denote the s-numbers of Γ_n. For each $j, 0 \leq j < \infty$, we have

$$\begin{cases} \|\Gamma_n - \Lambda_{n,k}\| = s_{n,k}, & \text{and} \\ |s_{n,j} - s_j| \leq \|\Gamma - \Gamma_n\| \to 0, \end{cases}$$

where the second inequality can be found in [17, p. 30], and from this inequality, it follows that, for all sufficiently large n,

$$s_{n,k-1} > s_{n,k} \geq \cdots,$$

so that $\operatorname{rank}(\Lambda_{n,k}) = k$. Since each $\Lambda_{n,k}(e_1)$ is a bounded sequence in ℓ^2, it has a weakly convergent subsequence. Without loss of generality, suppose that $\Lambda_{n,k}(e_1)$ itself converges weakly to $(l_1, l_2, \ldots)^T \in \ell^2$. Then

$$l_{n,m} \to l_m, \quad \text{as } n \to \infty, \tag{3.2}$$

for each $m = 1, 2, \ldots$. Set

$$\tilde{\Lambda} := \begin{pmatrix} l_1 & l_2 & l_3 & \cdots \\ l_2 & l_3 & \cdots & \\ l_3 & \vdots & & \\ \vdots & & & \end{pmatrix}. \tag{3.3}$$

Since the minors of $\Lambda_{n,k}$ converge to the corresponding minor of $\tilde{\Lambda}$, we have $\operatorname{rank}(\tilde{\Lambda}) \leq k$. Furthermore, for any $y, z \in \ell^2$, we have

$$((\Gamma - \tilde{\Lambda})y, z) = ((\Gamma - \Gamma_n)y, z) + ((\Gamma_n - \Lambda_{n,k})y, z) + ((\Lambda_{n,k} - \tilde{\Lambda})y, z),$$

and thus,

$$\|\Gamma - \tilde{\Lambda}\| \leq \overline{\lim_{n\to\infty}} \|\Gamma_n - \Lambda_{n,k}\| = s_k.$$

It now follows from Theorem C that $\tilde{\Lambda} = \Lambda_k$ is the best approximant from $G^{[k]}$ to Γ with $\operatorname{rank}(\Lambda_k) = k$. By Theorem A, we also have

$$r_{n,k}(z) = \sum_{j=1}^{\infty} l_{n,j} z^{-j} = \frac{c_n(z)}{d_n(z)} = \frac{c_{n,1}z^{k-1} + c_{n,2}z^{k-2} + \cdots + c_{n,k}}{z^k + d_{n,1}z^{k-1} + \cdots + d_{n,k}}$$

and

$$r_k(z) = \sum_{j=1}^{\infty} l_j z^{-j} = \frac{c(z)}{d(z)} = \frac{c_1 z^{k-1} + c_2 z^{k-2} + \cdots + c_k}{z^k + d_1 z^{k-1} + \cdots + d_k}.$$

We may now apply (2.4) to both $r_{n,k}(z)$ and $r_k(z)$, noting that the coefficient matrices in (2.4) for both situations are nonsingular. Thus, it follows from (3.2) that

$$d_{n,m} \to d_m, \quad \text{as } n \to \infty,$$

for $1 \leq m \leq k$. So, applying (2.5) to both $r_{n,k}(z)$ and $r_k(z)$, we also have

$$c_{n,m} \to c_m, \quad \text{as } n \to \infty,$$

for $1 \leq m \leq k$. This immediately gives

$$\|r_{n,k} - r_k\|_\infty \to 0,$$

and hence,

$$\|\Lambda_{n,k} - \Lambda_k\| \leq \|r_{n,k} - r_k\|_\infty \to 0,$$

which contradicts (3.1) and completes the proof of the theorem. □

In the following, we give an example to demonstrate the sharpness of Theorem 3.1, in the sense that if the s-number s_k is not simple, then the best Hankel approximation from $G^{[k+1]}$ is not necessarily continuous. This example is a modification of the one given in [19].

Example 3.1. Let $f(z) = z^{-1} - z^{-4}$. Then we have

$$\Gamma_f = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & \cdots \\ 0 & 0 & -1 & 0 & \cdots & \\ 0 & -1 & 0 & & & \\ -1 & 0 & & & & \\ 0 & \vdots & & & & \\ \vdots & & & & & \end{pmatrix},$$

and the s-numbers of Γ_f are given by

$$s_0 = \frac{\sqrt{5}+1}{2}, \quad s_1 = s_2 = 1, \quad s_3 = \frac{\sqrt{5}-1}{2}, \quad s_k = 0 \quad \text{for } k \geq 4.$$

So, for $k = 1$ or $k+1 = 2$, the best Hankel approximant of Γ_f from $G^{[2]}$ is

$$\Lambda_2 = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 0 & & \\ 0 & & & \\ \vdots & & & \end{pmatrix}$$

and its rational symbol is given by $r_2(z) = z^{-1}$. Now, if we consider

$$b_\alpha(z) := \frac{\alpha z - 1}{z - \alpha} = 1 + (\alpha - 1)\frac{z+1}{z-\alpha}$$

and $f_\alpha(z) := f(z)b_\alpha(z)$, we have

$$\lim_{\alpha \uparrow 1} \|\Gamma_f - \Gamma_{f_\alpha}\| \leq \lim_{\alpha \uparrow 1} \|f - f_\alpha\|_\infty = 0.$$

On the other hand, it is easy to verify that the best Hankel approximant of Γ_{f_α} from $G^{[2]}$ is $\Lambda_{\alpha,2} = \Gamma_{r_{\alpha,2}}$, where

$$r_{\alpha,2}(z) := r_2(z)b_\alpha(z).$$

Observe, however, that

$$\begin{aligned} \|\Lambda_2 - \Lambda_{\alpha,2}\| &= \|\Gamma_{(b_\alpha(z)-1)/z}\| \\ &\geq \|\Gamma_{b_\alpha - 1}\| \\ &= \|\Gamma_{b_\alpha}\| = 1. \end{aligned}$$

Hence, best Hankel approximation from $G^{[2]}$ is not continuous at f. It is also obvious that

$$\|r_{\alpha,2} - r_2\|_\infty = 2 \not\to 0 \quad \text{as } \alpha \uparrow 1.$$

4. Near-best Hankel approximants via truncations. Results on Hankel operators and Hankel-type approximation have been widely used in dealing with many problems in engineering such as systems reduction, systems identification, robust stability, and H^∞-control [9], [14], [12], [25], [21], [5]. When the Hankel operators associated with the systems have infinite rank, it is generally quite difficult to compute the best Hankel approximants. The only exceptions are systems of special types such as those with delays, etc., as discussed in [11], [8], [25]. Hence, some kinds of truncations, such as truncations of a balanced realization, are considered an intermediate step to finding the best approximation (cf. [15], [16], [24], [27]). In this section we give a general study of truncated Hankel operators, through which near-best Hankel approximants can be computed.

Let Γ_n be a sequence of bounded Hankel operators that converge to a given Hankel operator Γ, and let $\Lambda_{n,k}$, Λ_k be their corresponding best Hankel approximants from $G^{[k]}$. Then we have

$$\begin{aligned} (4.1) \quad s_k(\Gamma) &\leq \|\Gamma - \Lambda_{n,k}\| \leq \|\Gamma - \Gamma_n\| + \|\Gamma_n - \Lambda_{n,k}\| \leq s_{n,k}(\Gamma_n) + \|\Gamma - \Gamma_n\| \\ &\leq s_k(\Gamma) + 2\|\Gamma - \Gamma_n\|. \end{aligned}$$

Moreover, if $s_{k-1}(\Gamma) > s_k(\Gamma) \geq \cdots$, it follows from Theorem 3.1 that

$$\|\Lambda_{n,k} - \Lambda_k\| \to 0.$$

Therefore, we may call $\Lambda_{n,k}$ a near-best Hankel approximant of Γ.

In the following discussion, we assume that Γ is compact. It is then well known that its s-numbers $s_k(\Gamma)$ converge to 0. Hence, a sequence of Hankel operators Γ_n with finite rank can be chosen so that Γ_n converges to Γ. An application of the Carathéodory-Fejér (CF) algorithm to Γ_n yields its near-best Hankel approximant $\Lambda_{n,k}$. A discussion of the computation of $\Lambda_{n,k}$ from finite rank Γ_n is given in [30, §3], where the CF theorem is used to study the near-best rational Chebyshev approximation. In the following, we discuss the convergence of near-best approximants through a fairly general class of truncations.

Let Γ be given as in (2.2) and set

$$\gamma_n := (h_1, h_2, \ldots, h_n)^T.$$

Suppose that a sequence $\mathcal{M}$ of $n \times n$ matrices M_n, $n = 1, 2, \ldots$, is chosen. We consider the column vectors

$$\gamma_{n,n} := M_n \gamma_n = (h_{n,1}, h_{n,2}, \ldots, h_{n,n})^T \tag{4.2}$$

and the corresponding Hankel matrix

$$\Gamma_{n,\mathcal{M}} = \begin{pmatrix} h_{n,1} & h_{n,2} & \ldots & h_{n,n} & 0 & \ldots \\ h_{n,2} & & \iddots & \iddots & & \\ \vdots & \iddots & \iddots & & & \\ h_{n,n} & & & & & \\ 0 & & & & & \\ \vdots & & & & & \end{pmatrix}. \tag{4.3}$$

We call $\Gamma_{n,\mathcal{M}}$ the truncated Hankel operators of Γ relative to $\mathcal{M}$. It is obvious that different choices of $\mathcal{M}$ give different sequences of truncated Hankel operators $\Gamma_{n,\mathcal{M}}$. In the following, we discuss three kinds of truncations that are useful for computational purposes.

Example 4.1. Denote by $\mathcal{I}$ the sequence of identity matrices $I_{n\times n}$. Then the truncated Hankel operators $\Gamma_{n,\mathcal{I}}$ of Γ relative $\mathcal{I}$ are the truncated matrices

$$\Gamma_{n,\mathcal{I}} = \begin{pmatrix} h_1 & \ldots & h_n & 0 & \ldots \\ \vdots & & \iddots & \iddots & \\ h_n & \iddots & \iddots & & \\ 0 & \iddots & & & \\ \vdots & & & & \end{pmatrix}$$

considered in our earlier work [5] and [6].

Example 4.2. Let $0 < r < 1$ and $\mathcal{B}_r$ be the sequence of matrices

$$\mathcal{B}_{n,r} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & r & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & r^{n-1} \end{pmatrix}.$$

Then the truncated Hankel operators $\Gamma_{n,\mathcal{B}_r}$ associated with Γ are the infinite matrices

$$\Gamma_{n,\mathcal{B}_r} = \begin{pmatrix} h_1 & rh_2 & \dots & r^{n-1}h_n & 0 & \dots \\ rh_2 & & \iddots & \iddots & & \\ \vdots & \iddots & \iddots & & & \\ r^{n-1}h_n & \iddots & & & & \\ 0 & & & & & \\ \vdots & & & & & \end{pmatrix}.$$

As we will see, the importance of this type of truncation is that for some suitable choices of r_k and $n = n_k$, $\Gamma^n_{\mathcal{B}_{r_k}}$ always converge to the compact Hankel operator Γ.

Example 4.3. Consider the sequence $\mathcal{C}$ of matrices

$$C_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1-\frac{1}{n} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1-\frac{n-1}{n} \end{pmatrix}.$$

Then the truncated Hankel operators $\Gamma_{n,\mathcal{C}}$ associated with Γ is

$$\Gamma_{n,\mathcal{C}} = \begin{pmatrix} h_1 & \left(1-\frac{1}{n}\right)h_2 & \dots & \left(1-\frac{n-1}{n}\right)h_n & 0 & \dots \\ \left(1-\frac{1}{n}\right)h_2 & & \iddots & \iddots & & \\ \vdots & \iddots & \iddots & & & \\ \left(1-\frac{n-1}{n}\right)h_n & \iddots & & & & \\ 0 & & & & & \\ \vdots & & & & & \end{pmatrix}.$$

By an application of Theorem B, it is easy to see that $\Gamma_{n,\mathcal{C}}$ converges to Γ.

In the following, we give some sufficient conditions on Γ under which the truncated Hankel operators discussed above converge to Γ.

PROPOSITION 4.1. *Suppose that the symbol function* $\sum_{i=1}^{\infty} h_i z^{-i}$ *of the Hankel operator* $\Gamma = \{h_{i+j-1}\}_{1\le i,j<\infty}$ *is in the Wiener class* W. *Then*

$$\|\Gamma - \Gamma_{n,\mathcal{I}}\| \to 0.$$

Proof. By Theorem B, we have

$$\|\Gamma - \Gamma_{n,\mathcal{I}}\| \le \left\| \sum_{i=1}^{\infty} h_i z^{-i} - \sum_{i=1}^{n} h_i z^{-i} \right\|_{\infty} \le \sum_{i=n+1}^{\infty} |h_i|,$$

which converges to 0. □

PROPOSITION 4.2. *Let* Γ *be compact. Then there exists a sequence* $r_k \uparrow 1$ *as* $k \to \infty$ *and a sequence of positive integers* n_k *such that*

$$\|\Gamma - \Gamma_{n_k,\mathcal{B}_{r_k}}\| \to 0.$$

Proof. Since $\Gamma = \{h_{i+j-1}\}_{1\le i,j<\infty}$ of compact, it is known (cf. [18]) that the Hankel operators $\Gamma_r := \{r^{i+j-2}h_{i+j-1}\}_{1\le i,j<\infty}$ converge to Γ as $r \to 1^-$. For each fixed r, $0 < r < 1$, we have

$$\begin{aligned}
\|\Gamma_r - \Gamma_{n,\mathcal{B}_r}\| &\le \left\| \sum_{k=1}^{\infty} r^{k-1}h_k z^{-k} - \sum_{k=1}^{n} r^{k-1}h_k z^{-k} \right\|_\infty \\
&= \left\| \sum_{k=n+1}^{\infty} r^{k-1}h_k z^{-k} \right\|_\infty \\
&\le \sup_{|z|=1} \left| \sum_{k=n+1}^{\infty} r^{k-1}h_k z^{-k} \right| \\
&\le \left(\sum_{k=n+1}^{\infty} r^{2(k-1)} \right)^{1/2} \left(\sum_{k=n+1}^{\infty} |h_k|^2 \right)^{1/2} \\
&\le \|\Gamma\| \frac{r^n}{\sqrt{1-r^2}},
\end{aligned}$$

so that

$$\lim_{n\to\infty} \|\Gamma_r - \Gamma_{n,\mathcal{B}_r}\| \to 0.$$

Therefore, we can choose a sequence $r_k \to 1^-$ and integers of n_k such that

$$\|\Gamma - \Gamma_{n_k,\mathcal{B}_{r_k}}\| \to 0. \qquad \square$$

PROPOSITION 4.3. *If the symbol function* $f(z) = \sum_{i=1}^{\infty} h_i z^{-i}$ *of* Γ *is continuous on the unit circle, then*

$$\|\Gamma - \Gamma_{n,\mathcal{C}}\| \to 0.$$

Proof. The Fejér sequence of functions $G_n(z)$ associated with $f(z)$ is given by

$$\begin{aligned}
G_n(z) &= \frac{1}{n} \left(\sum_{k=1}^{n} \left(\sum_{i=1}^{k} h_i z^{-i} \right) \right) \\
&= h_1 z^{-1} + \left(1 - \frac{1}{n}\right) h_2 z^{-2} + \cdots + \left(1 - \frac{n-1}{n}\right) h_n z^{-n}.
\end{aligned}$$

Since $f(z)$ is continuous, we have

$$\|f - G_n\|_\infty \to 0, \quad \text{as } n \to \infty.$$

Moreover, from the fact that $\Gamma_{n,\mathcal{C}} = \Gamma_{G_n}$, we have

$$\|\Gamma - \Gamma_{n,\mathcal{C}}\| \le \|f - G_n\|_\infty \to 0, \quad \text{as } n \to \infty.$$

This completes the proof of the proposition. $\square$

If Γ already has finite rank, then the precise rate of convergence of $\Gamma_{n,\mathcal{I}}$ to Γ is given in [5]. However, one cannot say too much about the convergence rate in general. Indeed, for any sequence $\varepsilon_k \downarrow 0$, there exists a Hankel operator Γ that satisfies $s_k(\Gamma) \ge \varepsilon_k$ for all k (cf. [23], [31]). Moreover, an explicit formulation of such a Γ is

given in [7]. Consequently, for any sequence $\varepsilon_n \downarrow 0$, there is a compact Hankel operator Γ such that

$$\|\Gamma - \Gamma_{n,\mathcal{M}}\| \geq \varepsilon_n, \qquad n \geq 1,$$

for any sequence $\mathcal{M}$ of $n \times n$ matrices M_n. A very important problem is, therefore, how to select an appropriate truncation so as to give a sufficiently good rate of convergence. Once the truncations $\Gamma_{n,\mathcal{M}}$ are determined, the CF algorithm or Kung's algorithm [21] can be applied to find the near-best Hankel approximants $\Lambda_{n,k}$.

REFERENCES

[1] V. M. Adamjan, D. Z. Arov, and M. G. Krein, *Infinite Hankel matrices and generalized Carathéodory-Fejér and Riesz problems*, Functional Anal. Appl., 2 (1968), pp. 1–18.

[2] ———, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31–73.

[3] N. I. Achiezer, *Lectures on the Theory of Approximations*, 2nd rev. ed., Izdat, Nauka, Moscow, 1965; 1st ed., Frederick Ungar, New York, 1956. (In English.)

[4] C. Carathéodory and L. Fejér, *Über den zusammenhang der extremen von harmonischen funktionen mit ihren koeffizienten and über der Picard–Landauschen Satz*, Rend. Circ. Mat. Palermo, 32 (1911), pp. 218–239.

[5] C. K. Chui, X. Li, and J. D. Ward, *System reduction via truncated Hankel matrices*, Math. Control Signal Systems, 4 (1991), pp. 161–175.

[6] ———, *Rate of uniform convergence of rational functions corresponding to best approximants of truncated Hankel operators*, Math. Control Signal Systems, 5 (1992), pp. 67–79.

[7] ———, *On the convergence rate of s-numbers of compact Hankel operators*, Circuits Systems Signal Process., 2 (1992), pp. 353–362.

[8] R. F. Curtain, K. Glover, and J. Lam, *Reduced order models for distributed systems based on optimal Hankel-norm approximations*, in Proc. 5th Virginia Polytechnic Inst. and State Univ./American Inst. for Aeronautics and Astronomics Sympos. on Dynamics and Control of Large Structures, L. Meirovitch, ed., Blacksburg, VA, June 12–14, 1985, pp. 231–244.

[9] R. F. Curtain and A. C. M. Ran, *Explicit formulas for Hankel norm approximations of infinite-dimensional systems*, Integral Equations Operator Theory, 12 (1989), pp. 455–469.

[10] C. Foias and A. Tannenbaum, *On the Nehari problem for a certain class of L^∞-functions appearing in control theory*, J. Funct. Anal., 74 (1987), pp. 146–159.

[11] C. Foias, A. Tannenbaum, and G. Zames, *On the H^∞-optimal sensitivity problem for systems with delays*, SIAM J. Control Optim., 25 (1987), pp. 686–706.

[12] B. A. Francis, *A Course in H_∞ Control*, Springer–Verlag, New York, 1986.

[13] F. R. Gantmacher, *Theory of Matrices*, Vol. 2, Chelsea, New York, 1959.

[14] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L_∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[15] K. Glover, J. Lam, and J. R. Partington, *Balanced realization and Hankel-norm approximation of systems involving delays*, in Proc. 25th IEEE Conf. Decision Control, Athens, GA, December 1986, pp. 1810–1815.

[16] K. Glover, R. F. Curtain, and J. R. Partington, *Realization and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.

[17] I. C. Gohberg and M. G. Krein, *Introduction of the Theory of Linear Non-selfadjoint Operators in Hilbert Space*, American Mathematical Society, Providence, RI, 1969.

[18] P. Hartman, *On completely continuous Hankel matrices*, Proc. Amer. Math. Soc., 9 (1958), pp. 862–866.

[19] E. Hayashi, L. N. Trefethen, and M. H. Gutknecht, *The CF table*, Constr. Approx., 6 (1990), pp. 195–223.

[20] J. W. Helton and D. F. Schwartz, *The best approximation to a vector-valued continuous function from the bounded analytic functions*, preprint.

[21] S. Y. Kung, *Optimal Hankel-norm model reductions: scalar systems*, in Proc. Joint Automatic Control Conference, 1980, paper PA8-D.

[22] Z. Nehari, *On bounded bilinear forms*, Ann. of Math., 65 (1957), pp. 153–162.

[23] R. OBER, *A note on a system theoretic approach to a conjecture by Peller–Khrushcher: The general case*, IMA J. Math. Control Inform., 7 (1990), pp. 35–45.

[24] ———, *Balanced parametrization of classes of linear systems*, SIAM J. Control Optim., 29 (1991), pp. 1251–1287.

[25] J. R. PARTINGTON, *An introduction to Hankel operators*, Cambridge University Press, London, 1988.

[26] V. V. PELLER, *Hankel operators and continuity properties of the operators of best approximation*, Leningrad Math. J., 2 (1991), pp. 139–160.

[27] L. PERNEBO AND L. M. SILVERMAN, *Model reduction via balanced state-space representation*, IEEE Trans. Automat. Control, 27 (1982), pp. 382–387.

[28] I. SCHUR, *Über Potenzreihen, die im innern des einheitskresises beschränkt sind*, J. Reine Angew. Math., 148 (1918), pp. 122–145.

[29] T. TAKAGI, *On an algebraic problem related to an analytic theorem of Carathéodory and Fejér*, Japan J. Math., 1 (1924), pp. 83–93; 2 (1925), pp. 13–17.

[30] L. N. TREFETHEN, *Rational Chebyshev approximation on the unit disc*, Numer. Math., 37 (1981), pp. 297–320.

[31] S. R. TREIL, *Moduli of Hankel operators and a problem of V. V. Peller and S. V. Krushschev*, Soviet Math. Dokl., 32 (1985), pp. 293–297.

[32] L. Y. WANG AND G. ZAMES, *Lipschitz continuity of H_∞ interpolation*, Systems Control Lett., 14 (1990), pp. 381–387.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 32, No. 6, pp. 1782–1791, November 1994

016

ON ONE IDENTIFICATION PROBLEM FOR DISTRIBUTED CONTROLLABLE SYSTEMS*

BENZION SHKLYAR† AND VICTOR BAKHMUTSKY‡

Abstract. An identification problem for distributed control systems in the class of approximate null-controllable systems is considered. A criterion of identifiability of linear autonomous distributed systems has been proved, and is illustrated by an example of a system with small nonlinearity. Further development and generalization of the obtained results are discussed.

Key words. linear systems, distributed systems, systems with delays, identifiability, observability, controllability

AMS subject classifications. 93C25, 93D15

1. Introduction.

The main purpose of identification of control systems is to determine a mathematical model of a control object by using measured experimental data. The real control objects are described, as a rule, by nonlinear distributed-parameter systems. In the first approximation these objects can be modeled by the linear autonomous systems with distributed parameters.

This paper is devoted to the problem of determining the input parameters for a class of controllable systems. Controllability is a fundamental property of dynamic systems [1], therefore it is natural to use controllable systems for identification. In this paper we will use the class of approximately null-controllable systems. The obtained general results are used in an illustrative example of an electro-mechanical control system described by Minorsky's equation.

2. Problem statement.

Let X, Y, Z be Banach space. Consider the equation

$$\dot{x}(t) = Ax(t) + Gbu(t), \tag{1}$$

$$x(0) = x_0, \tag{2}$$

$$y(t) = Cx(t), \qquad 0 \le t \le t_1, \tag{3}$$

where $x(t)$ is the current state; x_0 is the initial state; $u(t)$ is the piecewise continuous control, $0 \le t \le t_1$; A is a linear operator whose domain $D(A)$ is dense in X; $G : Z \to X$ is a linear bounded operator; $b \in Z$; $C : X \to Y$ is a linear bounded operator; $x, x_0 \in X, y \in Y, u \in R^1$. We assume that problem (1)–(2) is uniformly well posed [2]. It follows from this assumption that A generates a strongly continuous semigroup $S(t)$ of operators in the class C_0 [2]. We consider only weak solutions of the above equation.

We assume A to have an additional property:

(i) there exists a number $T \ge 0$ such that the attainability set $K(t)$ of (1)-(2) is independent of t if $t > T$ $\forall b \in Z$.

* Received by the editors October 13, 1992; accepted for publication (in revised form) July 16, 1993. This work was supported by the Ministry of Science of Israel.

† Department of Mathematics and Computer Science, Bar-Ilan University, Ramat-Gan 52 900, Israel (shklyar@bimacs.bitnet).

‡ Center for Technological Education Holon, affiliated with Tel-Aviv University, Holon, Israel.

If $x \in X$ and $f \in X^*$, we write (x, f) instead of $f(x)$.

We assume that operators A, G and the initial state x_0 are known and b is unknown. The main purpose of the given identification problem is to determine b using experimental data. If b can be uniquely determined, (1)–(3) is said to be identifiable.

The identification problem of this kind was investigated in [3] for an important particular case: when X is a Hilbert space, A is a symmetric coercive operator with domain $D(A) = V \subset X \subset V^*$($V$ is also Hilbert space, V is dense in X), $Z = X, G = I, b \in V^*, Y = R^m, m \geq 1$, and some additional assumptions (see [3, pp. 474–475]). It is proved in [4] that under assumptions of [3], property (i) holds for $T = 0$.

In the present paper the criterion of identifiability for (1)–(2) has been established. We can find the unknown b by the least square fit. If (1) with the determined b is not approximately null-controllable, it is possible to approximate b by a b_ϵ such that (1) with b replaced by b_ϵ is approximately null-controllable.

3. Necessary and sufficient identifiability conditions.

DEFINITION 3.1. *Equation* (1)–(3) *is said to be identifiable on* $[0, t_1]$ *if* b *is uniquely determined by output* (3), $0 \leq t \leq t_1$.

Consider the equation

$$\dot{x}(t) = Ax(t), \tag{4}$$

$$y(t) = Cx(t), \qquad 0 \leq t \leq t_1. \tag{5}$$

DEFINITION 3.2. *Equation* (4)–(5) *is said to be observable on* $[0, t_1]$ *with respect to* G *if* z_0 *is uniquely determined by output* (5), $0 \leq t \leq t_1$.

We consider $t_1 > T$ and denote by σ the spectrum of the operator A. The identifiability criterion of equation (1)–(3) will be proved in this paragraph.

THEOREM 3.3. *Equation* (1)–(3) *is identifiable on* $[0, t_1]$ *if and only if:*

1. *equation* (4)–(5) *is observable on* $[0, t_1]$ *with respect to* G;
2. $u(t) \not\equiv 0$ *on the* $[0, t_1]$.

Proof. We first prove the sufficiency. The weak solution of (1)–(2) is given by the following formula [2]:

$$x(t) = S(t)x_0 + \int_0^{t_1} S(t-\tau)Gbu(\tau)d\tau. \tag{6}$$

By the problem statement, $S(t), G$, and x_0 are assumed to be known. Hence from (3) and (6) it follows that (1)–(3) is identifiable on $[0, t_1]$ if and only if the identity

$$C\int_0^{t_1} S(t\tau)Gbu(\tau)d\tau \equiv 0, \qquad 0 \leq t \leq t_1 \tag{7}$$

implies that the b is equal to 0. Denote by $R(\xi) = (\xi I - A)^{-1}$ the resolvent of operator A [2], $\xi \notin \sigma$, $U(\xi) = \int_0^{t_1} u(t)\exp(-\xi t)dt$. By the independence of the attainable set of (1)–(3) of t for $t > T$, we obtain the identity

$$C\int_0^{t_1} S(t-\tau)Gbu(\tau)d\tau \equiv 0, \qquad 0 \leq t < \infty \tag{8}$$

from (8) [4]. Applying to (8) the Laplace transform, we obtain

$$CR(\xi)GbU(\xi) \equiv 0 \tag{9}$$

for any complex $\xi \notin \sigma$. Since $u(t) \not\equiv 0$ on $[0, t_1]$, from the last identity we obtain

$$CR(\xi)Gb \equiv 0, \qquad \forall \xi \notin \sigma. \tag{10}$$

Applying the inverse Laplace transform to (10), we obtain

$$CS(t)Gb \equiv 0, \qquad 0 \leq t \leq t_1. \tag{11}$$

From (11), by the observability of (4)–(6) on $[0, t_1]$ with respect to G, we have $b = 0$. The sufficiency has been proved.

Now we prove the necessity. The necessity of the second condition of the theorem is obvious. Now assume that the first condition of the theorem does not hold, i.e., there exists a $b, b \neq 0$, such that (11) holds. Then (8) follows from (11) and this fact contradicts the identifiability of (1)-(3) on $[0, t_1]$. □

The explicit observability conditions for various particular cases of (4)–(6) are given in many publications (see, for instance, [6]–[10]). We give below a sufficient condition of observability with respect to G for (4)–(6).

We assume that operator A has the following additional properties:

(ii) the operator A has a purely point spectrum σ that is either finite or has no finite limit points and each $\lambda \in \sigma$ has a finite multiplicity;

(iii) there is a time $T \geq 0$ such that for each $x_0 \in X$ and $t > T$ the function $x(t) = S(t)x_0$ is expanded in a series of eigenfunctions and associated functions of A converging uniformly with respect to t on an arbitrary interval $[T_1, T_2], T_2 > T_1 > T$ for a certain grouping of terms.

THEOREM 3.4. *Equation* (4)–(6) *is observable on* $[0, t_1]$ *with respect to* G *if*

1. *the system of equations*

$$Ax - \lambda x = 0, \qquad Cx = 0 \tag{12}$$

has only the trivial solution for each $\lambda \in \sigma$;

2. *the system of eigenvectors and associated vectors of operator* A^* *is complete*;

3. *the equation* $Gx = 0$ *has only the trivial solution.*

Proof. We assume that identity

$$CS(t)Gz \equiv 0, \quad 0 \leq t \leq t_1, \quad z \in Z \tag{13}$$

is true.

In virtue of independence of the attainable set of (1)–(3) of t when $t > T$ we obtain from (13) [4] the identity

$$CS(t)Gz \equiv 0, \qquad 0 \leq t < \infty. \tag{14}$$

Applying the Laplace transform to (14) we obtain

$$CR(\xi)Gz \equiv 0, \qquad \forall \xi \notin \sigma. \tag{15}$$

It is known [5] that each $\lambda_j \in \sigma, j = 1, 2, \ldots$ is a pole of the resolvent $R(\xi)$ and the function $R(\xi)Gz$ has the Laurent expansion

$$R(\xi)Gz = \gamma_{j1}(\xi - \lambda_1)^{-\beta_j} + \cdots + \gamma_{j\beta_j}(\xi - \lambda_1)^{-1} + R_j(\xi) \tag{16}$$

in a neighborhood of λ_j, where the operator-valued function $R_j(\xi)$ is holomorphic in this neighborhood.

Let φ_{ij} and ψ_{ij} be the eigenvectors and associated vectors of A and A^*, respectively, corresponding to an eigenvalue $\lambda_i \in \sigma$ such that

$$(17) \quad (\varphi_{ij}, \psi_{kl}) = \delta_{ik}\delta_{jl}, \quad i,k = 1,2,\ldots, \quad j = 1,2,\ldots,\beta_i, \quad l = 1,2,\ldots,\beta_k.$$

By assumption (iii) and (17) we obtain [4]

$$(18) \quad \gamma_{jl} = \sum_{k=\beta_j-l+1}^{\beta_j} (Gz, \psi_{jk})\varphi_{jk+l-\beta_j}, \quad l = 1,2,\ldots,\beta_j, \quad j = 1,2,\ldots.$$

It follows from (15)–(18) (see [4, p. 329]), that

$$(19) \quad \sum_{l=1}^{\beta_j-p} (Gz, \psi_{jl+p})C\varphi_{jl} = 0, \quad p = 0,1,\ldots,\beta_j - 1, \quad j = 1,2,\ldots.$$

Since (12) has only the trivial solution (condition 1 of Theorem 3.4) we obtain from (19) that

$$(20) \quad (Gz, \psi_{jl}) = 0, \quad j = 1,2,\ldots, \quad l = 1,2,\ldots,\beta_j,$$

and using completeness of vectors ψ_{jl} (condition 2 of Theorem 3.4) we obtain from (20) that $Gz = 0$. Hence $z = 0$ (see condition 3 of Theorem 3.4). □

Note. If X, Y are Hilbert spaces, A is a self-adjoint operator, $Z = X, G = I$ and A has the properties given in [3, pp. 474–475], the trivial solvability of (12) is the criterion of observability for (1)–(3) and coincides with condition (i) of Theorem 4 from [3, p. 475].

4. Identifiability of delay systems. Now we consider an important class of (1)–(3), to which results of [3] cannot be applied.

Consider the differential-difference system

$$(21) \quad \dot{v}(t) = A_0 v(t) + A_1 v(t-h) + cu(t),$$

$$(22) \quad v(0) = x_0, \quad v(\tau) = \varphi(\tau) \quad \text{a.e. on } [-h,0],$$

$$(23) \quad w(t) = Kv(t), \qquad 0 \le t \le t_1,$$

where $v, v_0 \in R^n, u \in R^1, A_j, j = 0,1$ are constant nxn-matrices, $c \in R^n, 0 < h, \varphi(.) \in L_2^n[-h,0]; y \in R^p, K$ is pxn-matrix. System (21)–(23) is an important particular case of (1)–(3) [2], [4], [11], [12]. Denote $x(t) = \{v(t), v_t(.)\}$, where $v_t(.) = v(t+\tau), -h \le \tau \le 0, t \ge 0, X = M_2^n[-h,0] = R^n X L_2^n[-h,0]$. It is known [11], [12] that if $v(t)$ is the solution of (21)–(22), then $x(t) \in M_2^n[-h,0]$ for all $t \ge 0$ and $x(t)$ is the solution of (1)–(2), where

$$X = M_2^n[-h,0] = R^n X L_2^n[-h,0], \quad Y = R^p, \quad Z = R^n;$$

$$(24) \quad Ax = \{A_0\varphi(0) + A_1\varphi(-h), \dot{\varphi}(.)\}$$

with the domain

$$D(A) = \{x = \{v_0, \varphi(.)\} : v_0 = \varphi(0), \dot{\varphi}(0) = A_0\varphi(0) + A_1\varphi(-h)\}$$

and with the spectrum

$$\sigma = \{z \in \mathbf{C} : zI - A_0 - A_1 xp(-zh) = 0\},$$

where $\mathbf{C}$ is the complex plane. For each $v_0 \in R^n$,

$$Gv_0 = \{v_0, 0\} \tag{25}$$

and operator $C : M_2^n[-h, 0] \to R^p$ is given by the formula

$$Cx = Kv_0, \quad \forall x = v_0, \quad \varphi(.) \in M_2^n[-h, 0]. \tag{26}$$

The operator A in this case is not self-adjoint and does not posses the properties described in [3], therefore it is impossible to apply results of [3] to this case, but we can apply Theorem 3.3 of the present paper.

DEFINITION 4.1. *System* (21)–(23) *is said to be identifiable on* $[0, t_1]$ *if* c *is uniquely determined by output* (33), $0 \le t \le t_1$.

DEFINITION 4.2. *System* (21)–(23) *with* $u(\tau) \equiv 0$ *is said to be relatively observable on* $[0, t_1]$ *if* $v(0)$ *is uniquely determined by output* (23), $0 \le t \le t_1$, *and by known function* $\varphi(.) \in L_2^n[-h, 0]$.

The criterion of relative observability for system (21)–(23) with $u(\tau) \equiv 0$ was proved in [6].

The relative observability of (21)–(23) with $u(\tau) \equiv 0$ is equivalent to observability of (4)–(6) with respect to operator G, where operators A, C, and G are defined by (24)–(26). The corresponding initial problem is uniformly well posed on $[0, t_1]$ and the attainable set for this system is independent of t if $t > nh$ [13].

It is well known that the solution of (21)–(22) is given by the formula [14]

$$v(t) = F(t)v_0 + \int_0^t F(t-\tau)A_1\varphi(\tau - h)d\tau + \int_0^t F(t-\tau)cu(\tau)d\tau, \tag{27}$$

where $F(t)$ is the fundamental matrix of (21) [14].

For (21) we consider its defining equation [15] as follows:

$$\begin{aligned} Q_{i+1j} &= A_0 Q_{ij} + A_1 Q_{i-1,j}, \qquad i, j = 0, 1, \ldots; \\ Q_{00} &= I, \quad Q_{ij} = 0, \quad i < 0, \quad j < 0; \\ X_{ij} &= Q_{ij}^T K^T, \qquad i, j = 0, 1, \ldots. \end{aligned} \tag{28}$$

It follows from Theorem 3.3 and the results of [6] that the following theorem is valid.

THEOREM 4.3. *For* (21)–(23) *to be identifiable on* $[0, t_1]$ *it is necessary and sufficient for* $t_1 > nh$, *that*

$$\operatorname{rank}\{X_{ij}, i, j = 0, 1, \ldots, n-1\} = n; \tag{29}$$

$$u(t) \not\equiv 0 \quad \text{on } [0, t_1].$$

Using (27) we consider a functional

$$J(c) = \int_0^{t_1} \|z(t) - q(t)\|^2 dt, \tag{30}$$

where

$$q(t) = \int_0^t KF(t-\tau)cu(\tau)d\tau,$$

$$z(t) = w(t) - K\left(F(t)v_0 - \int_0^t F(t-\tau)A_1\varphi(\tau-h)d\tau\right),$$

where $w(t), 0 \le t \le t_1$ is a known function obtained as a result of measurements, v_0 and $\varphi(\tau)$ are known initial conditions. The identification problem formulated in this section is to minimize functional (30) with respect to c.

To solve this problem we should find the derivative (gradient) $J'(c)$. We denote

$$P(t)c = \int_0^t KF(t-\tau)cu(\tau)d\tau, \qquad 0 \le t < \infty. \tag{31}$$

Calculating the derivative, we obtain the following from $J'(b) = 0$:

$$\left(\int_0^{t_1} P^T(t)P(t)dt\right)c = \int_0^{t_1} P^T(t)z(t)dt.$$

It follows from identifiability of system (21)–(23) that matrix

$$W = \int_0^{t_1} P^T(t)P(t)dt$$

is positive. Hence there exists W^{-1} and we have

$$\begin{aligned} c &= \left(\int_0^{t_1} P^T(t)P(t)dt\right)^{-1} \int_0^{t_1} P^T(t)z(t)dt \\ &= \left(\int_0^{t_1}\int_0^t F^T(t-\tau)K^TKF(t-\tau)u^2(\tau)d\tau dt\right)^{-1} \\ &\quad \cdot \int_0^{t_1}\int_0^t F^T(t-\tau)K^Tu(\tau)d\tau z(t)dt. \end{aligned} \tag{32}$$

It is possible to use the formula [16]

$$F(t) = \sum_{i=0}^{\infty}\sum_{j=0}^{k} Q_{ij}\frac{(t-jh)^i}{i!}, \quad t \in [kh,(k+1)h), \quad k = 0,1,\ldots \tag{33}$$

for calculating the fundamental matrix $F(t)$.

The criterion of approximate null-controllability on $[0,t_1]$ for (21)–(22) is [4], [17] (this criterion was first obtained in [17] for neutral systems):

$$\operatorname{rank}\{zI - A_0 - A_1\exp(-zh), c\} = n \quad \text{for all } z \in \sigma. \tag{34}$$

If vector c obtained by (32) does not satisfy (34), then for each $\delta > 0$ we can find a vector c_δ such that $\| c - c_\delta \| < \delta$ and (21)–(23) with vector c_δ, respectively, is approximate null-controllable on $[0,t_1]$.

Note. In this case c depends continuously on the initial state (22). It is well known [3] that when a space of unknown vectors c is an infinite-dimensional one, the identifiability does not ensure that vector c depends continuously on the initial state.

4.1. Example. Consider, for instance, Minorsky's equation [18]

$$\kappa^{(2)}(t) + 2r\dot{\kappa}(t) + \omega^2\kappa(t) + 2q\dot{\kappa}(t-1) = \epsilon\dot{\kappa}^3(t) + bu(t), \tag{35}$$

where r, q, w, b, and ϵ are constants and ϵ is small value,

$$\kappa(0) = \kappa_0, \quad \dot{\kappa}(0) = \dot{\kappa}_0; \quad \kappa(\tau) = \varphi(\tau), \quad \dot{\kappa}(t) = \dot{\varphi}(\tau) \quad \text{a.e. on } [-1, 0].$$

This equation describes the behavior of some electro-mechanical control systems. Linearizing (35), we obtain the linear time-invariant equation with delay

$$\kappa^{(2)}(t) + 2r\kappa(t) + \omega^2\kappa(t) + 2q\dot{\kappa}(t-1) = bu(t). \tag{36}$$

This equation is reduced to system (21)–(23), where

$$A_0 = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -2r \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ 0 & -2q \end{bmatrix}, \quad c = \begin{bmatrix} 0 \\ b \end{bmatrix}, \quad v = \begin{bmatrix} \kappa \\ \dot{\kappa} \end{bmatrix},$$

$$\sigma = \{z \in \mathbf{C} : z^2 - 2rz - \omega^2 - 2qz\exp(-z) = 0\}.$$

We denote by $f_0(t)$ the solution of (36) on the interval $[0, \infty)$ when $u(t) \equiv 0$ with initial data $f_0(0) = 1, f_0^{(1)}(0) = 0$, and we denote by $f_1(t)$ the solution of (36) with initial data $f_1(0) = 0, f_1^{(1)}(0) = 1$ on interval $[0, \infty)$. It follows from (27) that the solution $x(t)$ of equation (36) is given by the formula

$$\kappa(t) = f_0(t)\kappa(0) + f_1(t)\dot{\kappa}(0) - 2q\int_0^t f_1(t-\tau)\varphi(t-\tau)d\tau + \int_0^t f_1(t-\tau)bu(\tau)d\tau. \tag{37}$$

Let $z(t)$ be a function observed on the output of the electro-mechanical system described by (35) with known parameters r, ω, q, ϵ, and an unknown parameter b. The identification problem in this case is the selection of parameters b such that the deviation

$$J(b) = \int_0^{t_1} \|(z(t) - x(t)\|^2 dt$$

is minimal. Here $x(t)$ is the solution of (36) defined by (37). Computing $X_{ij}, i, j = 0, 1, \ldots$ by means of (28) we obtain

$$X_{00} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad X_{10} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Hence (29) holds, and if $u(t) \not\equiv 0$ on the $[0, t_1]$ and $t_1 > 2$, then by virtue of Theorem 4.3, (36) is identifiable on $[0, t_1]$. So

$$\int_0^{t_1}\int_0^t f_1^2(t-\tau)u^2(\tau)d\tau \neq 0.$$

Using the method of the last section we obtain

$$b = \left(\int_0^{t_1}\int_0^t f_1^2(t-\tau)u^2(\tau)d\tau\right)^{-1}\int_0^{t_1}\int_0^t f_1(t-\tau)y(\tau)d\tau, \tag{38}$$

where $y(t)$ is the known function defined by the formula

$$y(t) = z(t) - f_0(t)\kappa(0) - f_1(t)\dot{\kappa}(0) + 2q\int_0^t f_1(t-\tau)\varphi(t-\tau)d\tau.$$

Note. We can calculate functions $f_0(t)$ and $f_1(t)$ for any $t \in [j, j+1), j = 0, 1, 2, \ldots$ by (33) or by means of the step method. We have

$$f_0(t+j) = f_0(t)f_0(j) + f_1(t)\dot{f}_0(j)\int_0^t f_1(t-\tau)2qf_0(\tau+j-1)d\tau,$$

$$f_1(t+j) = f_0(t)f_0(j) + f_1(t)\dot{f}_1(j)\int_0^t f_1(t-\tau)2qf_1(\tau+j-1)d\tau, \qquad 0 \le t < 1,$$

and, for instance, if $\mu^2 = r^2 - \omega^2 > 0$, then

$$f_0(t) = \frac{\exp(-rt)}{\mu} r\mathrm{sh}(\mu t) + \mu\mathrm{ch}(\mu t), \qquad f_1(t) = \frac{\exp(-rt)}{\mu}\mathrm{sh}(\mu t).$$

We can easily calculate $f_0(t)$ and $f_1(t), 0 \le t < 1$, when $r \le \omega$. It follows from (34) that (35) is approximately null-controllable on the $[0, t_1], t_1 > 2$, if and only if

$$\operatorname{rank}\begin{bmatrix} \lambda & 1 & 0 \\ \omega + 2q\exp(-\lambda) + 2r & 0 & b \end{bmatrix} = 2.$$

Hence (36) is approximately null-controllable on the $[0, t_1], t_1 > 2$ if and only if $b \neq 0$. If we have obtained $b = 0$, we can take $b = \gamma$, where γ is an arbitrary small, and (36) with $b = \gamma$ will be approximately null-controllable on $[0, t_1]$, for all $t_1 > 2$. Since the theorem on approximate null-controllability by the linear approximation holds [19], (it is possible to prove this statement by applying the results of [19] to (35) and using the equivalence of approximate null-controllability and exact null-controllability for hereditary systems [20], [21]) (35) with b, obtained from (37), will be approximately (locally) null-controllable on $[0, t_1]$ for all $t_1 > 2$. We consider the obtained equation as the approximate solution of the identification problem. The solution of the obtained equation can differ little from that of (35) due to the smallness value of the ϵ in (35).

5. Conclusion. In the present paper we have investigated and proved an identifiability criterion represented by the observability concept for the distributed parameter system (1)–(3), where the operator A is not necessarily self-adjoint and the solution of this system is not necessarily expanded into a series for $t \ge 0$. The system with delays is investigated as an example of a distributed system where the operator A is not self-adjoint and the solution of this equation can be expanded into a series only for $t \ge nh$. We obtained an explicit formula for the solution of the formulated identification problem.

In the present paper, as well as in [3], only the case $u(t) \in R^1$ is considered. This restricts the application of Theorems 3.3, 3.4, and 4.3. However, it is possible to apply these theorems for multi-input equations (1)–(3).

Let us consider term $GBu(t)$ instead of term $Gbu(t)$ in (1)–(3), where $u(t) \in R^r, r > 1$, and $B : R^r \to Z$ is the linear bounded operator; that is, $Bu = \sum_{j=1}^r b_j u_j$ for all $u = \mathrm{col}(u_1, u_2, \ldots, u_r), b_j \in Z, j = 1, 2, \ldots, r$. Since the control object acts with the different control functions $u(t)$, for each natural $k, 1 \le k \le r$, it is possible to

set $u_k(t) \not\equiv 0, u_j(t) \equiv 0$ on $[0, t_1], j = 1, \ldots, r, j \neq k$ and to measure a corresponding output $y(t)$. Using Theorems 3.3, 3.4, and 4.3, we can find each vector $b_k, k = 1, 2, \ldots, r$.

The identification problem for operator A is not investigated in the present paper. This problem is essentially more general and it was considered for a linear one-dimensional parabolic partial differential equation [22].

The illustrative example concerning the electro-mechanical nonlinear control system described by Minorsky's equation is investigated. In this problem we really found the linear approximation of Minorsky's equation, which solves the formulated identification problem.

Formula (37) is the main tool for the above purpose. In the authors' opinions, it is interesting to obtain the solutions of nonlinear systems with small nonlinearity, like Minorsky's equation, as an expansion into a series with respect to a small parameter. One method of such expansion is called Poincare's method and it is developed and applied to obtain stationary and transient solutions of ordinary differential equations [23], [24]. This expansion may be used instead of (37) in the identification problem.

REFERENCES

[1] M. Mesarovic and E. Takahara, *General System Theory: Mathematical Foundations*, Academic Press, New York, San Francisco, London, 1975.

[2] M. Krein, *Linear Differential Equations in a Banach Space*, Nauka, Moscow, 1967. (In Russian.)

[3] T. Kobayashi, *Determination of unknown functions for a class of distributed parameter systems,* SIAM J. Control. Optim., 17 (1979), pp. 469–476.

[4] B. Shklyar, *Controllability of linear systems with distributed parameters,* Differential Equations, 27 (1991), pp. 326–335.

[5] V. Hutson and G. Pym, *Applications of functional analysis and operator theory*, Academic Press, London, New York, Toronto, Sydney, San Francisco, 1980.

[6] R. Gabasov, F. Kirillova, R. M. Zhevnyak, and T. V. Kopeikina, *Conditional observability of linear systems*, Problems Control Inform. Theory, 1 (1972), pp. 217–238.

[7] A. Olbrot and S. Zak, *Controllability and observability problem for linear functional differential systems,* Found. Control Engrg., (1980), pp. 79–89.

[8] T. Kobayashi, *Initial state determination for distributed parameter systems,* SIAM J. Control Optim., 14 (1976), pp. 934–943.

[9] Y. Sakawa, *Observability and related problems for partial differential equations of parabolic type.* SIAM J. Control, 13 (1975), pp. 14–27.

[10] B. Shklyar, *On observability of linear systems with concentrated delays,* Soviet Phys. Dokl., 24 (1979), pp. 711–713.

[11] A. Manitius and R. Triggiani, *Function space controllability of linear retarded systems: A derivation from abstract operator conditions,* SIAM J. Control Optim., 14 (1976), pp. 599–645.

[12] J. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1977.

[13] H. Banks, M. Jacobs, and C. Langenhop, *Characterization of the controlled states in W_2^1 of linear retarded systems,* SIAM J. Control, 13 (1975), pp. 611–649.

[14] R. Bellman and K. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.

[15] F. Kirillova and S. Churakova, *On the problem of relative controllability of systems with aftereffect,* Differential Equations, 3 (1967), pp. 436–445.

[16] B. Shklyar, *The relative controllability of systems of neutral type with delayed arguments,* Differential Equations, 10 (1974), pp. 1116–1121.

[17] B. Shklyar, *On the controllability theory for neutral systems,* Izv. Akad. Nauk BSSR Ser. Fiz.-Mat., (1980), pp. 117–118. (In Russian.)

[18] N. Minorsky, *Self-Excited Mechanical Oscillations*, J. Appl. Phys., No 19 (1948), pp. 332–338.

[19] R. Underwood and D. Young, *Null controllability of nonlinear functional differential equations,* SIAM J. Control Optim., 17 (1979), pp. 753–772.

[20] F. COLONIUS, *On approximate and exact null controllability of delay systems,* Systems Control Lett., 5 (1984), pp. 209–211.

[21] A. W. OLBROT AND L. PANDOLFI, *Null controllability of a class of functional differential systems,* Internat. J. Control, 7 (1988) pp. 193–208.

[22] S. KITAMURA AND S. MAKAGIRI, *Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type*, SIAM J. Control Optim., 15 (1977), pp. 785–802.

[23] A. PROSKURYAKOV, *Poincare's method in the nonlinear oscillation theory*, Nauka, Moscow, 1977. (In Russian.)

[24] V. BAKHMUTSKY, *On application of Poincare's method for investigation of nonlinear oscillations*, Izv. Akad. Nauk USSR, (1961), pp. 84–90. (In Russian.)